Cognitive Processes in the Perception of Art
W.R. Crozier and A.J. Chapman (editors)

THE AUDITORY IMAGE: A METAPHOR FOR MUSICAL AND PSYCHOLOGICAL RESEARCH ON AUDITORY ORGANIZATION

Stephen McAdams

Institut de Recherche et Coordination Acoustique/Musique

INTRODUCTION

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 pounding the street with metal and your ears with painful pressure waves. Do we merely hear these sources as a collection of 'sound events' (p. 50)? Or do we hear each of these complex sound constellations as an 'object'? I would opt for the latter claim. I do not 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 behaviour 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 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 the 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 towards this parsing behaviour that we have 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 coloured 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 the same orchestra being 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 behaviours.

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 function 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, for example, rapid staccoto 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 volitional act of perceptual organization that may take place within each listener, one might ask the following questions: (1) What might possibly be paid attention to 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?; and (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?

Several concepts have been introduced in these opening paragraphs which need to be explicated further, such as the formation and distinction of auditory source images (objects), simultaneous and sequential auditory organization, and attentional processes. I discuss these concepts further in an attempt to clarify what we have to work with in approaching answers to or rephrasings of the questions posed.

THE METAPHOR: AUDITORY IMAGE

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 to me in describing the results of auditory organizational processes to composers, musicians and psychologists. In particular, and directed toward my main 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. While my own work to date has been limited to the study of images deriving from sound stimulation, 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 a coherence in its acoustic behaviour. The notion of coherence is necessary, if rather 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, all of the physical sources listed in the first paragraph were quite complex acoustically and some even involved multiple sources of sound. But each of these could be perceived as a whole, as a single image. Certainly we could listen only to the jangling medallions or the clicking nails of the right forefoot. But the temporal nature of the pattern as a whole is what gives us the coherent auditory image of a domesticated trotting dog.

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-levelled approach to auditory organization. We can consider a single trumpet tone as an image and speak of its properties as a tone, for example, 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 as a grouping and parsing of sound events into multi-tiered musical images.

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

SEQUENTIAL ORGANIZATION

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

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

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

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

1. It is possible to focus one's attention on a give stream and follow it through time; this means that a stream, by definition, exhibits temporal coherence.

2. The parsing of a sequence into smaller streams takes a certain amount of time to occur; it generally takes several notes into a compound melody line until the separate registers are relegated to different streams. It appears that the perceptual organizing processes assume things are coming from one source until they accumulate enough information to suggest a different interpretation of how the world is behaving.

3. It is easily possible to order the events of a stream in time, but it is more difficult to determine the relative order of events across streams. Two streams resulting from the same sequence of notes appear to overlap in time, but it is hard to say exactly how they are related temporally. Since temporal ordering of notes is an essential determinant of a melody, this means that a melody is, by definition, a stream, that is a melody has a perceptual unity (temporal coherence). This also implies that not just any arbitrary sequence of tones constitutes a melody; if the sequence is not temporally coherent, it is not heard as a melody (but maybe as two or more melodies).

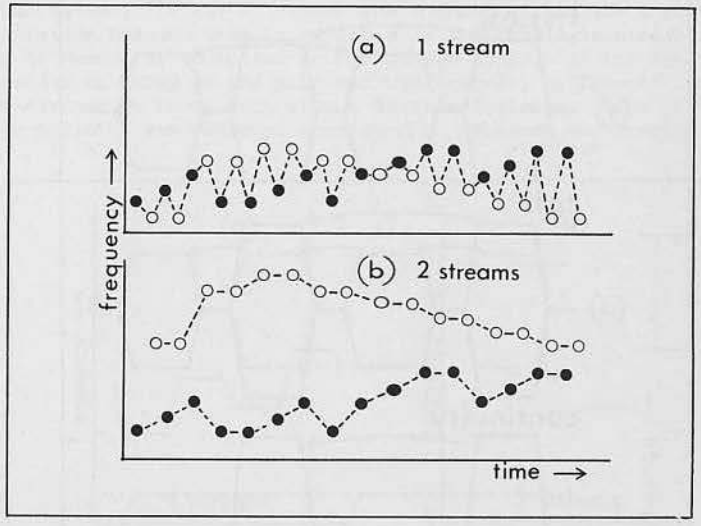
4. A given event can be a member of one or of another concurrent stream but not both simultaneously. As mentioned above, one might switch between hearing an event, belonging to one organization and then to another. The important point here is that several parsing schemes cannot be used at the same time.

The main acoustic factors which have been found to be used by the perceptual systems to build descriptions of streams include frequency, rate of occurrence of events (tempo), amplitude, and spectral content and form, that is the frequencies present in a complex tone and their respective amplitudes. It is not possible to go into great detail about all of these factors. Simple illustrations are given here and the reader is referred to review articles (Bregman, 1978, 1982; McAdams and Bregman, 1979).

Frequency Separation

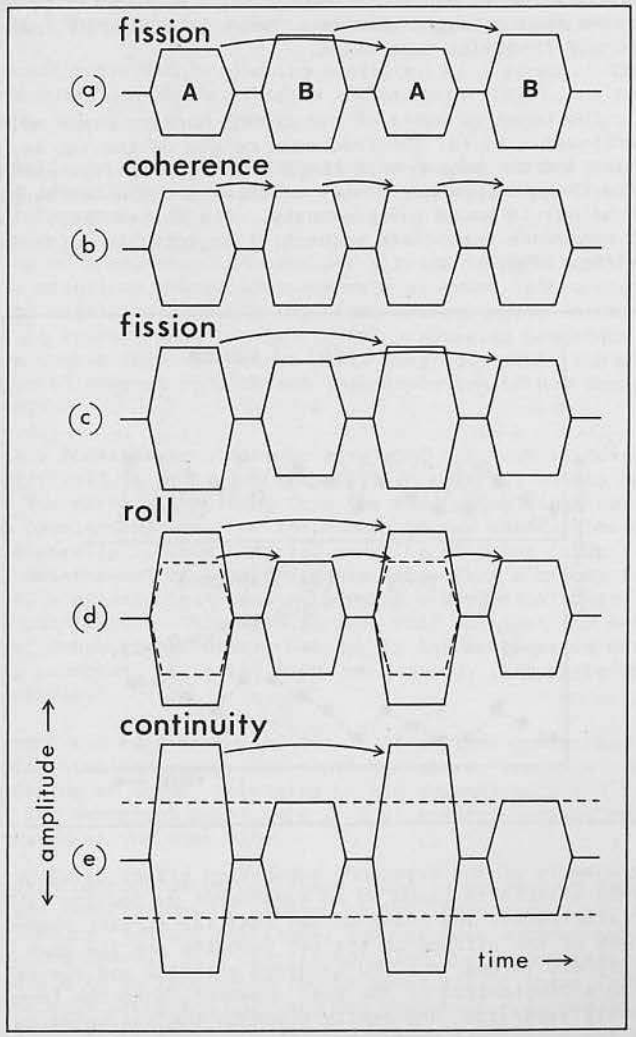
It has been shown repeatedly with sine tone sequences that the relative frequency separation between tones influences the formation of stream organization (Bregman, 1981; Bregman and Campbell, 1971; Bozzi and Vicario, 1960; Dannenbring and Bregman, 1976; Deutsch, 1975; Van Noorden, 1975, 1977; Vicario, 1965, 1982). At a given tempo, tones that are farther apart in frequency are more likely to be heard in separate streams than those that are closer together. Also, at a given frequency separation the role of tempo is such that faster sequences have more of a tendency to split into multiple streams than slower sequences. There is a kind of trade-off between tempo and frequency separation.

FIGURE 1 The tones of parts of two common nursery rhyme melodies are interleaved. In (a) the frequency ranges of the two melodies are similar and the sequence is heard as one, unfamiliar melody. In (b) the frequency ranges of the melodies are non-overlapping and each melody is heard independently. The dotted lines indicate temporal coherence (perceived sequential organization). (derived from Dowling, 1973)



A compelling example of the frequency separation effect is illustrated in Figure 1. This example is based on an experiment by Dowling (1973) where he interleaved (alternated) the notes of two familiar nursery rhyme melodies. When the ranges of the pitches of the two melodies are the same (see Figure 1a) it is difficult to hear out the separate melodies and one melody is heard which is a combination of the two. However, when the frequency ranges are sufficiently separated, one easily discerns the two melodies as is indicated in Figure 1b. In this and similar succeeding figures the dashed lines between tones indicate temporal coherence. For example, in Figure 1b the third tone is perceived as following the first tone rather than the second tone. In Taped Example 1[1], the two melodies are played at four

FIGURE 2 Illustration of the different percepts resulting from the alternation of two sinusoidal tones of identical frequency and duration, when the amplitude of tone A is varied relative to that of tone B. The shapes in the figure represent the amplitude envelopes of the tones. (from Van Noorden, 1975)

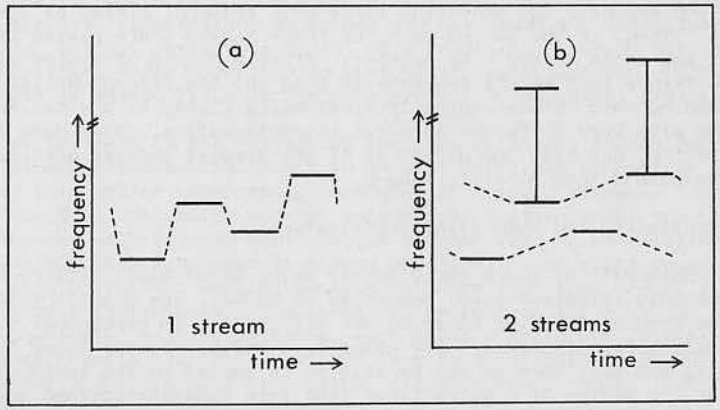


different separation values; the first and last are as shown in Figure 1. Here the identification of the melody as a whole entity is dependent on being able to separate its elements from the other melody and hearing them as a group. The streams that are formed on the basis of frequency separation are the melodies.

Amplitude Differences

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

FIGURE 3 Effect of differences in spectral composition on sequential organization. In (a) all tones are sinusoidal and the frequency separation between them is adjusted so that a single stream percept may be heard, as indicated by the dotted lines. In (b) the third harmonic is added to one pair and the spectral difference causes two streams to form, each with a different timbre. Each of these 4-tone patterns was recycled continually. (McAdams and Bregman, 1979)

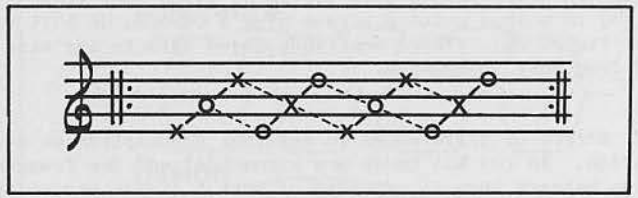


Spectral Form and Content

The last factor to be discussed that contributes to stream formation is spectral form and content. A stimulus sequence can be constructed where the spectral composition is very similar (all tones are sinusoidal, as in Figure 3a) and the frequency variation from tone to tone is small enough so that the sequence can be heard as one stream. By adding a harmonic (the third, in this case) to certain tones in this sequence, those tones are made to form a separate stream (Figure 3b). The solid vertical bar denotes the

fusion of the spectral components into one percept. In Taped Example 2, you may hear first the stimulus cycle in Figure 3a and then the cycle in Figure 3b. Note also that the tempo of the new streams is half that of the original stream. This illustrates that perceived rhythm is also dependent on the stream organization, that is rhythm may be considered as a quality of a given stream.

FIGURE 4 When all of these tones are played by the same instrument ascending pitch triplets are heard (solid lines). But when two instruments with different spectral forms each play the X's and O's respectively, descending triplets are heard (dotted lines).(from Wessel, 1979)



Another example of stream formation based on spectral form is illustrated in Figure 4 which is taken from Wessel (1979). In the first part of Taped Example 3 you may hear the ascending three-note sequence played by one instrument. Then, in the second part,the tones marked X are played by one instrument and those marked O by another. After a couple of cycles of the three-note figure (and as the sequence is sped up) the percept splits into two overlapping sets of descending triplets being played by the two separate instruments with very different spectral characteristics. Note here that not only the rhythm, but also the direction of the triplet changes when the sequence is parsed into multiple images.

Spectral Continuity and Sequential Organization

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

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

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

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

SIMULTANEOUS ORGANIZATION

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

My recent work has been concerned with determining the processes that contribute to the formation and distinction of concurrent source images. And of particular musical interest, the relation between these processes and the derivation of the perceptual qualities of sources. At least four classes of acoustic cues may be shown to contribute to auditory image formation: (1) coherence of amplitude modulation across a sub-group of spectral components belonging to the same source; (2) coherence of frequency modulation across a spectral sub-group; (3) stable resonance structure forming the amplitudes of a spectral sub-group; and (4) localization in space of a spectral sub-group. The first three have been shown to contribute to 'spectral fusion', that is the perceptual fusion of spectral components into a unified percept or source image.

Amplitude Modulation Coherence

By amplitude modulation, I mean the low-frequency modulations we consider as the amplitude envelope (attack and decay functions, and fluctuations in the intensity) of natural sounds. In a musical situation, this would also include tremolo. Two aspects of the coherence of amplitude behaviour of spectral components are important for grouping decisions: onset synchrony of spectral components and amplitude fluctuations across these components during a sustained tone.

Onset Synchrony. It has been demonstrated several times that when the onsets of the partials of a tone complex are asynchronous by as little as 20-30 msec, the perceived fusion decreases and the ability to hear out individual partials increases (Bregman and Pinker, 1978; Dannenbring and Bregman, 1976; Rasch, 1978, 1979). The minor asynchronies observed in the partials of musical instrument tones are generally less than 20 msec (Grey and Moorer, 1977). Helmholtz (1885, 1954) reminded us in the last century that with natural tones all of the partial tones tend to start together, swell uniformly and cease simultaneously. 'Hence no opportunity is generally given for hearing them separately and independently' (p. 60).

With computer synthesis techniques, one has easy control over the relative onsets of individual partials. In Taped Example 6 you may hear a series of tones which are identical except for the synchrony of onset of the partials. The extremes of the series are represented schematically in Figure 5. Beginning with perfect synchrony the exponential envelope of the inharmonic partials of this bell-like tone are progressively desynchronized until you can hear each partial separately. Note that the change is one from a fused rich sound to a more chord-like quality, even though the frequency relations are identical. Here, the coherence of the amplitude behaviour is progressively destroyed which results in the destruction of the image's unity (or, alternatively, in the creation of a multiple image).

Kubovy and Jordan (1979) produced a similar kind of effect by suddenly changing the phase relation of one harmonic of a 12-harmonic tone relative to the phases of the rest of the harmonies. In this case the single harmonic is separately audible as a pure tone if the phase difference is at least 30 degrees. It seems likely that a sudden phase shift like this might be interpreted as an onset asynchrony by the auditory system, thus causing the single harmonic to be heard as a separate source.

Amplitude Fluctuations. Increased fusion can also be obtained even in inharmonic tone complexes by imposing a common amplitude modulation on the components. Von Békésy (1960) reported getting fused images lateralized to

one side of the head when he presented sinusoids of differing frequencies (one to each ear over headphones) and imposed an identical low-frequency (5-50 Hz) sinusoidal amplitude modulation (100% modulation depth) on each tone component.

FIGURE 5 The role of onset synchrony in spectral fusion. When the amplitude envelopes of all spectral components of this inharmonic sound start synchronously (a), a single, fused, bell-like image occurs. When the onsets of these partials are spread out in time (b), each partial is heard as a separate source image. (from McAdams and Wessel, 1981)

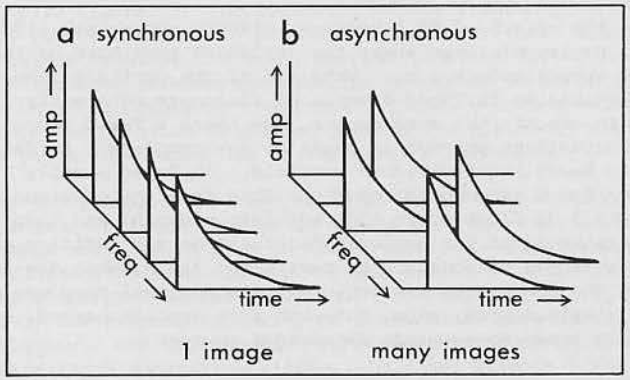
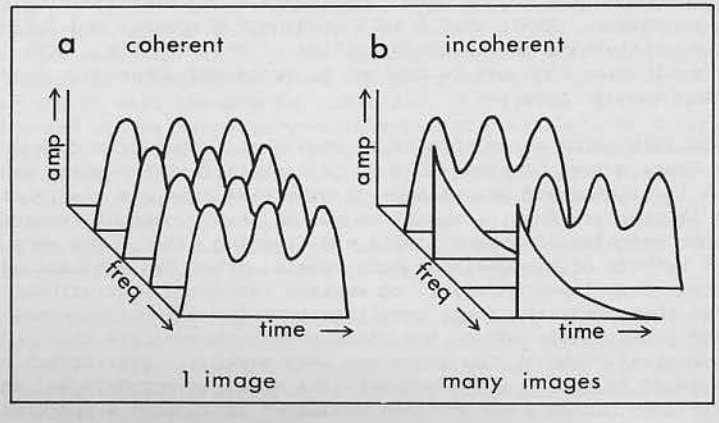


FIGURE 6 The role of amplitude modulation coherence in spectral fusion. When the amplitudes of all partials fluctuate together (a) a fused image is heard, but when the onsets are asynchronous and the amplitude fluctuations are incoherent, several images are heard. (from McAdams and Wessel, 1981)



As far as I am aware no systematic study has been done on the necessary modulation depth or the limitations of the modulation frequencies which can cause tone complexes to fuse. Bregman, Abramson and Darwin (1983) performed an experiment in which a pure tone alternated with a two-tone complex (a stimulus configuration similar to that shown in Figure 6). The pure tone and the upper of the two tones in the complex were amplitude modulated at 100 Hz. The modulation rate of the lower of the two tones varied between 95 and 105 Hz. The best fusion of the tone complex occurred when the modulation rates were the same and the modulation waveforms were in phase. The frequency relation between the tones of the complex did not affect the fusion. Only the coherence of the amplitude modulation had an effect on the fusion. This same result was reported by von Békésy and implies that it is the coherent fluctuation that is the unifying criterion rather than the harmonicity of the side bands created by the modulation.

To illustrate the effect of AM coherence, McAdams and Wessel (1981) synthesized a series of tones where the amplitude envelopes of the partials fluctuated at approximately 3 Hz. When all of the partials (the same inharmonic partials as in Taped Example 6) fluctuate coherently (see Figure 6a), that is in exactly the same manner, one hears a fused tone. However, when these fluctuations are not in phase or are completely different (see Figure 6b) one hears the individual partials. In Taped Example 7 you may hear again the fused exponential envelope tone from the previous example, then a coherent 3 Hz fluctuation (50% modulation depth) and then different mixtures of dephasing of the periodic fluctuations and addition of the exponential envelopes on some of the partials. Again, when the amplitude behaviours of the components are coherent, the spectral components are fused into a single source image, but when they are incoherent, or unrelated, they tend to be heard as separate sinusoidal sources.

Frequency Modulation Coherence

The types of frequency modulation I am referring to here include musical vibrato (periodic modulation), jitter (aperiodic modulation)[2], and slow pitch glides such as in voice inflections or musical portamento. In all natural, sustaining vibration sources, any perturbation, periodic or otherwise, of the fundamental frequency is imparted proportionally to all of the harmonies. There are, of course, minor departures due to various non-linearities in such acoustic systems, but in general as the fundamental frequency changes, all of the harmonics change with it maintaining their harmonic relations. Thus, what I call 'coherent frequency modulation' is modulation maintaining the frequency ratios of the partials. With the computer synthesis, this can be applied to sustained inharmonic tones as well as to harmonic tones.

It is difficult to do an experiment to show that frequency modulation actually fuses a tone complex. But it is certain that for music synthesis it adds a liveliness and naturalness to otherwise dead and electronic sounding images. In fact, I was first put on this course of research by hearing the examples of McNabb (1981) and Chowning (1980) who were trying different methods of synthesizing sung vowels. They had synthesized all of the spectral form (speech format) parameters correctly, but without modulation the tones were still unsatisfactory for musical purposes. When jitter and vibrato were added, the vocal sounds became life-like and imbued with the musical richness the composers were seeking. This effect can be heard in Taped Example 8 where a vowel-like sound is unmodulated, then modulated, then unmodulated and then modulated again with a spectral change

to a different vowel. Note that the natural voice quality goes away, and one can actually hear out individual harmonics, when the modulation is not shown.

It can be seen that if a frequency modulation (vibrato or jitter) is imposed on the partials of a harmonic tone complex such that the ratios are not maintained, the complex 'defuses'. In one experiment (McAdams, 1983), I asked the question whether coherence could be had simply by moving all the harmonics in the same direction at the same time or whether it was really necessary to maintain the frequency ratios. Subjects were asked to compare among harmonic complex tones with different modulation schemes. One tone had a modulation that maintained constant frequency ratios among the 16 harmonics. The other tone had a modulation that maintained constant frequency differences among the harmonics. These stimuli are illustrated schematically in Figure 7. The logarithm of the frequency variation is plotted as a function of time. Note that on a logarithmic scale, constant ratios maintain a constant distance, whereas constant differences do not. We know that the basilar membrane resolves frequency components in the inner ear roughly on a log frequency continuum. Thus a constant ratio modulation would maintain the relative distances between the places of maximum stimulation on the membrance due to the various harmonics.

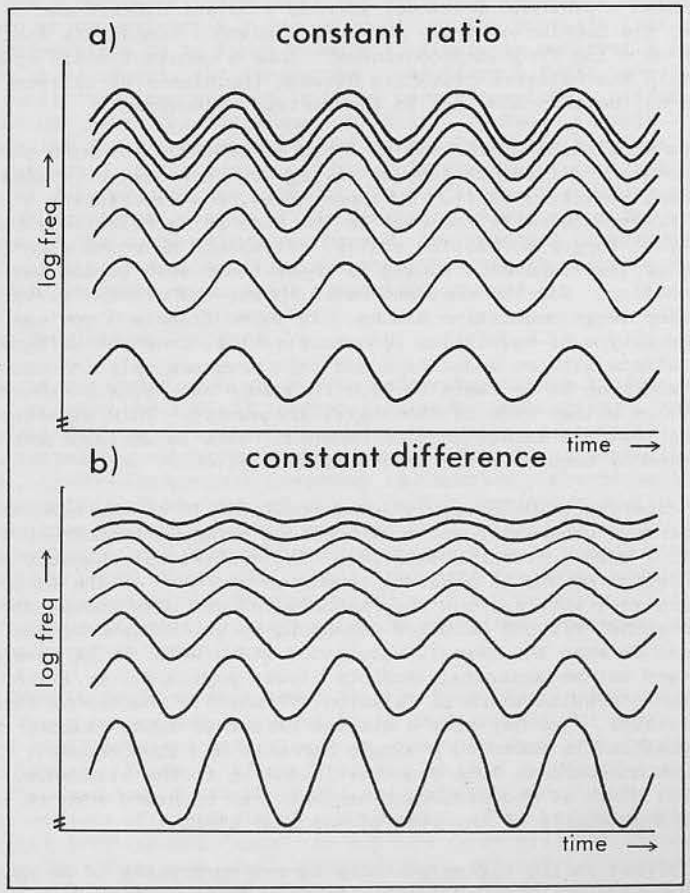
When the rms modulation width[3] was at least 12 cents[4], listeners more often chose the constant difference tone as having more sources, or images, or distinguishable entities in it. In these tones, one experiences a modulating fundamental with the rest of the tone being relatively stationary particularly at larger modulation widths. It should be noted also that the frequencies of the components making up these tones move in and out of a harmonic relation. For the constant ratio tones, the percept is very unified even at rather large modulation widths. In Taped Example 9 you may hear one series of each type of modulation (constant ratios, constant differences). The series starts with no modulation and then progressively increases the modulation width up to 56 cents (3.3%,a frequency excursion of about a quarter tone on either side of the centre frequency). This experiment demonstrates that the maintenance of constant ratio is an important part of the definition of coherence for frequency modulation.

In another experiment (McAdams, 1983c) I modulated 15 of the harmonics of a 16-component tone coherently and modulated one harmonic incoherently. In these tones, I used a jitter modulation. The statistical characteristics of the modulation on the 15 coherent harmonics and that on the incoherent harmonic were very similar, but the random waveforms were independent. Several perceptual effects resulted depending on which harmonic was modulated and on what the overall modulation width was. Either certain partials stand out as separately audible (lower partials), or a kind of 'choral effect' results where an illusion of multiple sources is heard (higher partials). You may hear a similar effect in Taped Example 10. The vibrato modulation is added to a single harmonic and then removed. This is done for each harmonic in turn from the lowest up to the sixteenth. Note that even the pitch of the sixteenth harmonic can be heard when it is modulated independently of the rest of the tone complex.

The choral effect in the higher partials is not surprising if we stop to imagine the behaviour of several instruments playing sustained tones simultaneously: five violins for example. Each acoustic source has its own independent jitter modulating all of its harmonics. When we add all of

the sources together we get these random movements of the frequencies beating against one another creating quite a complex situation acoustically. In addition, as one moves into the higher harmonics, the patterns of stimulation on the basilar membrane move closer and closer together until they are heavily overlapping.

FIGURE 7 The role of constant ratio frequency modulation in spectral fusion. A spectrographic diagram of constant frequency ratio and constant frequency difference modulations is plotted on a log frequency scale. In (a) a fused, modulating image is heard. In (b) the lowest frequency separates perceptually from the rest, which are perceived as barely modulating.



In these regions, the incoherent movement of adjacent harmonics is creating a complex stimulation for any given auditory nerve. You can imagine that if enough of these violins are playing the same pitch, there is a limit to how many sources you can pick out. The difference between 15 and 16 violins is very small indeed and after about 8 to 10, we generally just hear 'many'.

I have reported previously (McAdams, 1980, 1981, 1982a, 1982b, 1983a, 1983c) that FM serves not only to group simultaneous components into a source image, but serves as a cue to distinguish concurrent sources as well. The presence of independent modulation patterns on separate sub-groups gives two types of cues for the presence of multiple sources: (1) adjacent partials belonging to separate sources are incoherently modulating with respect to one another; and (2) the modulation across the partials belonging to a single source is coherent. It seems likely to me that the auditory system makes use both of the local incoherence between partials to detect the presence of multiple sources and global coherence among partials to accumulate the appropriate spectral components into a source image. This may be one reason soloists, particularly opera singers, use vibrato to the extent they do, that is to separate themselves from the rest of the ensemble. Of course, as mentioned before, if things are too crowded temporally and spectrally the system may have trouble distinguishing individual source images and tracking their behaviour. This would be due to the limitations of spectral and temporal resolution in the auditory system.

Another cue related to frequency that interacts to a certain extent with modulation coherence is the harmonicity of the frequency components. This is particularly evident with sustained sounds. A harmonic series, in most contexts, gives an unambiguous pitch sensation. Sustained inharmonic sounds tend to elicit a perception of multiple pitches. In many cases, the perception of multiple pitches can be interpreted as the presence of multiple sources (Cutting, 1976; Scheffers, 1983).

In a laboratory situation, unmodulated, sustained harmonic sounds can be perceptually analysed into their harmonics for harmonic numbers up to the fifth, sixth or seventh depending on the fundamental frequency. This means listeners can reliably hear out individual harmonics and identify their pitches. But a pilot study I have recently performed suggests that when these tones are modulated, listeners are no longer able to hear out the harmonics, indicating that the image is fused and unanalysable perceptually. In this case the only pitch heard is the pitch of the fundamental. This may be interpreted as support for the notion that the grouping processes (including *spectra fusion*) influence our perception of the qualities of source images (including pitch). There remains, however, the possibility that pitch detection also influences source image formation under certain conditions. It is still unclear at this point whether it is the presence of a number of pitches that indicates the number of sources, or whether the presence of *multiple harmonic series* indicate multiple sources and give rise to multiple pitches. I am more inclined toward the latter interpretation given the preliminary result of reduction of perceptual analysability of harmonic complexes in the presence of frequency modulation.

Spectral Form

Most sustaining musical sound sources have resonance structures that are relatively stable, or very slowly changing, compared to the frequency fluctuations mentioned in the previous section. These structures are due to resonant cavities that filter the sound before it radiates into the air, for

example, vocal cavities, body resonance for string instruments and tube resonance for winds. Each resonance has a particular frequency to which it responds the greatest. The frequency is related to the volume of the cavity, and the size of the opening. Other frequencies are attenuated (made less intense), or are not passed as easily, relative to this resonant frequency. Also, different shapes and the nature of the walls of the resonant cavities influence which frequencies near the resonant frequency are allowed to pass. When it allows a larger number of frequencies around the resonant frequency to pass we say it has a larger bandwith.

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

Now let us imagine what happens when a singer sings with vibrato. All of the frequencies are moving back and forth in a coherent manner. And what happens to their relative amplitudes? Well, since the resonances come after the point in the system where the vibrato is introduced, the amplitudes must follow the contour of the formant structure. This is illustrated schematically in Figure 8. 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 while the excursions 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. In a sense, we can consider that as the frequencies modulate, their amplitudes change such that each partial *traces* a small portion of the *spectral envelope*, that is, the frequency-amplitude curve describing the overall spectral form. This complex coupling of frequency and amplitude modulation serves to define the spectral contour and in certain cases may actually reduce the ambiguity of the resonant identify of the sound source.

In Figure 9 is another spectral form. 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 well defined and accordingly, the perception of the vowel sound would be weak if at all existent. However, when the spectral components are made to modulate in frequency, 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. This has been verified experimentally, particularly for higher fundamentals where definition of spectral form is lacking (McAdams, 1982a, 1983c).

Another experiment has shown that if the spectral envelope with the frequency modulation, that is, the amplitudes of the components remain constant, a kind of timbral modulation occurs. With vowel envelopes one hears a whistling sound which seems associated with the perceptual decomposition of the higher formants. This occurs for modulation widths in excess of 1/8 to 1/4 tone

(25 cents to 50 cents). This result has some interesting musical possibilities for sound synthesis methods based on formant structures. This is discussed later.

But now let us imagine the following perception problem. The ear receives a complex spectrum as shown in Figure 10. There is no modulation on any of the components. Listeners sometimes report hearing certain vowels embedded in this complex and report as many as 6 - 8 different pitches. You may hear six such configurations in Taped Example 11. Without some cue to help us group the elements of this complex, it sounds like a tone mass. If, however, we add some frequency modulation coupled to the resonance structures the elements are grouped perceptually. We hear the images more clearly as being a certain vowel at a certain pitch. What the auditory system may then have access to spectrally is represented in Figure 11. It now knows that there are three sources each with different vowel quality and pitch. The stimuli actually presented are notated in Figure 12. You may hear these same configurations of three vowels at three pitches, but with vibrato this time, in Taped Example 12.

FIGURE 8 The dotted line represents the spectral form created by a 3-formant resonance structure. 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 the formant shapes and in the case of f_1 and f_2 these shapes 'point' toward the formant peak.

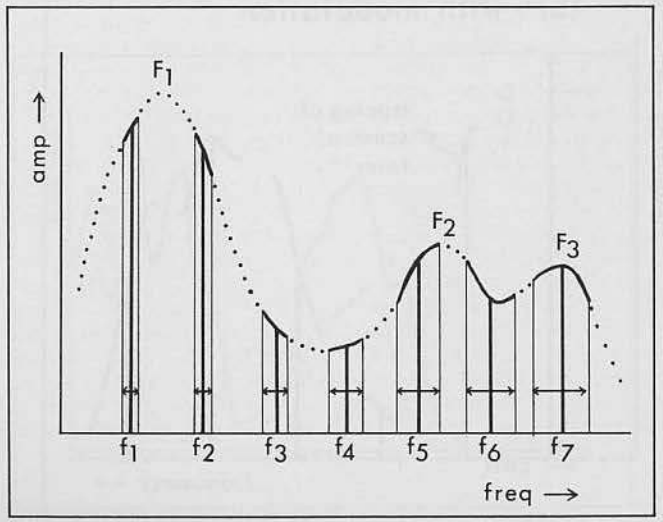


FIGURE 9 The vowel /a/ is plotted with a high fundamental frequency where there are few harmonics present. Without modulation (a), the inferred spectral form (dashed line) would be very different from the actual spectral form (dotted line). With modulation (b), the spectral slopes give a much clearer indication of the spectral form.

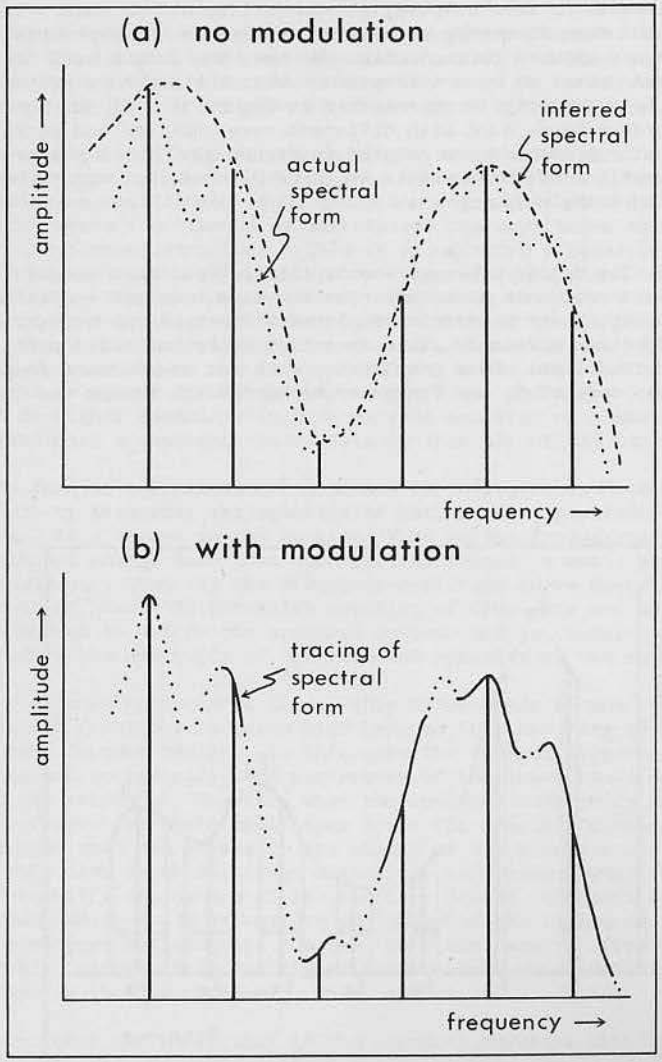


FIGURE 10 Complex spectrum resulting from several unmodulated sustaining harmonic sources.

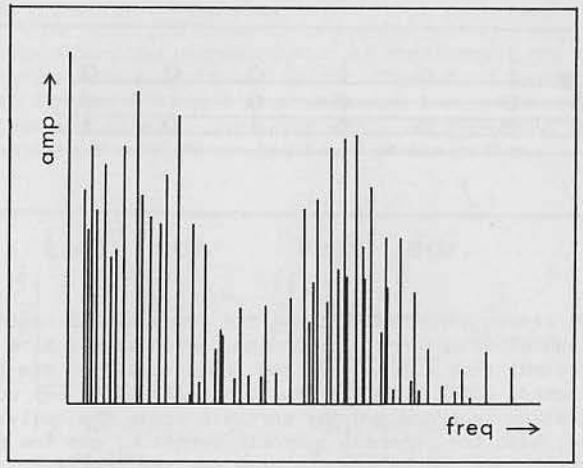


FIGURE 11 Spectral forms extracted by the auditory system when the individual sources are modulated, thereby increasing the information pertaining to the number and nature of the sources.

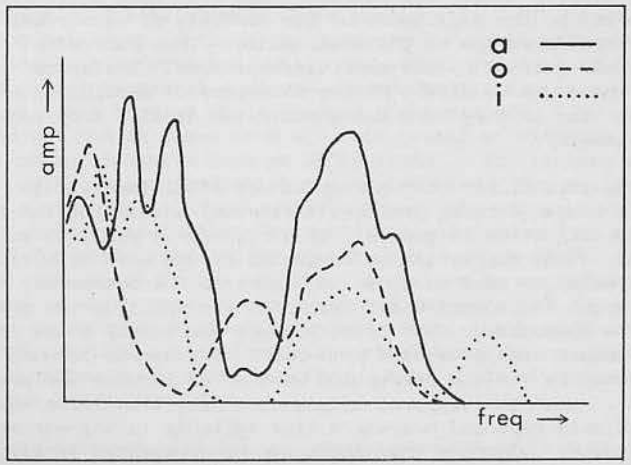
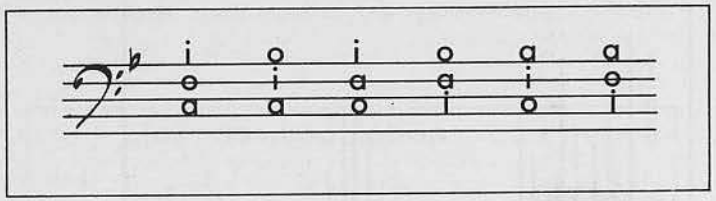


FIGURE 12 Permutations of three vowels at three different pitches. These chords may be heard in Taped Examples 11 and 12.



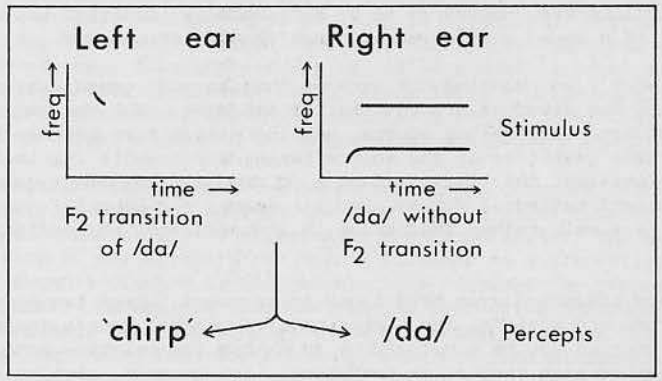
It is important to remark here that without the grouping information to select a certain subset of spectral components, one does not have access to the particular spectral form which gives the vowel quality. The overall spectral form is heard, which does not really correspond to any vowel[6]. But when the modulation is added and the partials trace the individual spectral envelopes, both the coherent harmonic behaviour and the reduced ambiguity of spectral form can be used to hear out the vowels.

In an experiment using stimuli similar to these, listeners judged the vowels to be more prominent and the pitches less ambiguous when the 'sources' were modulating. But there was also a surprising result. In some conditions all three vowels at their respective pitches were modulated coherently, maintaining the exact routes between all harmonics of all vowels. Given the putative criterion of frequency modulation coherence for grouping, I expected it to be more difficult to hear out the vowels in this situation. However, there was no difference between modulating the vowels coherently or incoherently between them. But remember that even when their frequencies are moving coherently, the amplitudes of the partials of each vowel are tracing the spectral envelope of the vowel alone. Thus each vowel is still being unambiguously defined by the amplitude movement. Listeners' results indicated that there was no effect of the coherence of modulation of the three vowels. As long as they were being modulated at all, they were more prominent perceptually.

This suggests the possibility that perception of vowel identity is independent of a source forming process, that vowel identification (or speech sound identification in general) is performed in parallel with source image processing. This suggestion is supported by the work of Cutting (1976). The stimulus he used is shown in Figure 13. A schematic, two-formant speech sound (consonant-vowel syllable) was split up and presented dichotically over headphones. The first formant and steady state portion of the second formant were presented to one ear. The second formant transition responsible for the perception of the /d/ phenome was presented to the other ear. When the temporal alignment of the transition sound was appropriate, subjects reported hearing a /da/ syllable in the ear with the steady state portion. When the transition was not presented at all, subjects reported hearing a /ba/ most of the time. So the transition in the opposite ear was contributing to the percept of the /da/ syllable, the whole of which was located in one ear. However, many subjects also reported

hearing a chirp sound (the percept elicited by the second formant transition alone) in the opposite ear as well. A two source percept resulted where one element of acoustic information was contributing to the qualities of both sources simultaneously.

FIGURE 13 Schematic representation of the formant trajectories in the stimulus used by Cutting (1976). The second formant transition for the /da/ syllable was presented to one ear and the rest of the signal was presented to the other ear. The resulting percepts and their perceived located are indicated at the bottom.



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

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

Spatial Location

Sounds in the environment stimulate two sense organs, our two ears. Except for sound sources lying in the median place, equidistant from the two ears, the signals received at the two eardrums are slightly different. When a source lies to the right side of the head, the sound reaches the right ear first. It is also more intense in the right ear at higher frequencies since the head casts a sound shadow which attenuates the sound before it arrives at the left ear. Also, the two pinnae (the outer ears) modify the sound significantly depending on the azimuth and elevation of the sound source relative to them. These modifications, that is, the introduction of certain reflections due to the structure of the outer ear, would again be slightly different for each ear when the sound was not in the median plane. The pinna modifications are considered to be particularly important for detecting the elevation of a sound and for making back/front distinctions.

At any rate, what I am pointing out here is that in most cases, the sounds arriving at the two eardrums are not exactly the same. But the two sources of stimulation are heard as one source, and the disparities between the ears influence certain qualities of the source image, for example its location in azimuth and elevation, the characteristics of the acoustic environment (large reverberant cathedral *versus* open air space, for example; and I am sitting next to a wall rather than being in the middle of the auditorium,) etc.

When sounds are presented over headphones or speakers, where two or more physical sources are present, the placement of a virtual source image between the physical sources can be simulated by adjusting the relative intensities and onset times at each speaker or headphone. For example, over headphones (where the sound is usually heard inside the head but can be moved from side to side) a time onset difference of only 0.6 msec is sufficient to make the sound appear to come entirely from the leading earphone (cf. Mills, 1972). If the onset difference is increased to a few msec one begins to hear two separate events (cf. Schubert, 1979; editors's comments pp 255-257). With speakers in a moderately reverberant room, the difference must be in the order of 30 msec for separate events to be perceived. Thus, in the temporal domain, there are some rather narrow limits to the fusion of separate events into a single image. Also, if the spectral and temporal characteristics of the events are quite different, the disparities necessary to hear the events as separate are smaller.

Often, in room listening, echoes (which are repeated versions of the same sound, but coming from different directions and transformed by the room acoustics) do not affect the apparent location of the sound source. They are ignored in this respect by the auditory system, but contribute instead to the perception of the acoustic properties of the environment. This is called the 'precedence' effect to indicate that the event which precedes its replications (echoes) is used to determine the nature and location of the source, and the replications that follow are perceptually fused with the direct sound.

Now let us consider the extent to which spatial location helps us to distinguish among sound sources. Cherry (1953) studied the ability of listeners to attend to and extract meaningful information from a speech stream embedded in a noisy background (a cocktail party) of many other speakers (people speaking loudly, not loudspeakers). When at the party itself. we can use the location of the source, speech characteristics of the

voice being listened to, semantic constraints and additional visual information, such as lip reading, hand gestures and the like, to reconstruct information masked or covered up by the noisy environment. When we listen to the same person speaking in the same environment but recorded on a stereo tape recorder (in order to preserve the directionality of the various sources) intelligibility decreases somewhat but a significant amount can still be distinguished. However, when only one channel of sound is presented, intelligibility of the target speech stream is reduced drastically.

This suggests that the sounds that mask the target signal in the monoaural case are not as effective as maskers when the auditory system can relegate them to separate spatial locations. Thus, spatial location is a cue that can be used to attend to and follow the emanations of a given sound source. There is a well-developed area of study in 'classical' psychoacoustics addressing the improvement in detection that occurs when two ears are used instead of one. This improvement is called a masking level difference (MLD cf. Durlach, 1972; Jeffress, 1972), since it is a difference between monaural and binaural listening conditions in the level that is necessary to hear a target signal in the presence of a masking signal.

There arise certain situations in music where a composer desires a great complexity of material and, at the same time, a great clarity or distinguishability of the elements in that complexity. The spatial separation of key elements can help a listener to differentiate between them. A good example of this is the recent piece '*Repons* ' by Pierre Boulez (1981/1982). He has six soloists playing percussive instruments (for example piano, xylophone, harp) which are arranged around the perimeter of a rectangular hall. Each soloist's sound is modified electronically and sent to six speakers also in the perimeter of the hall. A small orchestra is placed in the centre of the hall. The audience is distributed around the orchestra and thus positioned between the acoustic orchestra and the electronically modified soloists. At one point near the beginning of the piece the transformed sounds from each soloist are echoed many times and sent bouncing around to the speakers. What results is a marvellously rich, crystalline texture of timbral arpeggios that nonetheless has a kind of clarity due to the timbral elements being distributed in space. After hearing the first performance of this work, I then heard a stereophonic recording of the same concert. The carefully woven, multi-mirrored reflections and the intricacy of the electronic modifications of the soloists' notes were reduced to a dull, ambient mush. The spatialization had been an essential compositional element in the hearing of the piece. This was something akin to seeing a photograph of a Monet painting of the water-lilies at Giverny in low-contrast black and white. In the case of '*Repons*', two electronic 'ears' were not enough to preserve the clarity because in a concert hall the human ears are attached to a movable globe and the act of moving itself helps to distinguish different sources by creating a continually changing disparity between the ears which is unique for each source of sound. This changing disparity is a very strong cue for the invariance in location of the source.

Coherent Behaviour of Sound Objects and Simultaneous Organization

All of the factors discussed above contribute to a general coherence of the elements belonging to a physical sound source. And this coherence may be considered in turn as a by-product of the behaviour of the physical system producing the sound. I propose that much of the organization our perceptual

systems perform is based on (but by no means limited to) a learning of the normal behaviour of physical objects in the world around us. I would not want to limit the possibilities of perceiving to such object-based learning, but I would suggest that this whole realm of normal perceiving heavily influences our perception of music. Further, I have found, in my own perceptual analyses of several pieces of music, that extensions of these various criteria of 'behavioural coherence' have proven useful in predicting when different kinds of reorganization of the physical objects are possible, for example, the recombination of several instruments into a fused composite timbre, and so on.

To summarize briefly the elements treated in this section, there are (at least) four factors contributing to the fusion and separation of simultaneous source images. These are: (1) a common (or closely correlated) global amplitude modulation (low frequency amplitude fluctuations and amplitude envelope); (2) a common (or closely correlated) frequency modulation which maintains the frequency ratios among the components (periodic: vibrato, aperiodic: jitter, or slow pitch glides: inflection); (3) a complex coupling of amplitude and frequency modulation which defines a spectral envelope, implying a stable resonance structure (such as vowel formants); and (4) a common spatial location (the dynamic maintenance of similar time, intensity and spectral disparities at the two ears for all elements of a source whether the source is moving or the head is moving).

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

For Taped Example 15, the sound of an oboe was analysed by phase vocoder. You would first hear the original oboe sound on the tape. The output of this analysis describes the amplitude and frequency behaviour of each harmonic. From these data the sound can be resynthesized either exactly as analysed or with certain modifications. In this case, the even and odd harmonics are sent to separate channels in order to be played over different loudspeakers. Initially the same vibrato and jitter patterns are imposed on the two groups of sounds. Then, slowly, the frequency modulation pattern on the even harmonics is decorrelated from the pattern on the odd harmonics. This is illustrated in Figure 15.

As you may hear, the initial image of an oboe between the speakers gradually pulls apart into two images and in the two speakers: one of a soprano-like sound an octave higher (the even harmonics) and one of a hollow, almost clarinet-like sound at the original pitch (the odd harmonics). Following this, each channel is played separately and then the two channel version is played again. It is extremely important that the levels of the two channels be properly adjusted for the effect to work. This example was used in a composition by Roger Reynolds '*Archipelago*' (1983) and was realized at IRCAM by Lancino. Here we have a case where the coherence of frequency

modulation at the beginning of the sound overrides the spatial separation of the two subsets of harmonics and one hears a single image of the oboe, more or less localized between the speakers. But as the modulations become incoherent, the images move to their rightful places and the sounds now appear to come from where they were originally coming from.

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

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

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

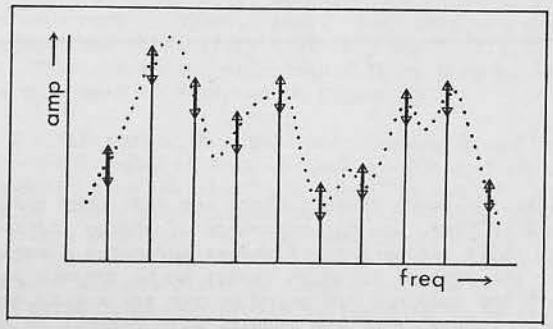
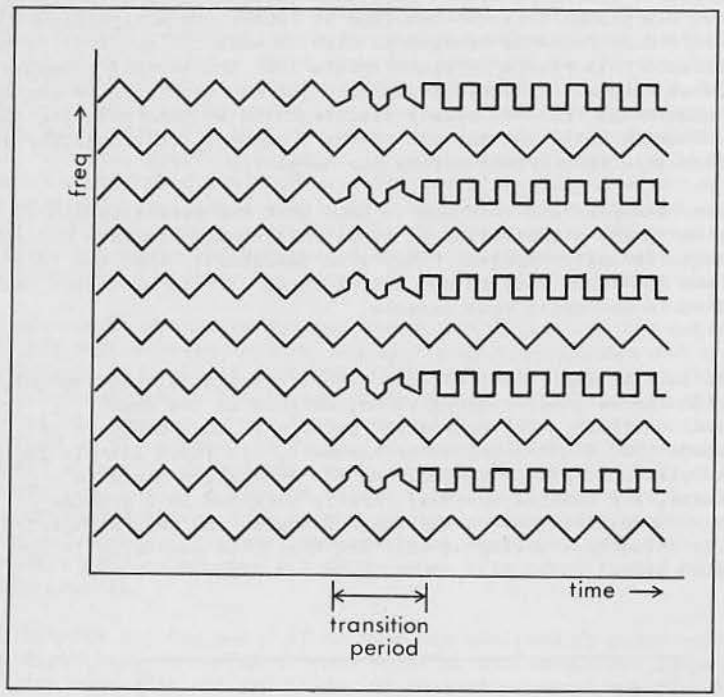


FIGURE 15 The parsing of even and odd harmonies of an oboe sound. Initially, all harmonics are modulated coherently. Then the even harmonics are slowly decorrelated from the odd harmonics until they have an independent modulation pattern. The triangle and square waves are only used for easy visualization.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



The important similarities among these factors are that they are dynamic, that is they change with time, and the coherence of change indicates a common source origin while incoherence of change indicates diverse source origins. I consider the spatial location factor to be dynamic because in normal listening both the head and the sound source are moving to some extent and the dynamic coherence of the result, with respect to the two ears, becomes a relatively unambiguous cue for place of origin of the sound source. Indeed, 'place' becomes an invariant quality of the sound image when this coherence is maintained.

What I think we need as a general and subdividable principle is the notion

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

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

INTERACTIONS BETWEEN SEQUENTIAL AND SIMULTANEOUS ORGANIZATION

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

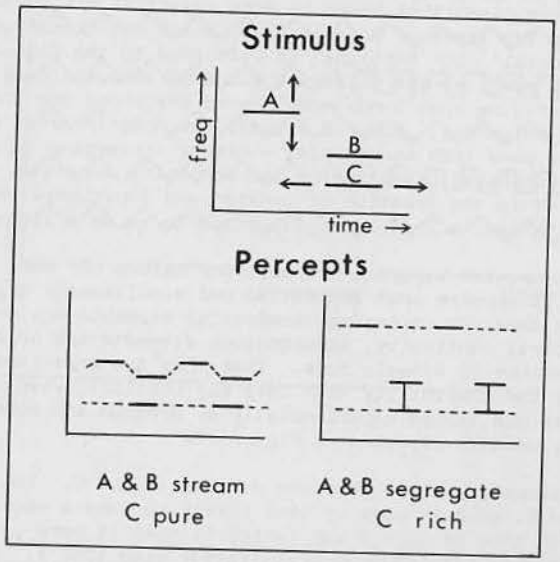
In a sense, there are two separate propositions before our ears in this kind of counterplay. It appears that sequential and simultaneous organization are determined by separate criteria: sequential elements are organized according to spectral continuity, simultaneous elements are organized according to coherence of dynamic cues. That they are organized by separate criteria suggests the possibility that they may conflict, even compete, with one another. This was tested experimentally by Bregman and Pinker (1978). The stimulus they used is depicted in Figure 16.

A pure tone A alternated with a two-tone complex B and C. The frequency of tone A was varied to make it more or less likely to form a sequential stream with B. The onset time of tone C was varied to make it more or less likely to fuse, that is form a simultaneous organization with tone B. Thus, both tones A and C were competing for the membership of tone B. Two judgements were collected from subjects: (1) whether A and B formed one stream or two; and (2) whether C was perceived as being more pure or more rich. When A and B were in close frequency proximity (tending to be perceived as one stream) and B and C were asynchronous (tending to be perceived separately), judgements indicated that A and B more often formed one stream and C was perceived as being more pure. When A and B were distant in frequency (tending to be perceived as two streams) and B and C were synchronous (tending to be perceived as a group), judgements indicated that A was heard in a stream by itself and C was perceived as being richer (being fused with

B). These two cases may be heard in Taped Example 17 in the order described above.

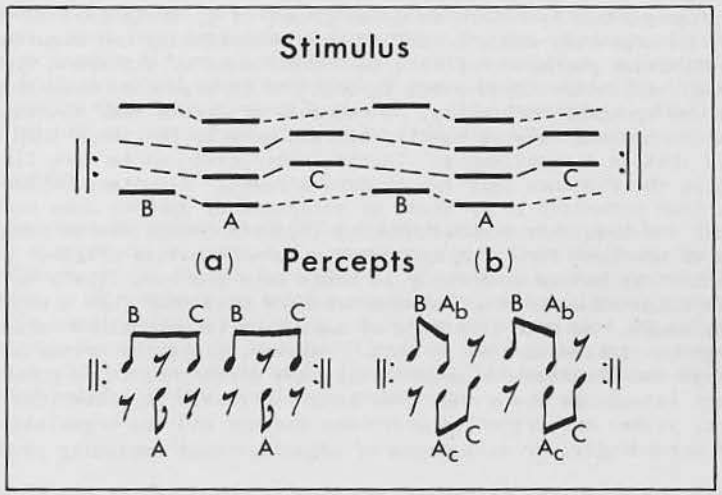
This was an important experiment in two respects. It demonstrated the separate kinds of organization and their interaction in the final perceptual result. And it also demonstrated that a perceived quality of source, for example the timbre of C, was dependent on how the elements were organized. When tone B was not grouped (fused) with C, the latter was more pure. But when they were fused, C was perceived as being rich. Thus timbre, and, as I demonstrated with the oboe in Taped Example 16, pitch are properties of source images that are derived after the concurrent elements have been organized into those images.

FIGURE 16 The stimulus used by Bregman and Pinker (1978) to demonstrate the competition of sequential and simultaneous organizations. Tone B is potentially a member of a sequential organization with A or a simultaneous organization with C. See text for more detailed information.



Another example to demonstrate the organizational dependency of timbre is schematized in Figure 17. This sound configuration is played in two separate contexts which greatly affect the perceived timbre of tone A. In the first part of Taped Example 18, tone A (4 components) is initially played alone. And then tones B and C (two components each) gradually fade in. At first the percept is that shown in Figure 17 as (a). Tones B and C form one stream and tone A forms another. Note the timbre of tone A. But as the intensities of B and C approach that of A, the components of A are captured and pulled into separate organizations by B and C. At this point the original timbre of A is more difficult to hear. It is replaced by a double timbre and the rhythmic pattern is as shown in percept (b).

FIGURE 17 Depending on the intensity of tones B and C and on the context, A may be heard as a fused tone with its own timbre, or it may be split into two simultaneous tones, A_b and A_c as indicated by the dotted lines in the stimulus diagram. In the former case, the resulting rhythm is (a); in the latter case, the rhythm is (b).



But now let us reverse the procedure. The second part of Taped Example 18 starts with tones B and C and the tone A is gradually faded in. This time it never reaches (for my ears anyway) a point where it is pulled apart by tones B and C. Somehow the crescendo movement of A seems to keep it stable as a separate stream and stays at percept (a). So with as simple a sound configuration as this we already have some rather complex effects of context on the way the elements are organized and the timbres are perceived. Demonstrations such as this support the proposition of Bregman (1977, 1980) that perception is an active process of composing the sensory data into some kind of interpretation of the way the world is behaving. The

composition process draws from a large number of elements in the perceptual field that interact in complex ways to produce a final percept (Bregman and Tougas, 1979).

Where this becomes interesting musically is in its implication that, with computer synthesis techniques, processes of horizontal and vertical musical organization can be carried into the sound microstructure. The composer can play with the processes of perceptual organization that underly the heard musical surface. This sets up the possibility of composing situations where sequential and simultaneous organizations compete for individual spectral components to be part of the structure of a musical image. With a careful consideration (or better yet, embodiment and subsequent intuitive use) of these principles of perceptual organization, the composer has access to a whole realm of mutability of the heard 'image'. Convergences and divergences of the musical functionalities of individual elements allow the development of microstructural (and pre-perceptual) ambiguity. These possibilities are evidenced in the last Taped Example (no. 19), created by Rodet with the CHANT computer-synthesis program.

SUMMARY

I have proposed that there are separate groups of criteria that determine the way one organizes acoustic information sequentially and simultaneously. This distinction perhaps reflects the involvement of different types of perceptual mechanisms. Sequential information is organized according to criteria of spectral continuity. A sequence of events that maintains spectral continuity is more easily followed as a source image than a sequence that is discontinuous. In the latter case one is more likely to reorganize the sequence into two or more streams. Simultaneous information is organized according to criteria of coherence of dynamic cues such as amplitude and frequency modulation that indicate common source origin, the tracing of spectral form that indicates a common spatial origin. Spectral components that behave coherently in these ways are more likely to be heard as originating from a common source and will thus form a unified, fused auditory image. As this coherence of behaviour is maintained across time, the image can follow in time as well. However, since the criteria for sequential and simultaneous organization are different, it is possible to construct situations where they come into conflict. In situations of conflict, either one criterion overrides another and the organization follows accordingly, or situations of organizational ambiguity result.

Given the extensibility of the auditory image metaphor and the principles of continuity and coherence, it should be possible to develop a psychologically relevant theory of musical attention and organization that covers the range from the formation of the image of a single event to the accumulation of the 'image' of a musical form, passing through many intermediate levels of organizational polyvalence (each element is potentially a member of several concurrent organizations) and the construction of composite musical objects that have a complex evolution through time. I feel that structural and functional ambiguities are a very important part of musical organization and if well understood can be used with great effectiveness and power in musical composition.

So where do we stand with respect to the initial questions? As concerns what might possibly be attended to as a musical image, a good start has been made with an understanding of certain basic principles of sequential and simultaneous organization, at least at the level of source organization.

There remains much work to be done on the effects of higher level organizing principles, such as underlying metric and rhythmic structure and underlying harmonic structure, on what can be followed throught time as a coherent entity and on what can be grouped as musically meaningful conglomerates or composite images. With respect to the second and third questions, the principles outlined here are certainly an important group of processes that are involved in the act of auditory organization. Again, what needs to be further researched (as much in musical as in psychological paradigms) is the extent to which the active, creative involvement of the listener can play a role in the organizing of complex constellations of sound events into musical images.

As I am prone to reiterate ad nauseum, each listener still carries into the musical situation 'normal' tendencies of hearing that are going to act as defaults in the organization of musical sound. However, what is most compelling as a result of all of the research on auditory organization is the fact that the will and focus of the listener play an extraordinarily important role in determining the final perceptual results. Musical listening (as well as viewing visual arts or reading a poem) is and must be considered seriously by any artist as a creative act on the part of the participant. As mentioned previously, perceiving is an act of composition, and perceiving a work of art can involve conscious and willful acts of composition. What this proposes to the artist is the creation of forms that contain many possibilities of 'realization' by a perceiver, to actually compose a multipotential structure that allows the perceiver to compose a new work within that form at each encounter. This proposes a relation to art that demands of perception that it be creative in essence.

ACKNOWLEDGMENTS

Some of the research reported in this chapter was conducted while the author was supported, in part, by a research grant from the Minister of Foreign Affairs of the French Government. I would like to express appreciation to Al Bregman for kindly communicating unpublished research to me and for a continually stimulating, dialogue on principles of auditory organization. I continue to draw heavily from his example and his work. Finally, I would like to thank Wendy Lindbergh, Tod Machover and David Wessel who provided helpful critques of this manuscript.

FOOTNOTES

1. A cassette tape of the sound examples described in this text is available by writing to the author.

2. Vibrato is approximately sinusoidal modulation with modulation rates between about 3-10 Hz. Jitter is an aperiodic modulation generally found in all natural sustained-vibration sources like voices, bowed strings and winds: the frequency spectrum of the modulation itself usually has a low-pass characteristic, that is has a greater predominance of lower frequencies, particularly those below 30-50Hz (McAdams, 1983).

3. Rms stands for root-mean-square which is a measure of the overall deviation from the centre frequency of a given partial. Several

different experiments have shown that this is a good measure since vastly different modulation waveforms, like vibrato and jitter, can be equated for amount of deviation with this measure (Hartmann and Klein, 1983 ; McAdams, 1983c).

4. 1 cent = 1/100 of a semitone; 12 cents is approximately 0.7% of the frequency.

5. This, however, has been shown by Cohen (1980) to be dependent to some extent on the form of the amplitude envelope. Inharmonic (and by implication multi-pitched) sounds are most often heard as fused, single sources when they have exponentially decaying amplitude envelopes as one finds with many kinds of struck sound sources, for example, strings, bars, tubes, plates, etc.

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

REFERENCES

Barriere, J.B. (1983).'Chréode' for computer generated tape. Paris: IRCAM.

Békésy, G. von. (1960). Experiments in Hearing. New York: McGraw-Hill.

Boulez, P. (1981/1982). 'Repons' for orchestra, soloists, live electronics and computer generated tapes. London: Universal Editions.

Bregman, A.S. (1977). Perception and behavior as compositions of ideals Cognitive Psychology, 9, 250-292.

Bregman, A.S. (1978). The formation of auditory streams. In J. Requin (Ed.) Attention and Performance. Volume 7. Hillsdale, New Jersey: Lawrence Erlbaum.

Bregman, A.S. (1980). The conceptual basis of perception and action. In Perception and Cognition II: Presentations on Art Education Research. Volume 4. University of Montreal: Concordia.

Bregman, A.S. (1981). Asking the 'what for' question in auditory perception. In: M. Kubovy and J. Pomerantz (Ed.), Perceptual Organization. Hillsdale, New Jersey: Lawrence Erlbaum.

Bregman, A.S. (1982). Two-factor theory of auditory organization. Journal of the Acoustical Society of America, 72, S10 (A).

Bregman, A.S., Abramson, J. and Darwin, C. (1983). Spectral integration based on common amplitude modulation. Unpublished manuscript, McGill University, Montreal.

Bregman, A.S. and Campbell, J. (1971). Primary auditory stream segregation and the perception of order in rapid sequences of tones. Journal of Experimental Psychology, 89, 244-249.

Bregman, A.S. and Pinker, S. (1978). Auditory streaming and the building of timbre. Canadian Journal of Psychology, 32, 19-31.

Bregman, A.S. and Tougas, Y. (1979). Propagation of constraints in auditory organization. Unpublished manuscript, McGill University, Montreal.

Bozzi, P. and Vicario, G. (1960). Due fattori di unificazione fra note musicali: La vincinanza temporale e la vincinanze tonale. Rivista de Psychologia, 54, 235-258.

Cherry, E.C. (1953). Some experiments on the recognition of speech with one and with two ears. Journal of the Acoustical Society of America, 25, 975-979.

Chowning, J.M. (1980). Computer synthesis of the singing voice. In: Sound Generation in Winds, Strings, Computers. Royal Swedish Academy of Music, Publication number 29, Stockholm.

Cohen, E.A. (1980). The influence of non-harmonic partials on tone perception. Unpublished doctoral Dissertation, Stanford University.

Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review, 83, 114-140.

Dannenbring, G.L. and Bregman, A.S. (1976). Stream segregation and the illusion of overlap. Journal of Experimental Psychology/Human Perception and Performance, 2, 544-555.

Deutsch, D. (1975). Two-channel listening to musical scales. Journal of the Acoustical Society of America, 57, 1156-1160.

Dowling. W.J. (1973). The perception of interleaved melodies. Cognitive Psychology, 5, 322-337.

Durlach, N.I. (1972). Binaural signal detection: Equalization and cancellation theory. In: J.V. Tobias (Ed.), Foundations of Modern Auditory Theory. Volume 2. New York: Academic Press.

Grey, J.M. and Moorer, J.A. (1977). Perceptual evaluations of synthesized musical instrument tones. Journal of the Acoustical Society of America, 62, 1493-1500.

Hartmann, W.M. and Klein, M.A. (1980). Theory of frequency modulation detection for low modulation frequencies. Journal of the Acoustical Society of America, 67, 935-946.

Helmholtz, H.L.F von (1954). On the Sensations of Tone as a Physiological Basis for the Theory of Music (Dover, New York, from 1885 edition of English translation by A.J. Ellis).

IRCAM: Un Portrait. (1983). Recorded disc with accompanying text. Paris: IRCAM.

Jeffress, L.A. (1972). Binaural signal detection: Vector theory. In J.V. Tobias (Ed.) Foundations of Modern Auditory Theory. Volume 2. New York: Academic Press.

Julesz, B. (1971). Foundations of Cyclopean Perception. Chicago, Illinois: University of Chicago press.

Kubovy, M. and Jordan, R. (1979). Tone-segregation by phase: On the phase sensitivity of the single ear. Journal of the Acoustical Society of America, 66, 100-106.

McAdams, S. (1980). The effects of spectral fusion on the perception of pitch for complex tones. Journal of the Acoustical Society of America, 68, S109 (A).

McAdams, S. (1981). Auditory perception and the creation of auditory images. Paper presented at Informatica e Composizione Musicale, Festival Internazionale di Musica Contemporanea, La Biennale di Venezia, Venice.

McAdams, S. (1982a). Contributions of sub-audio frequency modulation and spectral envelope constancy to spectral fusion in complex harmonic tones. Journal of the Acoustical Society of America, 72, S11(A).

McAdams, S. (1982b). Spectral fusion and the creation of auditory images. In: M. Clynes (Ed.) Music, Mind and Brain: The Neuropsychology of Music. New York: Plenum.

McAdams, S. (1983a). Acoustic cues contributing to spectral fusion. Proceedings of the International Congress of Acoustics, Paris, France, 3, 127-130.

McAdams, S. (1983b). L'image auditive: Un métaphore pour la recherche musicale et psychoacoustique. In: 'Un Portrait': IRCAM. Paris: IRCAM.

McAdams, S. (1983c). Spectral fusion, spectral parsing and the formation of auditory images. Unpublished doctoral dissertation, Stanford University, California.

McAdams, S. and Bregman, A. (1981). Hearing musical streams. Computer Music Journal, 3, 26-43.

McAdams, S. and Wessel, D. (1981). A general synthesis package based on principles of auditory perception. International Computer Music Conference, Denton, Texas.

McNabb, M. (1981). Dreamsong: The composition. Computer Music Journal, 5, 36-53.

Mills, A.W. (1972). Auditory localization. In: J.V. Tobias (Ed.), Foundations of Modern Auditory Theory. Volume 2. New York: Academic Press.

Noorden, L.P.A.S. van (1975). Temporal coherence in the perception of tone sequences. Unpublished doctoral dissertation. Tech, Hogeschool, Eindhoven, The Netherlands.

Noorden, L.P.A.S. van (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. Journal of the Acoustical Society of America, 61, 1041-1045.

Rasch, R. (1978). The perception of simultaneous notes such as in polyphonic music. Acustica, 40, 21-33.

Rasch, R. (1979). Synchronization in performed ensemble music. Acustica, 43, 121-131.

Reynolds, R. (1983). 'Archipelago' for orchestra and computer-generated tape C.F. Peters, New York.

Rodet, X. (1980a). CHANT Manual. Paris: IRCAM.

Rodet, X. (1980b). Time-domain formant-wave-function synthesis. In: J.C. Sinon, (Ed.) Spoken Language Generation and Understanding. Reidel, Dordrecht, Holland.

Scheffers, M. (1983). Sifting vowels: Auditory segregation and pitch perception. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Schubert, E.D. (1979). Psychological Acoustics. Stroudsberg, Pennsylvania: Dowden, Hutchinson and Ross.

Vicario, G.B. (1965). Vicinanza spaziale e vincinanza temporale nella segregazione di eventi. Rivista di Psicologia, 59, 843-863.

Vicario, G.B. (1982). Some observations in the auditory field. In: J. Beck (Ed.), Organization and Representation in Perception. Hillsdale, New Jersey: Lawrence Erlbaum.

Wessel, D. (1979). Timbre space as a musical control structure. Computer Music Journal, 3, 45-52.

Wessel, D. (1983). Timbral control in research on melodic patterns. Paper presented at the Fourth International Workshop on the Physical and Neuropsychological Foundations of Music, Ossiach, Austria.