TEMPORAL GESTALT PERCEPTION IN MUSIC

James Tenney

with Larry Polansky

Introduction

For the historian, *time* is not the undifferentiated "continuum" of the theoretical physicist, but a hierarchically ordered network of moments, incidents, episodes, periods, epochs, eras, etc.—i.e., time-spans whose conceptual "boundaries" are determined by the nature of the events or processes occurring within them (or of the historian's interpretation of these events or processes). Similarly for the musician, a piece of music does not consist merely of an inarticulate stream of elementary sounds, but a hierarchically ordered network of sounds, motives, phrases, passages, sections, movements, etc.—i.e., time-spans whose *perceptual* boundaries are largely determined by the nature of the sounds and sound-configurations occurring within them. What is involved in both cases is a conception of distinct spans of time—at several hierarchical levels—each of which is both internally cohesive and externally segregated from comparable time-spans immediately preceding and following it. Such time-spans (and the events or processes which define them) will here be called *temporal gestalt-units* (or "TGs").

In the years that have elapsed since the early papers on gestalt perception by Wertheimer, Köhler, and others[1, 2, 3], a considerable body of literature has accumulated which deals with the visual perception of *spatial* gestalt-units, although some of this literature remains highly speculative. Much less has been written (even of a speculative nature) about the perception of *temporal* gestalt-units. Some useful analogies have been drawn between visual and auditory perception, but such analogies provide little insight into the basic mechanisms of temporal gestalt perception, and many of the questions which might be the most relevant to *musical* perception have not even been asked by perceptual psychologists, much less answered. How, for example, are the perceptual boundaries of a TG determined? To what extent are the factors involved in temporal gestalt perception *objective*, bearing some measurable relation to the acoustical properties of the sounds themselves? Assuming that there are such objective factors, is their effect strong enough that one might be able to *predict* where the TG boundaries will be perceived, if one knows the nature of the sound-events that will occur?

In an effort to provide some tentative answers to such questions, a hypothesis of temporal gestalt perception will be proposed in Section 1 of this paper, and Section 2 will present some results of a computer analysis program based on this hypothesis. The program represents a simplified *model* of this aspect of musical perception, and some of the implications, limitations, and possible extensions of this model will be considered in Section 3. Although the hypothesis on which the model is based is very simple, it involves some unfamiliar concepts and terms that will have to be explained before the hypothesis will be comprehensible. Some of these concepts were first stated—albeit in rather embryonic form—in an earlier paper,[4] though these have evolved considerably in the intervening years.[5, 6] Others have emerged more recently, in the effort to organize the more general music-theoretical ideas into an explicit "algorithmic" form. Though I will not recount the history of the development of the model, I will try to describe the conceptual transformations of these earlier ideas in a way which parallels their actual historical development.

1. The Fundamental Hypothesis.

As in my earlier writings, I shall use the terms "element," and "clang," and "sequence" to designate TGs at the first three hierarchical levels of perceptual organization. An element may be defined more precisely as a TG which is not temporally divisible, in perception, into smaller TGs. A clang is a TG at the next higher level, consisting of a succession of two or more elements, and a succession of two or more clangs—heard as a TG at the *next* higher level—constitutes a sequence. In the earlier writings names were not given to TGs at levels higher than that of the sequence, but recently we have been using the terms "segment" and "section" for units at the next two higher levels. The TG at the highest level normally considered is, of course, coextensive with the *piece* itself, although situations are certainly conceivable where still larger gestalt-units might be of interest—e.g., the series of pieces on a concert, or the set of all pieces by a particular composer.

In *Meta + Hodos* (1961), I designated *proximity* (in time) and *similarity* (with respect to any or all other parameters) as the two "primary factors of cohesion and segregation" involved in musical perception (or, more specifically, in clang-formation) as follows:

> In a collection of sound-elements, those which are *simultaneous* or *contiguous* will tend to form clangs, while relatively greater separations in time will produce segregation, other factors being equal. Those which are *similar* (with respect to values in some parameter) will tend to form clangs, while relative dissimilarity will produce segregation, other factors being equal.

Aside from certain other differences between these early formulations and my more recent ideas (e.g. that two or more simultaneous elements do not necessarily constitute a clang, but more likely what I would now call a "compound element," several problems had to be solved before the current algorithm could be designed.[7] First, the principles, as stated, were not "operational," but merely descriptive. That is, although they were able to tell us something about TGs whose boundaries were already determined, they could say nothing about the process by which that determination was made. They described the *results* of that process, but not its *mechanism*. Second, "similarity" was not defined in any precise way, except by reference to "values in some parameter." The assumption here, of course, was that the similarity of two elements is an inverse function of the magnitude of the *interval* by which they differ in some parameter. This remains a plausible assumption, though it was never made explicit—but even such a correlation of similarity/dissimilarity with interval-magnitude does not, by itself, allow for the simultaneous consideration of more than one parameter at a time. This rather profound difficulty was implicit in the "other factors being equal" clause appended to the two statements. At the time, this qualification seemed necessary, in order to rule out cases where two or more parameters vary in conflicting ways, or where two or more "factors" function independently. Although this was a useful device for isolating and studying some important aspects of temporal gestalt perception, it imposed a very severe limitation on the range of musical examples whose gestalt structure might be predicted. In most real musical situations, other factors are manifestly *not* equal, and our perceptual organization of the music is a complex result of the combination and interaction of several more-or-less independent variables.

Third (and finally), these early formulations referred to *one* hierarchical level only—the grouping of elements into clangs—although it was obvious to me even then that the similarity-factor, at least, was of great importance in the perceptual organization of TGs at all higher levels. In a later paper,[8] an attempt was made to generalize these principles, restating them in a way that would be applicable to all hierarchical levels. Thus (from Proposition II, p. 4),

> The perceptual formation of TGs at any hierarchical level is determined by a number of factors of cohesion and segregation, the most important of which are *proximity* and *similarity;* their effects may be described as follows: relative temporal *proximity* [and] relative *similarities* of TGs at a given hierarchical level will tend to group them, perceptually, into a TG at the next higher level. Conversely, relative temporal *separation* and/or *differences* between TGs will tend to segregate them into separate TGs at the next higher level.

Although these later "propositions" served to extend the earlier formulations to higher levels, they suffered all of the other deficiencies of the later formulations: they were non-operational in character, imprecise with respect to the concept of "similarity," and restricted to one parameter (or factor) at a time.

The first of these problems has been solved by a shift of emphasis from the unifying effects of proximity and similarity to the segregative effects of temporal separation and parametric dissimilarity, and by a more careful consideration of these effects as they must occur in real time. In the ongoing process of perception in time, TG-boundaries are determined by successive *TG-initiations.* This obviously applies to the beginning of a TG, but also to the *end* of it, since the perception that it has ended is determined (in the monophonic case, at least) by the perception that a *new* TG at that same hierarchical level has *begun.* In this new light, the effect of the proximity-factor (at the element/clang level) might be restated as follows:

> In a monophonic succession of elements, a clang will tend to be initiated in perception by any element which begins after a time-interval (from the beginning of the previous element, i.e., after a *delay-time*) which is *greater than those immediately preceding and following it*, "other factors being equal."

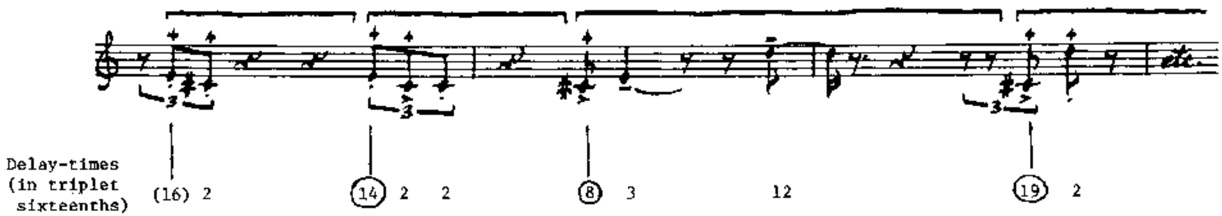
Thus, in mm. 24-28 of Varèse's DENSITY 21.5 (Example 1), where clang-initiations are determined almost entirely by the proximity-factor, it can be seen that the elements which initiate successive clangs are, in fact, invariably those whose delay-times are "greater than those immediately preceding and following" their own (the delay-times associated with each element are indicated in the example by the numbers below the staff, in triplet sixteenth-note units; those which are circled are for the clang-initiating elements). Note that the first occurrence of D (at the end of m. 25) does not initiate a new clang, in spite of its fairly long delay-time (12 units), because the delay-time which follows it is still longer (19 units). As stated above, the proximity-factor begins to take on a form that *is* "operational." In a musical situation where no other parameters are varying (say, a drum solo, at constant dynamic level), this principle can provide an unambiguous procedure for predicting clang-boundaries.

In an analogous way, the effect of the similarity-factor (at the element/clang level) may be reformulated as follows (and note that this statement can actually include the previous one as a special case, if the parameter considered is time, and the "interval" is a delay-time):

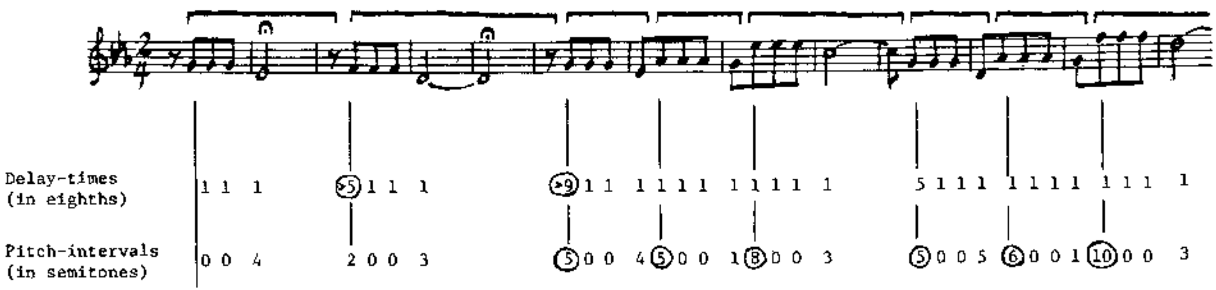
> In a monophonic succession of elements, a clang will tend to be initiated in perception by any element which differs from the previous element by an interval (in some parameter) which is *greater than those* (inter-element intervals) *immediately preceding and following it*, "other factors being equal."

This, too, is "operational," in that it suggests an unambiguous procedure for predicting clang-boundaries, though it is limited to special cases where only one parameter is varying at a time. Consider, for example, the first 12 measures of Beethoven's *Fifth Symphony*. Example 2 shows the melodic line, abstracted from all contrapuntal/textural complications—as it would be heard, say, in a piano transcription. Because of the considerable difference in tempo here, compared to the Varèse example—and thus in the actual duration of notated time-values—relative weights are used that give the value of 1 to the eighth-note (as well as to the semitone, as before). The clang-initiations during the first six bars are obviously determined by the proximity-factor alone, but beginning in m. 6, the proximity-factor can have no effect on the clang-organization (except in m. 9), because the delay-times are all equal. This passage is not heard simply as two clangs, however, but as a succession of clangs (indicated by the brackets above the staff), each consisting of four elements. And note that, for every clang-initiating element, the pitch-interval associated with it is greater than those immediately preceding and following it.

The parallelism of the proximity- and similarity-factors, as re-stated



Example 1. Clang-initiations determined by delay-times



Example 2. Clang-initiations determined first by delay-times (mm. 1–5), then by pitch-intervals (mm. 6–12)

above—and the fact that the second statement can be considered to include the first one as a special case—is extremely important. In both, it is the occurrence of a *local maximum in interval magnitudes* which determines clang-initiation. An *interval* is simply a *difference*, and whether this is a difference in starting-times, or pitch, or intensity—or any other attribute of sound—is not what is important. Rather, it is *relative differences* (in *any* parameter) that seem to be crucial. We live in a "universe of change," but whether a particular change marks the beginning of a new temporal gestalt-unit or simply another "turn" in the *shape* of the current one depends not only on its absolute magnitude, but on the magnitude of the changes which precede and follow it.

The restriction to one parameter (or factor) at a time, still implicit in the last formulation, remains to be overcome before our principle can be of much use in predicting clang-initiations in any but a very limited set of musical situations. What is needed is some way to combine or integrate the interval-magnitudes of *all* parameters into a single measure of change or difference. The solution to this problem involves a concept that has been employed by experimental psychologists for several decades now—that of a *multidimensional psychological or perceptual "space."*[9,10,11] The "dimensions" of this space are the several parameters involved in the perception and description of any sound, i.e., time, pitch, and intensity. Other parameters (e.g. timbre) could be added to this list, if they satisfy certain conditions, but I shall limit my discussion here to these three basic ones. The set of parametric values characterizing an element serve to locate that element at some "point" in this multi-dimensional space, and we can consider not only the intervals between two such points (one along each separate axis), but also a *distance* between those points, which takes into account the contribution of intervals in each individual parameter, but effectively combines these into a single quantity. Such a distance, or distance-measure—what a mathematician would call a "metric"[12]—may now be used in place of the less precise notions of "similarity" and "proximity."[13] In order to do this, however, two further questions had to be answered: first, how to *weight* the several parameters relative to each other (thereby "scaling" the individual dimensions) in a way that is appropriate to musical perception, and second, what kind of function to use in computing these distances.

The weightings referred to above are necessary for two reasons: first, because quantitative scales of values in the several parameters—and thus the numbers used to encode these values as input data to a computer program—are essentially arbitrary, bearing no inherent relation to each other; and, second, because we have no way of knowing, *a priori*, the relative importance of one parameter *versus* another, in its effects on TG-formation. As yet, no clear principle has been discovered for determining what the weights should be. The current algorithm requires that they be specified as input data, and the search for "optimum" weightings has so far been carried out purely on a trial-and-error basis. It now appears that such optimum weightings are slightly different for each piece analyzed, which suggests that there might be some correlation between these optimum weightings and statistical (or other) characteristics of a given piece, but the principles governing such correlations have yet to be determined.

Regarding the type of distance-measure to be used, there are many different functions which can satisfy the mathematical criteria for a metric, and therefore many distinct measures that might be used. A definitive answer to the question as to which of these metrics is the most appropriate to our musical "space" would depend on the results of psychoacoustic experiments that, to my knowledge, have never been done, although studies of other multidimensional perceptual or psychological spaces provide a few clues toward an answer.[14] The best-known metric, of course, is the Euclidean, but after trying this one, and noticing certain problems which seemed to derive from it, another was finally chosen for the algorithm. This second distance-measure is sometimes called the "city-block" metric, and an example of this metric vs. the Euclidean is shown graphically in Figure 1, for the two-dimensional case. When three or more dimensions are involved, the relations become difficult or impossible to represent graphically in two dimensions, but the relationships are the same. In the Euclidean metric, the distance between two points is always the square root of the sum of the squares of the distances (or intervals) between them in each individual dimension (in two dimensions, this is equivalent to the familiar Pythagorean formula for the hypotenuse of a right triangle). In the city-block metric, on the other hand, the distance is simply the sum of the absolute values of the distances (or intervals) in each dimension.[15]

One of the most important steps in the development of our model involved the decision to treat musical space as a metric space within which all the individual parametric intervals between two points might be integrated into a single measure of distance, and to use this distance, in turn, as a measure of relative "cohesion" (or "segregation") between two musical events. This made it possible to reformulate the basic principle of TG-initiation in a new way, which can be applied to virtually any musical situation, without the old restriction to variations in just one parameter at a time (though it is still limited to the element/clang level, and to monophonic textures), as follows:

A new clang will be initiated in perception by any element whose

Euclidean metric:

$$d = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$

"City-block" metric:

$$d = |y_2 - y_1| + |x_2 - x_1|$$

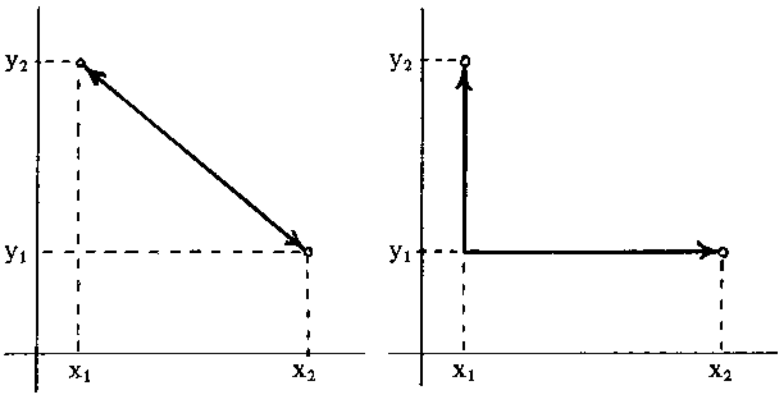
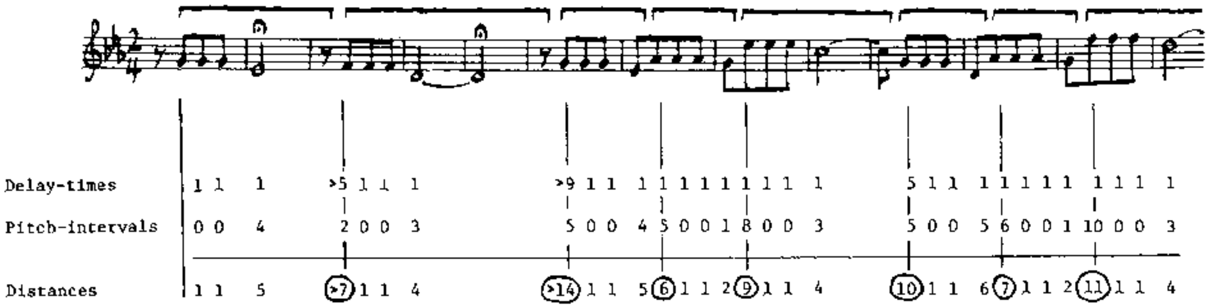


Figure 1. Euclidean vs. "city-block" distances.

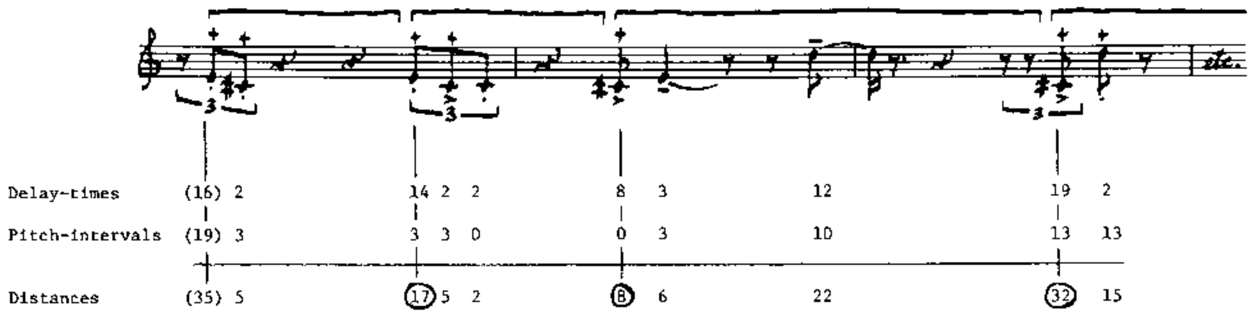
distance from the previous element is greater than the inter-element distances immediately preceding and following it.

If we now apply this principle to the Beethoven example considered earlier, using (again) relative weights for duration and pitch that give values of 1 for both the eighth-note duration and the semitone (Ex. 3), we see that this simple principle serves to predict or locate all of the clang-initiations involved in the passage (note that each inter-element distance, listed in the bottom row of figures, is simply the sum of the two (weighted) intervals associated with each element). As a second example, consider the Varèse passage quoted earlier (Ex. 4). Although in this case delay-times alone were sufficient to determine clang-initiation, we see that maxima in the distance-function will still predict the same boundaries. Again, our simple principle of clang-initiation seems to determine clang-boundaries in a reasonable way.

One final problem remained to be solved, before the current algorithm could be realized—that of extending this basic principle of clang-formation to higher levels. The discussion so far has been limited to TG-initiations at the element/clang level because the notion of a "distance" in the musical space can only be used properly as a difference between two *points* in that space. How might the "differences" between two clangs, sequences, or still higher-level TGs—which would correspond to clusters or *sets* of points—be defined? It has seemed to me that such differences are of three basic kinds, corresponding to three distinct aspects of our perception (and/or description) of these higher-level TGs, namely, differences of *state, shape,* and *structure.*[16] By "state" I mean the set of average or *mean values* of a TG (one for each parameter except time), plus its *starting-time.* The state of a TG might thus be compared to the "center of gravity" of an object in physical space, except that the temporal counterpart to mean parametric value is the *beginning* of the TG, rather than its "center." "Shape" refers to the contour or profile of a TG in each parameter, determined by changes in that parameter with time, and "structure" is defined as "relations between subordinate parts" of a TG—i.e., relations between its component TGs at the next lower level (or at several lower levels). Thus, the differences between any two TGs may be differences between their states, or between their shapes, or between their structures, or any combination of these. At the element-level, however, the differences that are reflected in the measure of distance are of the first kind only (differences between states) because we are not yet dealing with shape at the element-level, and because structure is assumed to be imperceptible at this level, by the very definition of "element" as "not temporally divisible, in perception, into smaller TGs" (see above).



Example 3. Clang-initiations determined by inter-element distances



Example 4. Clang-initiations determined by inter-element distances

It is not yet clear what role similarities and differences of shape and structure might have in temporal gestalt perception, but it is quite clear that state-differences have virtually the same effects at the higher levels that they have at the element-level. Consequently, shape and structure play no part in the current model, but state-differences (i.e. intervals and distances) are treated essentially the same way at all hierarchical levels, with just one additional refinement, not mentioned previously. Although the magnitude of change perceived when one element follows another is well-represented by the distance-measure defined above, the magnitude of change perceived in the succession of two clangs, sequences, or higher-level TGs is only partially accounted for by this distance. In addition, the changes perceived *at the boundary* between two TGs have an important influence on TG-initiation at higher levels. In order to deal with this, a distinction is made between "mean-intervals" and "boundary-intervals," as follows:

> A *mean-interval* between two TGs at any hierarchical level, in any parameter except time, is the difference between their *mean values* in that parameter; for the time-parameter, a mean-interval is defined as the difference between their *starting-times*. A *boundary-interval* between two TGs is the difference between the mean values of their *adjacent terminal components* (i.e. the final component of the first TG and the initial component of the second).

Note that a boundary-interval at one hierarchical level is a mean-interval at the next lower level.

An analogous distinction is made between "mean-distances" and "boundary-distances," as follows:

> The *mean-distance* between two TGs at any hierarchical level is a weighted *sum of the mean-intervals* between them, and the *boundary-distance* between two TGs is a weighted *sum of the boundary-intervals* between them.

Finally, mean- and boundary-distances are combined into a single measure of change or "difference" which we call "disjunction," defined as follows:

> The *disjunction* between two TGs, or the disjunction of a TG with respect to the preceding TG (at a given hierarchical level) is a weighted *sum of the mean-distance and the boundary-distances between them at all lower levels.*

Note that, whereas the weightings referred to in the definitions of mean- and boundary-distances are weightings across *parameters*, the weightings used in the definition of disjunction are weightings across *hierarchical levels*. In the program, these are set to decrease by a factor of 2 for each successively lower level considered. The disjunction between two sequences, for example (or the disjunction of the second sequence with respect to the first), involves—in addition to the mean-distance between them—one-half of the mean-distance between their adjacent terminal clangs, and one-fourth of the mean-distance between the adjacent terminal elements of those clangs.[17]

Now, at last, it becomes possible to state the fundamental hypothesis of temporal gestalt perception, on which the current model is based, as follows:

A new TG at the next higher level will be initiated in perception whenever a TG occurs whose disjunction (with respect to the previous TG at the same hierarchical level) is greater than those immediately preceding and following it.

2. *THE MODEL.*

A computer analysis program based on the hypothesis developed in the previous section has been written by Larry Polansky, and used to obtain hierarchical segmentations for several pieces.[18,19] It is beyond the scope of this paper to describe this program in any detail, but a few points must be noted before its results can be appropriately evaluated. The model has certain limitations, in terms of the kind of music it can deal with, as well as the musical factors which it considers, and it is essential that these limitations be clearly understood. First of all, it can only work with *monophonic* music. Although in principle the same concepts and procedures should be applicable to polyphonic music, there are certain fundamental questions about how we actually *hear* polyphonic music which will have to be answered before it will be possible to extend the model in that direction. In addition, and for the same reason, the algorithm is not yet able to deal with what might be called "virtual polyphony" in a monophonic context—that perceptual phenomenon which Bregman has called "stream segregation."[20] Real as this phenomenon is, I think it can only be dealt with, algorithmically, by a more extended model designed for polyphonic music.

The next two limitations of the algorithm are related to each other, in that both have to do with factors which are obviously important in musical perception, but which the current model does not even consider, namely *harmony* (or harmonic relations between pitches or pitch-classes), and *shape* (pattern, motivic/thematic relations). What the algorithm is capable of doing now is done entirely without the benefit (or burden) of any consideration of either of these two factors. Thus, although it is by no means a comprehensive model of musical perception, the very fact that it does so much without taking these factors into account is significant.

Still another type of limitation is inherent in certain basic procedures used by the program. For one thing, all higher-level TGs must contain *at least two* TGs at the next lower level (thus there can occur no one-element clangs, or one-clang sequences, etc.). Furthermore, no *ambiguities* regarding TG-boundaries are allowed: a terminal element might be the initial element in a clang *or* the final element in the preceding clang, but it cannot be both. A different approach to this problem, involving the notion of a "pivotal" TG—i.e., a TG which might function as *both* an initial component of a TG at the next higher level *and* as the final component of the preceding TG (at that same, higher level)—has recently been sketched, but has not yet been implemented.

Finally, the reader should be warned that the output of this program says absolutely nothing about the musical *function* of any of the TGs it finds. It merely partitions the overall duration of the piece into component TGs at several hierarchical levels. Questions of function are left entirely up to us, to interpret as we will. What the algorithm *does* purport to tell us is where the temporal gestalt boundaries are likely to be perceived—surely a prerequisite to any meaningful discussion of the musical "function" of the TGs determined by these boundaries.

Input data to the program are numbers representing the pitch, initial intensity, final intensity, duration, and rest-duration of each element in the score, plus weighting factors for each parameter, and certain constants for the particular piece or run (e.g. the total number of elements, the tempo of the piece, etc.). Numerical values for these parameters are encoded as follows: in order to avoid "roundoff" errors, the value of 1.2 (rather than 1.0) is used for the quarter-note at the specified tempo for the piece, with other note-values proportional to this. Thus, an eighth-note = .6, a triplet-eighth = .4, etc. These values are re-scaled, internally, to units of one-tenth of a second. Pitches are represented by integers, with the value of 1 usually assigned to the lowest pitch in the piece (although this is entirely arbitrary, since the program's operations involve only the intervals between pitches, not the pitches themselves). For intensity, integer values from 1.0 through 8.0 are used for the notated dynamic levels, *ppp* through *fff*, with decimal fractions for intermediate values, as during a gradual crescendo or diminuendo. In transcribing the score, these fractional values are derived by simple linear interpolations between the integer values.

At the element-level, then, three basic parameters are involved: time (or "delay-time," determining "proximity"), pitch, and intensity, and weights must be input for each of these. At the clang-level, and carrying through to all higher levels, a new parameter emerges which I considered important in musical perception, namely *temporal density* (or, more strictly, element-density, as a function of time). Provision was therefore made in the program for this parameter, although it has turned out to be unnecessary. Our best results on the pieces analyzed so far have been obtained with a weight of zero for temporal density. The program also allows for input data (and a weighting-factor) to be given for one more parameter, which we call "timbre," but which could be used for any other attribute of sound that seemed appropriate in a particular piece. It should be noted, however, that meaningful results can only be expected if this additional parameter is one in which values may be specified (or at least approximated) on what S. S. Stevens[21] has called an "interval scale." So far, it has only been used in a very primitive way, with scale values of either 0 or 1, to represent the "key-clicks" in measures 24–28 of Varèse's *Density 21.5*. Provision was originally made for specifying the weights in each parameter for mean- and boundary-intervals independently. As it turned out, however, the optimum weightings seemed to be the same in any given parameter for both types of interval, so they are now both given the same value.

The parametric weights used for the results shown in Examples 5–7 are as follows:

	duration	*pitch*	*intensity*	*timbre*
Varèse	1.0	0.67	6.0	20.0
Webern	1.0	0.5	6.0	0.0
Debussy	1.0	1.5	2.0	0.0

An input weight of 1.0 implies a time-unit of one-tenth of a second, a unit-interval of one semitone, or of one dynamic-level-difference (as between *mf* and *f*), depending on the parameter involved. The set of weights listed above may thus be taken to imply certain equivalences between intervals in the several parameters, at least with respect to their effects on TG-initiation. In the Debussy piece, for example, a delay-time of one-tenth of a second is equivalent to a pitch-interval of two-thirds of a semitone, and to one-half of one dynamic-level-difference. In the Varèse piece, on the other hand, a delay-time of one-tenth of a second is equivalent to a pitch-interval of 1.5 semitones, and to one-sixth of one dynamic-level-difference. The relatively large intensity-weights for both the Varèse and the Webern pieces confirms what one would already have expected—that both of these composers were using dynamics as a structural (rather than merely "expressive") parameter in these pieces. The differences between the pitch-weights for the three pieces are more difficult to explain. As noted earlier, it seems likely that correlations may eventually be found between these "optimum" weightings and some statistically measurable aspect of the pieces themselves, but no such correlations have yet been found.

The input data described above are used in a first "pass" through the program to compute inter-element intervals, distances, and disjunctions, and the latter are tested to determine the points of initiation of successive clangs, according to the fundamental hypothesis described in the previous section of this paper. The beginning of each new clang is assumed to define the end of the preceding clang, and when that clang's boundaries have thus been determined, the program computes and stores its starting-time, average pitch, and average amplitude—i.e. values which represent what I have called its "state." When there are no more elements to be considered, the program returns to the (temporal) beginning of the piece, but one hierarchical level higher. It then goes through successive clangs, computing and storing sequence-initiations and states. This procedure continues "upward" through progressively higher hierarchical levels until a level is reached at which there are not enough TG's to make a next-higher-level grouping possible (i.e., less than four). The program's architecture is thus hierarchically recursive; the computations are *essentially identical* at every level of TG organization, and this is one of the most attractive features of the model.

Results of the program for three pieces—Varèse's *Density 21.5;* Webern's Concerto, Op. 24, 2nd movement; and Debussy's *Syrinx*—are displayed in the form of graphically annotated scores in Examples 5 through 7. The segmentation given by the algorithm for each piece is indicated by the vertical lines above the staff-notation, each extending to a horizontal line corresponding to the hierarchical level of the largest TG initiated at that point. For the first two pieces, these results may be compared with analogous segmentations to be found in the analytical literature. In the case of *Density 21.5,* a segmentation which is both explicit and complete is available, and will be used for comparison—that given in a monograph by Jean-Jacques Nattiez.[22] For the Webern example, an analysis of the first "period" by Leopold Spinner[23] will be compared to the results of our program. The results for Debussy's *Syrinx* are given without any such comparisons, because we have not found any published analyses of this piece in which the segmentation is sufficiently explicit to justify a comparison.

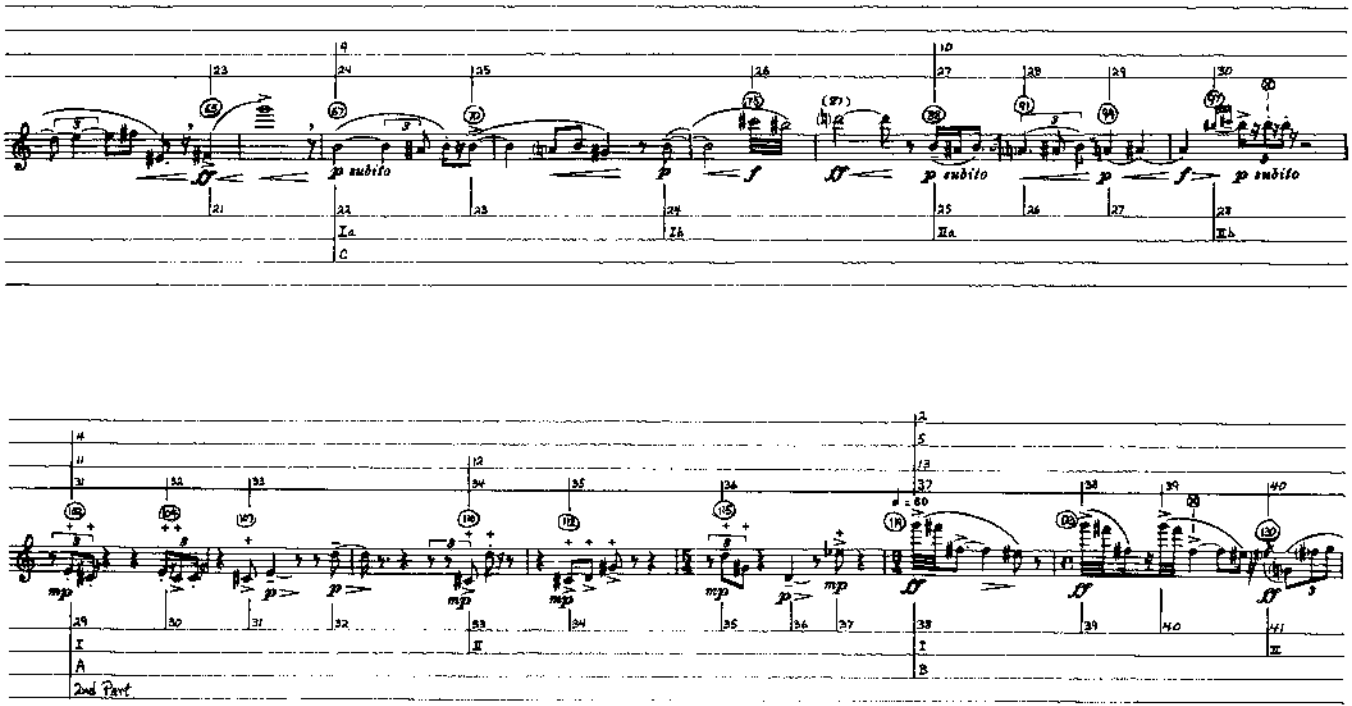
Edgard Varèse: DENSITY 21.5

The segmentation given by Nattiez for this piece is shown in the lower portion of Example 5, so that a direct, point-by-point comparison can be made. Here the correlations between the two partitionings are quite close—especially at the clang- and sequence-levels—although the two are not identical, of course, and the similarities diminish at higher levels. In fact, some 81% of the clang-initiations in our results, and 85% of the sequence-initiations (but only 44% of the segment-initiations) coincide with the corresponding boundaries in Nattiez's segmentation. There are no coincidences at any higher level. Some of the discrepancies between the two segmentations are fairly trivial, as where one of the two "models" simply interpolates an extra clang-break between two otherwise coincident boundaries (as at elements 8, 25, 54, 109, 117, 118, 140, 179, 224, 226, 233, and 241). A few differences result from the fact that Nattiez does not prohibit one-component TGs, as our model does. These occur in his segmentation in the form of "one-element clangs" beginning at elements 109, 117, and 118, and as sequences containing only a single clang, beginning at elements 22, 52, 74, and 97.

Even if we disregard such discrepancies as these, however, there will still remain a number of places where the two segmentations differ. Some of these probably have to do with the fact that neither harmonic nor motivic factors are considered by our algorithm. For example, the high-level TG-initiation which Nattiez locates at element 188 is clearly determined by the fact that the initial motivic idea of the piece suddenly returns at this point, and a model which included some consideration of motivic relations might well yield a result here more like Nattiez's. On the other hand, the strong element of surprise that this return of the initial motive evokes in my perception of the piece suggests that this motivic factor is here working very much "against the grain" of most of the other factors of TG-organization, and that an important part of the musical effect of this event in the piece depends on the fact that the motive recurs at a point which would *not* otherwise be a high-level boundary.

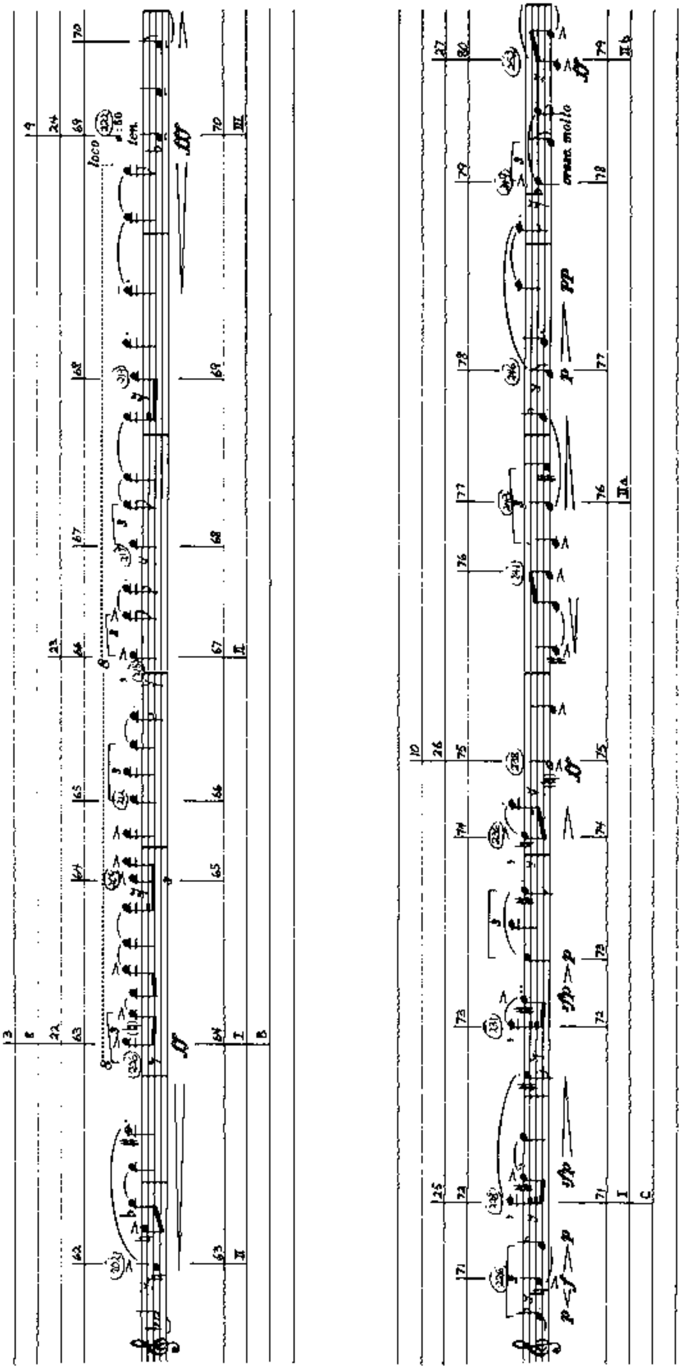
After all of the foregoing reasons for the differences between the two segmentations have been accounted for, a few discrepancies will remain which suggest that our weightings may not be quite "optimum" after all, or that they are simply different from those unconsciously assumed by Nattiez, or even that some aspect of our algorithm may need refining. Finally, however, I must say that I think our segmentation represents the perceptual "facts" here more accurately than Nattiez's at certain points. These would include the clang-initiations at elements 13, 20, and 75, and the sequence-initiations (and perhaps even the segment-breaks) at 177 and 238.

Example 5. Edgard Varèse: *Density 21.5*

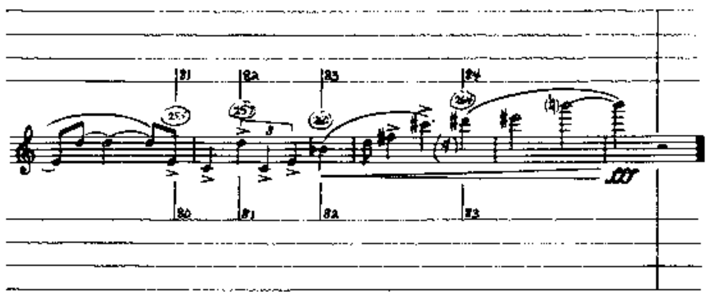


(Ex. 5, cont.)

(Ex. 5, cont.)



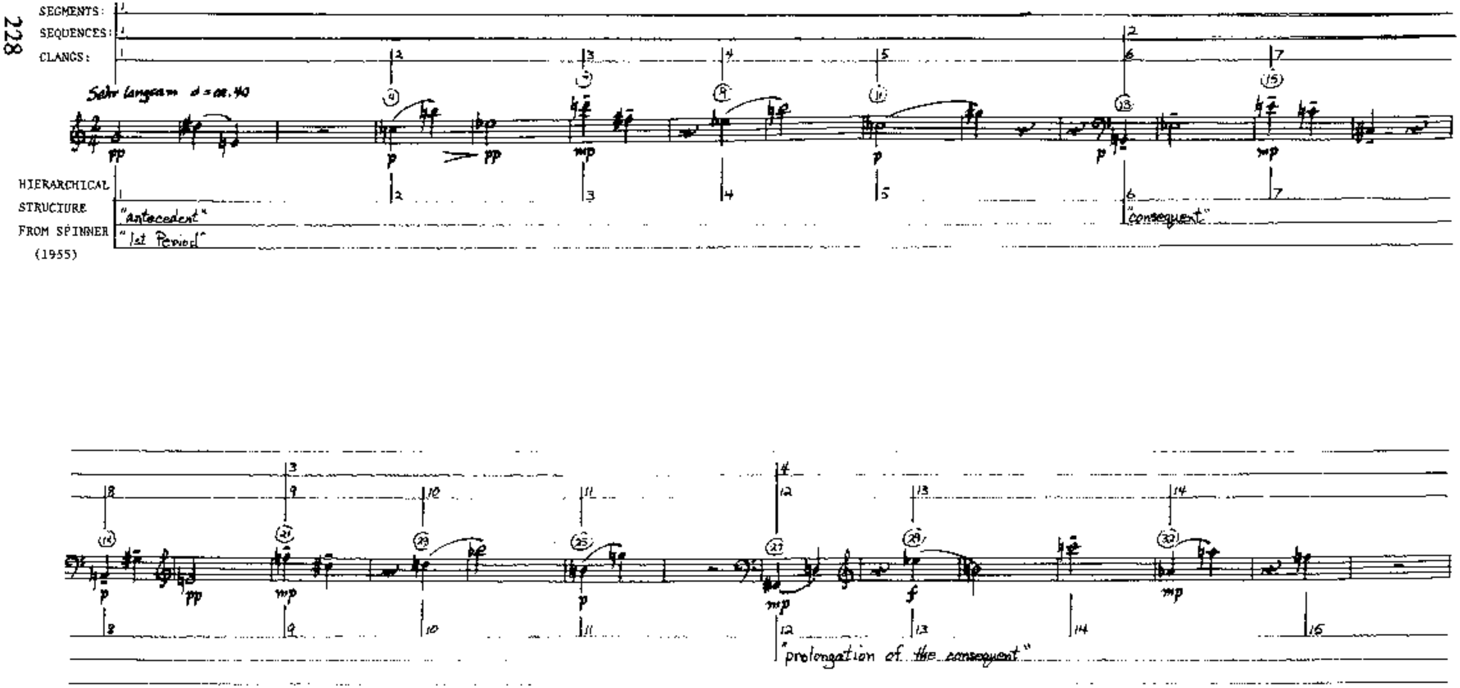
(Ex. 5, cont.)



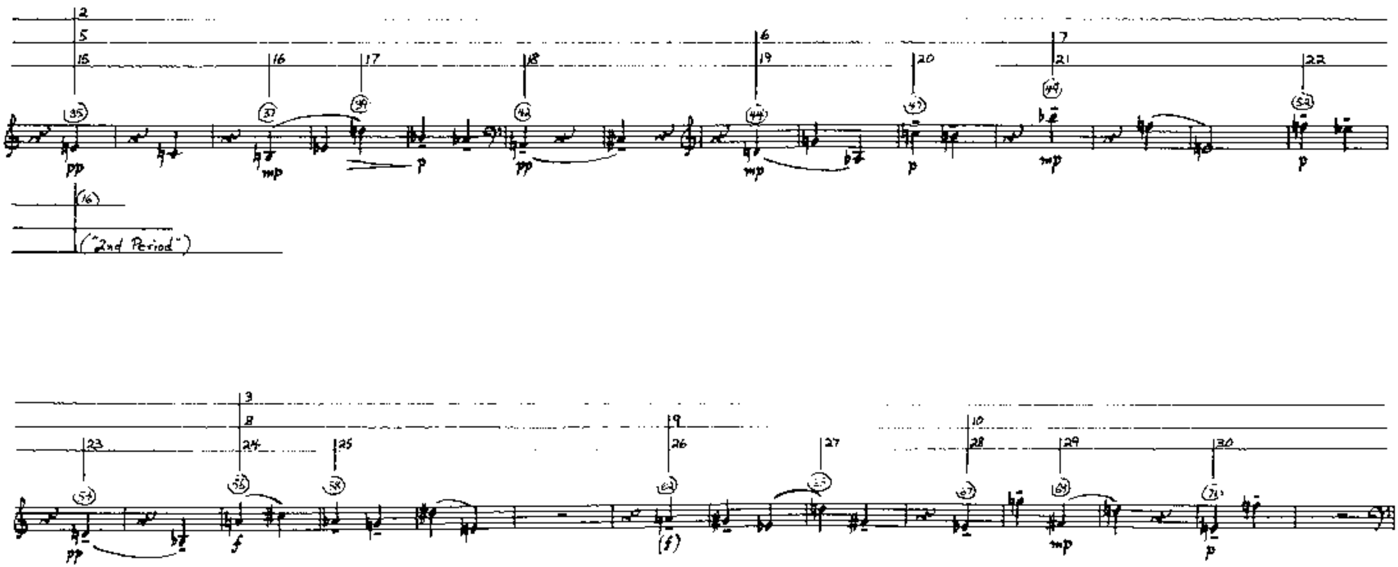
(Ex. 5, cont.)

Anton Webern: CONCERTO, Op. 24, 2nd movement (melodic line only)

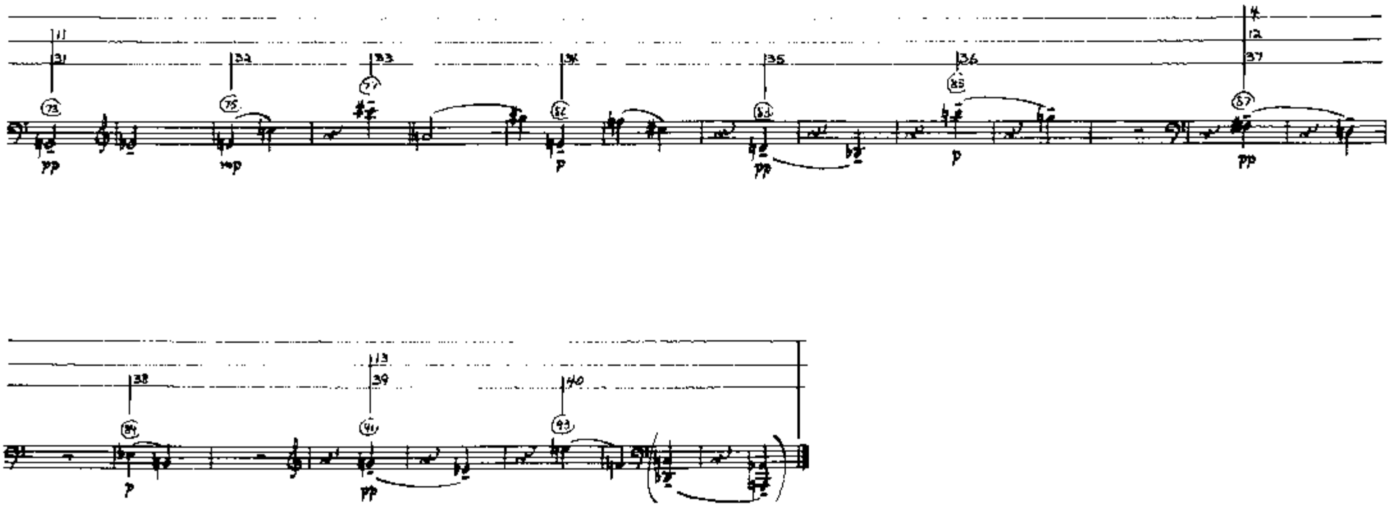
The segmentation given by our program for the first 28 bars of this piece is identical at every point but two with that assumed by Leopold Spinner in his "Analysis of a Period"[24] (see Example 6). Spinner's first "period" is equivalent to our first segment, and each of the three parts into which he divides this period ("antecedent", "consequent", and "prolongation of the consequent") begins at a point which coincides with one of our sequence-breaks (although the program further divides Spinner's "consequent" into two sequences). Our clangs are coincident with his "phrases" everywhere except at elements 31–34 (marked x in the lower part of line 2 of the annotated score), but the discrepancy here is easily explained. Spinner's concern in the analysis is to demonstrate a cohesive unity in the music resulting from the recurrences of a limited set of rhythmic motives, in addition to that deriving from serial pitch-relations. At the point in question, he notes the equivalence of a three-note motive beginning in m. 25 (element 31) with the motive which begins the movement (𝅗𝅥 ♩ ♩). To my ear, however, the oboe's high C in m. 25 sounds like the final element in the 3-element clang beginning in m. 23 (element 29), as our program determines it, rather than an initial element, as Spinner would have it.



Example 6. Anton Webern: Concerto, Op. 24, 2nd movement, (melodic line only)



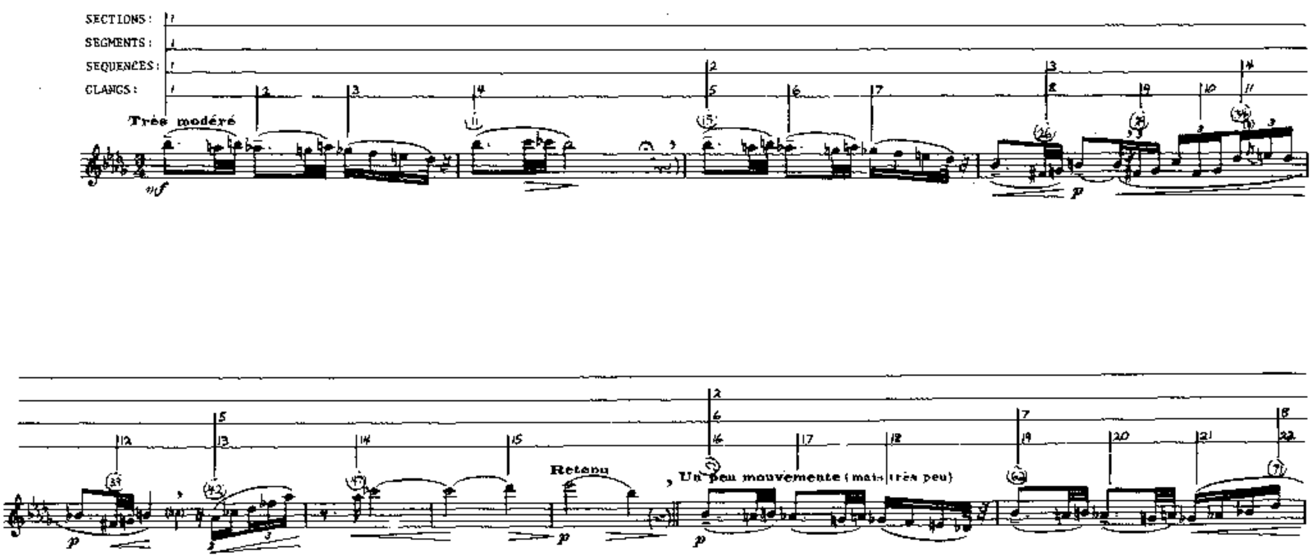
(Ex. 6, cont.)



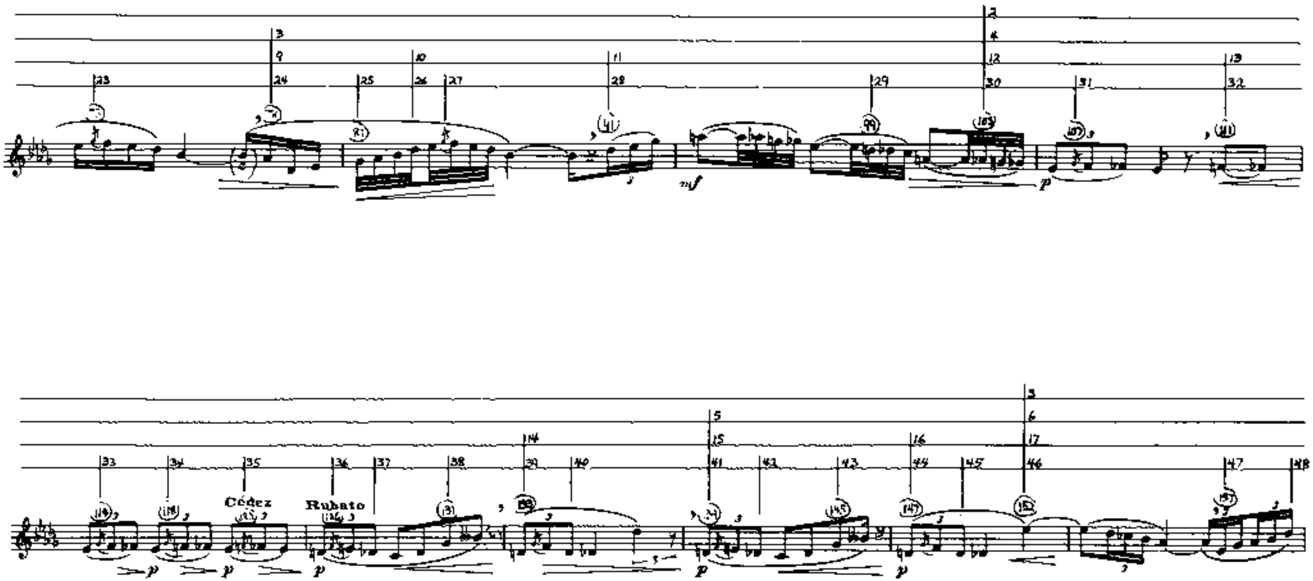
(Ex. 6, cont.)

Claude Debussy: SYRINX

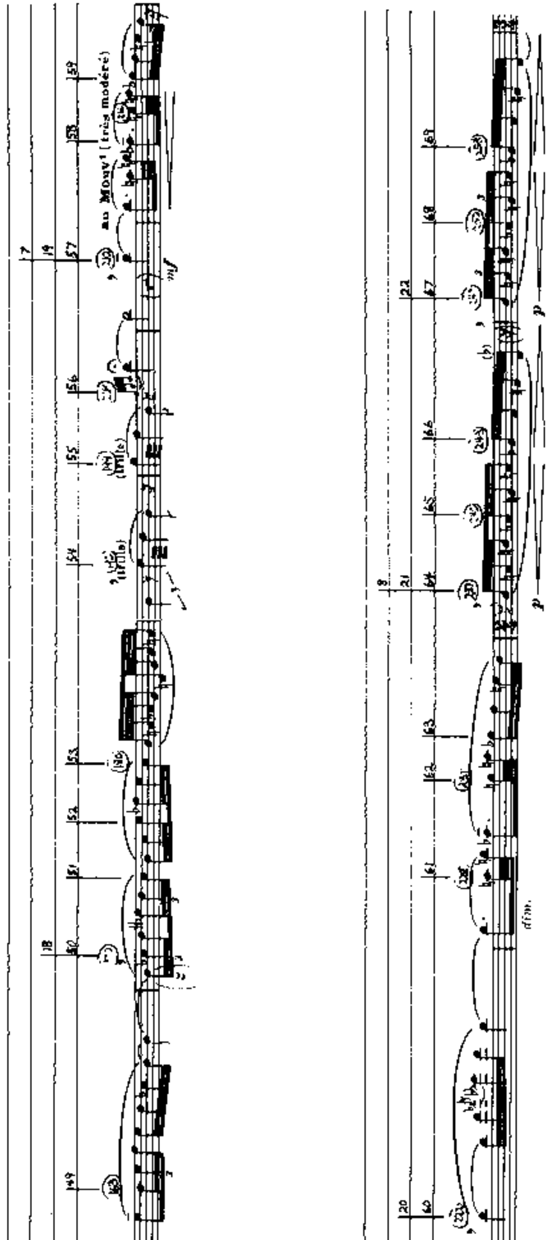
Our best results for this piece, using the parametric weights listed earlier, are shown in Example 7. In the absence of any other analysis with which these results might be compared, I shall leave it to the reader to decide whether–and to what extent–they correspond to the temporal gestalt organization he or she might make of this piece "spontaneously." I should point out, however, that the intention behind these analyses (at this stage in the development of the model) has not been to demonstrate a segmentation which is more accurate or "correct" than another, derived by alternative means–"spontaneous" or systematic. The music has been used primarily to *test the model*, and the only claim that might reasonably be made at this point is that our algorithm is remarkably effective, considering the simplicity of the hypothesis on which it is based. What the results of the model seem to show are aspects of the structure of these pieces that we all more-or-less take for granted. We have proceeded on a sort of faith in the commonality of our (*all* of our) perceptual "structurings" in this respect, and the validity of our model may ultimately stand or fall according to whether this faith was justified or not.



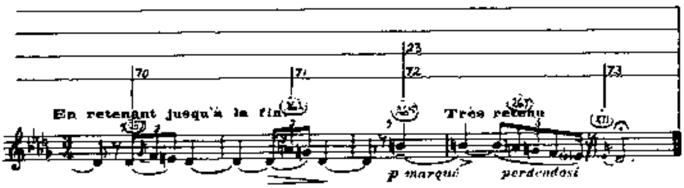
Example 7. Claude Debussy: *Syrinx*



(Ex. 7, cont.)



(Ex. 7, cont.)



(Ex. 7, cont.)

3. Applications, Implications, and Possible Extensions of the Model.

In spite of the rather severe limitations of this model, the degree to which its results correspond to segmentations arrived at by other means suggests that the "fundamental hypothesis of temporal gestalt perception" on which the model is based is at least a plausible formulation of an important principle of musical perception. As such, it may have useful applications for the composer as well as the theorist, since it can be used to create perceptually effective formal structures without recourse to traditional devices—"tonal" or otherwise. For example, serial, aleatoric, and stochastic compositional methods frequently result in textures which are *statistically homogeneous* at some fairly low hierarchical level. A typical negative response to this kind of formal situation (which I have elsewhere called "ergodic"[25]) is that, although "everything is changing, everything remains the same." Whether this is to be considered undesirable or not obviously depends on a number of purely subjective factors, including the expectations of the listener, the intentions of the composer, etc.—none of which are of concern to me at the moment. What is of concern, however, is the fact that the model outlined in this paper suggests a technique for *controlling* this aspect of musical form, when the composer's intentions make such control desirable. A piece becomes "ergodic" (with respect to some parameter) as soon as a hierarchical level is reached at which the *states* of successive TGs are indistinguishable—i.e. at that level at which the mean-intervals between successive TGs (in that parameter) are all effectively zero. In general, this can be shown to depend on the degree to which parametric ranges are *constrained* at the lower levels. That is, the more the total available range in some parameter is "used up" at a given level, the smaller will the average effective differences be *between* TGs at that level, and the more quickly will the texture approach "ergodicity" at the next higher level. The technical remedy for this is simply to *distribute* the total available ranges more evenly over as many hierarchical levels as needed to achieve the formal structure intended.

The model also has certain interesting implications regarding the nature of musical perception. One of the most surprising of these involves what might be called the "decision-delay" between the moment of initiation of a TG at any level and the moment at which this TG-initiation can be perceptually determined or "known." This is the result of several basic conditions inherent in the model, including (1) the fact that the TG-initiating effect of a given disjunction is dependent upon the disjunction which *follows* it (as well as the one which precedes it), (2) the fact that the measure of disjunction involves intervals between *mean* parametric values (i.e. "states") of those TGs, and that these mean values can only be determined after that TG has *ended,* and (3) that this, in turn, is determined by the perception that a *new* TG has begun at that level. The decision-delays resulting from these various conditions are shown schematically in Figure 2, where it can be seen that the delays are *cumulative* at progressively higher levels, and become quite long fairly quickly. The implications of this for musical perception are significant, especially for what they tell us about the importance and function of *memory* and *anticipation.* Clearly, the higher the level concerned, the greater will be the demands on short-term memory, if the TG-boundaries are to be determined at all, and the less certain these boundary determinations must be on a first hearing. On second and later hearings—i.e. with gradually increasing familiarity with a piece—these delays may be diminished, or finally eliminated altogether, to the extent to which TGs which have not yet occurred can be anticipated, via longer-term memory. Thus, while the indispensible importance of memory to musical perception is a matter of common agreement, and the anticipation of what is about to be heard in a familiar piece is surely a common experience, our model goes one step farther and suggests that the primary *function* of both memory and anticipation is *to diminish the delay between the moment of occurrence of a TG and the moment of recognition of its gestalt boundaries, and eventually to bring these into synchrony.*

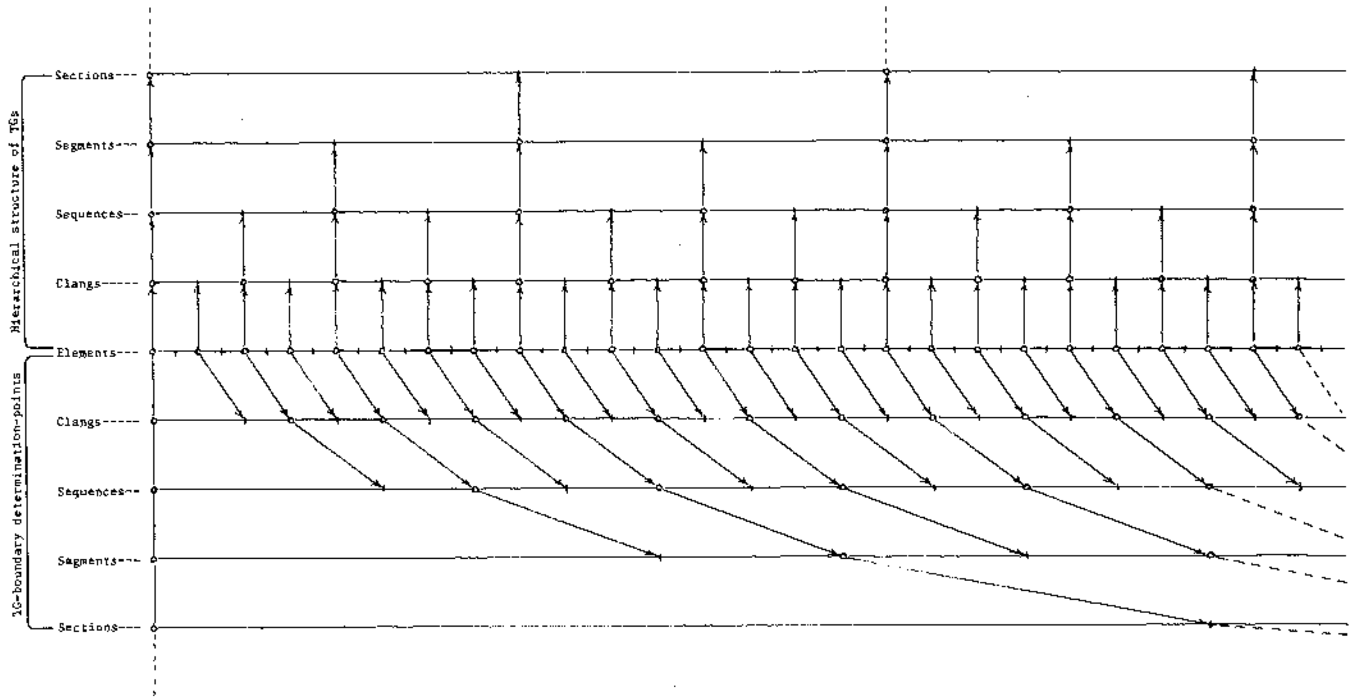


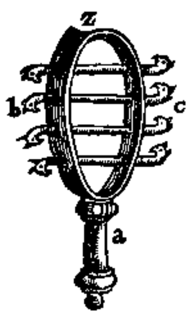
Figure 2. Schematic diagram of decision-delays implied by the model

The extent to which our temporal gestalt perception might be confused, if not utterly confounded, by these phase-shifting "decision-delays" might appear to throw into question the efficacy of the model described here, if it were not for the very considerable *information-reduction* implicit in the model. That is, the information that is retained, at a given hierarchical level, for determining TG-initiations at that level, is always less than (or at most, equal to) *half* of the information that was needed at the next lower level. The ratio of information-reduction here depends on the average number of components per higher-level TG, which is by definition at least two. In fact, the average, for the pieces analyzed so far, turns out to be slightly larger than three.

The algorithm described here obviously needs to be tested with other musical examples, so that considerable work remains to be done with the program in its present form. In addition, there are several extensions of the model which ought to be possible, and which promise to be important to the growth of our understanding of musical perception, and perceptual processes in general. One area in which such extensions are most immediately needed would involve the incorporation of harmonic and motivic factors in the workings of the algorithm. Another area would include whatever elaborations might be necessary to enable it to deal with polyphonic music. Still another would involve some method of dealing with ambiguous TG-boundaries in a more flexible and musically realistic way (perhaps using the notion of "pivotal" TGs, mentioned earlier).

Finally, it should be possible to extend the model "downward" to sub-element levels, which would not only eliminate the tedious process of transcription now required to specify input data to the program, but also be far more *accurate* than this process can ever be, in representing the sounds as we actually hear them. Such an extension would involve analog-to-digital conversion of the acoustical signal into numerical "samples," suitable for input to the computer program. These samples would then constitute the "elements" (or micro-elements) whose parameters would be subjected to computational procedures essentially the same as in the current algorithm. Element states (sample amplitudes and starting-times) and inter-element disjunctions would be computed, and used to determine the points of initiation of micro-TGs at the next higher level. Micro-clangs would probably correspond to individual periods of the original signal, and micro-sequences to groups of these periods delimited by the on-off behavior of the amplitude-envelopes and/or other modulation processes that might be present (vibrato, tremolo, etc.). Eventually, TGs will be found (probably at the micro-sequence or micro-segment level) whose boundaries correspond to those of the elements whose parameters are now given as input to the program. In the course of such a process, new "parameters" would emerge—pitch, and perhaps timbre—in the form of additional "states" not definable at the lowest level (where the only parameters were amplitude and time).

Many of the details of any "downward" extrapolation of the current model are still unclear, but I am convinced that such an extension to sub-element levels is an area of investigation well worth pursuing. Moreover, the conclusion seems justified that the basic procedures in this model will work, with perhaps only minor revisions, at *any* level of perceptual organization, and with "elements" whose description might involve other "parameters" than those relevant to sound. Thus, extrapolations of the model "upward" to TGs larger than individual pieces should be possible, as well as what might be described as extrapolations "outward" to temporal gestalt-units involving other modes of perception, or several different modes of perception simultaneously.



NOTES

1. Ellis, W. D., ed. *A Source Book of Gestalt Psychology,* Routledge and Kegan Paul Ltd.: London, 1955.
2. Köhler, W., *Introduction to Gestalt Psychology,* New American Library, 1959.
3. Koffka, K., *Principles of Gestalt Psychology,* Harcourt, Brace: New York, 1935.
4. Tenney, J., *Meta ≠ Hodos–A Phenomenology of 20th Century Music and an Approach to the Study of Form,* privately circulated monograph, 1961, published in 1964 by the Inter-American Institute for Musical Research, Gilbert Chase, editor. (This has long been out of print, and is now available only from the author.)
5. Tenney, J., "Form in 20th Century Music," in John Vinton, ed., *Dictionary of Contemporary Music,* E. P. Dutton: New York, 1971.
6. Tenney, J., "META Meta ≠ Hodos", *Journal of Experimental Aesthetics,* 1 (1977), pp. 1–10.
7. Some of these problems were noted by Wayne Slawson in his review of *Meta ≠ Hodos, Journal of Music Theory* 10 (1966), p. 156.
8. See note 6.
9. Attneave, F., "Dimensions of Similarity," *American Journal of Psychology* 63 (1950), pp. 516–556.
10. Shepard, R. N., "Attention and the Metric Structure of the Stimulus Space," *Bell Telephone Laboratories Technical Memorandum,* October, 1962.
11. Shepard, R. N., "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function" *Psychometrika* 27 (1962).
12. Cf. Beckenbach, E. & Bellman, R., *An Introduction to Inequalities,* New Mathematical Library, 1961.
13. It should be noted that Shepard uses the term "proximity" for what is here being called "distance".
14. Cf. notes 9 and 11.
15. The Euclidean and city-block metrics are themselves special cases of a more general class of distance-functions sometimes called the Minkowski metric which (in two dimensions) is of the form:

    $$d = ((x_2 - x_1)^R + (y_2 - y_1)^R)^{1/R} \text{ for } R \geqslant 1.$$

    Note that, when $R = 1$, this becomes the city-block distance-function, and when $R = 2$, it is equivalent to the Euclidean metric. It would be of interest to experiment with this parameter of the equation, in the context of the current algorithm. In particular, it might turn out that a value of R somewhere *between* 1 and 2 would be even more appropriate to the "space" of musical perception.
16. Cf. note 6.
17. Note that the sum of the weightings used to compute boundary-distances is always less than 1, but approaches this value as a limit when higher levels are being considered (i.e. $1/2 + 1/4 + 1/8 + \ldots < 1.0$).
18. A detailed description of the program, with a complete documentation of the relevant input and output data, is contained in an earlier research report,

written by this author (in collaboration with Larry Polansky) entitled "Hierarchical Temporal Gestalt Perception in Music: A 'Metric Space' Model," August, 1978.

19. Polansky, L., "A Hierarchical Gestalt Analysis of Ruggles' *Portals*," July, 1978.
20. Bregman, A. S. and Campbell, J., "Primary Auditory Stream Segregation and Perception of Order in Rapid Sequences of Tones," *Journal of Experimental Psychology* 89 (1971), pp. 244-249.
21. Stevens, S. S., "Mathematics, Measurement, and Psychophysics," in *Handbook of Experimental Psychology*, John Wiley and Sons, Inc.: New York, 1951.
22. Nattiez, J. -J., *"Densité 21.5" de Varèse: Essai d'Analyse Sémiologique*, Groupe de Recherches en Sémiologie Musicale, Faculté de Musique, Université de Montréal, Québec, 1975.
23. Spinner, L., "Analysis of a Period," *Die Reihe* 2, English Edition, 1958.
24. Ibid.
25. See note 6.