Apparent Motion in Music?

ROBERT O. GJERDINGEN
Center for Advanced Study in the Behavioral Sciences, Stanford

A great deal of the motion perceived in music is apparent rather than real. On the piano, for example, no continuous movement in frequency occurs between two sequentially sounded tones. Though a listener may perceive a movement from the first tone to the second, each tone merely begins and ends at its stationary position on the frequency continuum. Recent advances in the modeling of apparent-motion effects in vision provide a starting point for the modeling of the strong apparent-motion effects in music. An adaptation of the Grossberg-Rudd model of apparent motion in vision, when given input representing the strengths of pitch sensations positioned along a one-dimensional frequency continuum, can simulate important musical phenomena of auditory stream segregation, van Noorden’s melodic-fission/temporal-coherence boundaries, various Gestalt effects, aspects of dynamic attending, and Narmour’s predicted categorical distinction between musical intervals implying a continuation and those implying a reversal of direction.

I hear the melody and its accompaniment even when they are played by an old-fashioned clock where each tone is separate from the others.

Wertheimer, 1923

Introduction

A psychologist, recreating one of the classic demonstrations of Gestalt principles for a group of undergraduates, places two lights a short distance apart at the front of a lecture hall. After darkening the room, she briefly flashes first one light and then the other. The students see not two isolated flashes (the physical reality) but rather an integrated perception of motion from the first flash to the second. On the other side of campus an organist, demonstrating the subject of a Bach fugue for a music-appreciation class, presses the series of keys that sound the organ pipes in their proper sequence. The students hear not isolated, static tones (the physical reality)

Requests for reprints may be sent to Robert O. Gjerdingen, Center for Advanced Study in the Behavioral Sciences, 202 Junipero Serra Blvd., Stanford, CA 94305.
but rather an integrated perception of melody. The psychologist’s flashing lights produce a phenomenon known as apparent motion (Exner, 1875; Kolers, 1972; Körte, 1915; Wertheimer, 1912). Do the organism’s sounding pipes produce the same phenomenon? Is the “striking parallel” (Bregman, 1990, p. 21) between apparent motion in vision and in music more than just a tempting analogy?

Musicians, in their everyday speech, take it for granted that music moves. Melodies “go” here and there, instruments “run” up and down scales, vibrato “wobbles” around a central tone, trills “shake,” chords “leap” to higher or lower ranges, and so forth. The allied notion that tones leave behind a curvilinear trace of their motion has so deeply ingrained itself in musical discourse that “melody” and “melodic line” have become all but synonymous. Yet even if musicians—and listeners in general—concur in sensing motion in music, there is little agreement about the nature of that motion (Zuckerkandl, 1956). In viewing a classic visual demonstration of apparent motion, one can easily imagine an intermittently flashing light source that moves in normal three-dimensional space. In hearing the subject of a Bach fugue, however, it is not at all clear what is moving or where that motion takes place. So in the sense that a melody is not imagined as the product of an intermittently sounding, physically moving tone source, the analogy between apparent motion in vision and apparent motion in music breaks down.

If this often mentioned analogy (Bregman, 1990; Miller & Heise, 1950; van Noorden, 1975; Wertheimer, 1923) does not rest on high-level ascriptions of source and meaning, then perhaps it is grounded instead on shared low-level, Gestalt-like grouping processes that yield smooth, continuous motion traces from discontinuous inputs. This paper will explore the conjecture that many motion percepts in music may be a product of the same type of neural circuitry proposed by Grossberg and Rudd (1989) to explain apparent motion percepts in vision. The Grossberg-Rudd model is a massively parallel, multiplex, feed-forward neural-network system. Its nonlinear internal dynamics, though complicated in detail, endow the behavior of the system as a whole with a measure of simplicity. In the following sections, I will first present an overview of the Grossberg-Rudd model, stressing the emergent simplicity of its overall behavior, and then demonstrate how this motion-tracking system can account for various effects in music. These will include aspects of what more specialized studies have described as melodic fission (Miller & Heise, 1950), various Gestalt effects (Meyer, 1956), auditory stream segregation (Bregman & Campbell, 1971), temporal coherence (van Noorden, 1975), dynamic attending (Jones & Boltz, 1989), and the predicted categorical distinction between musical intervals implying a continuation and those implying a reversal of direction (Narmour, 1990).
The Basic Grossberg-Rudd Model Applied to Pitch

The stimuli that prompt the illusion of apparent motion are, as already illustrated, discontinuous in both space and time. A motion-tracking system must thus be capable of smoothing out both spatial and temporal discontinuities if it is to reintegrate discrete stimuli into a continuous percept of motion. The neural-network architecture of Grossberg and Rudd (1989) accomplishes the requisite reintegrations through the agencies of specialized processing levels. These levels, though originally described solely in terms of the visual system, can be generalized as stages of information processing that deal with (1) input, (2) temporal reintegration, (3) spatial reintegration, and (4) motion analysis. Figure 1 shows these four processing levels in schematized form.

**LEVEL ONE: INPUT**

Level one, when adapted for music input, approximates a one-dimensional continuum of virtual pitch sensations (Terhardt, 1974, 1984). That is, the model assumes that the auditory system, at some stage prior to the motion-tracking system, reduces the multiple signals produced by the many frequency components of a complex tone to a unitary signal of perceived pitch. That signal then excites only one point along the pitch continuum in the first level of the motion-tracking system. The assumption of pitch determination prior to motion analysis may, on the one hand, be an oversimplification, inasmuch as motion analysis probably contributes to the determination of pitch (Bregman, 1990). Moreover, the present model does not require a prior determination of pitch for the simulation of simple apparent-motion effects created by solo melodies or single bands of noise. On the other hand, more complex effects like the

---

**Fig. 1.** A schematization of four processing levels in the Grossberg-Rudd model of motion tracking.
tracking of separate voices in a multivoice musical texture are difficult to explain without either the prior unitization of pitch sensations or some more inclusive model of the interaction of motion analysis with the determination of pitch and timbre. In any case, the computer simulations described below all assume a strict form of pitch determination where a (simulated) musical tone excites, at any moment, only one of 980 separate pitch positions in level one, each position representing a 5-cent (½20th-of-a-semitone) interval of the four-octave range from C₂ to C₆.

**LEVEL TWO: TEMPORAL REINTEGRATION**

Level two receives signals from each level-one pitch position and processes them on the basis of their amplitude envelopes. In the simplest case, level two just smooths out rapid changes in signal intensities by averaging them over time (the special treatment of signal transients will be discussed later). Figure 2 illustrates how the output of the i-th level-two processing unit alters the angular rise and fall of the i-th level-one signal (assumed here to mirror the amplitude envelope of the simulated sound). This smoothing or temporal smudging can be modeled by various designs for neural group dynamics. For the computer simulations described below, however, the actual time averages were calculated directly by numerical approximation (Euler's method) to the differential equation

\[
d/dt x_i = -Ax_i + I_i,
\]

where \( x_i \) is the activation strength of the i-th pitch location at level two, \( A \) is its decay rate (\( A = 0.03125 \) in the simulations), and \( I_i \) is the signal strength from the i-th pitch location at level one.
LEVEL THREE: SPATIAL REINTEGRATION

If level two can be said to diffuse discrete temporal events over a wider span of time, level three diffuses discrete pitch events over a wider span of pitch (cf. Seibert & Waxman, 1989). Figure 3 depicts how a signal from the \( i \)th level-two unit diffuses across a broad log-frequency band of level-three units. The shape of that diffusion is assumed to approximate a Gaussian distribution whose height varies in proportion to the strength of the received level-two signal and whose standard deviation is an arbitrary musical interval of width \( W \). The spatial diffusion of level three, in conjunction with the temporal diffusion of level two, transforms a unitary level-one signal into a broadly distributed, smoothly rising and falling level-three response, as shown in Figure 4.

The activity of the \( i \)th processing unit in level one has no effect on the activities of other units in the same level. The same is true of units within level two. In level three, however, the diffusion from one level-two signal can overlap the diffusion from another level-two signal, producing a summed effect of both signals. For a given degree of diffusion, that is, for a given standard deviation of musical interval \( W \), the level-three response to tones more than \( 2W \) apart in pitch differs markedly from that to tones less than \( 2W \) apart (Grossberg, 1977). Figure 5 illustrates the case of a melodic interval where the first tone is a simulated \( A_4 \) (440 Hz), the second tone is an octave higher at \( A_5 \) (880 Hz), and the value of \( W \) equals a minor third (chosen as a rough approximation of one critical bandwidth in this frequency range). Because the melodic interval of an octave (=4\( W \)) far exceeds \( 2W \), there is very little overlap between the two Gaussian distribu-

![Diagram](https://example.com/diagram.png)

Fig. 3. An output signal from the \( i \)th level-two unit diffuses across a broad pitch band of level-three units. The diffusion is assumed to approximate a Gaussian distribution.
Fig. 4. The spatial diffusion of level three, in conjunction with the temporal diffusion of level two, transforms a unitary level-one output signal into a broadly distributed, smoothly rising and falling level-three response.

Fig. 5. The response on the front left is to a simulated $A_4$ (440 Hz), the subsequent response on the back right to a simulated $A_5$ (880 Hz). Given a level-two-to-level-three signal diffusion resulting in Gaussian distributions whose standard deviations ($W$) equal a minor third, the melodic interval of an octave ($=4W$) is greater than the critical distance of $2W$. In consequence, there is little overlap between the two Gaussian distributions.

Consequently the figure gives a visual impression of the sequential rise and fall of two distinct phenomena. Figure 6, by contrast, illustrates the case of a smaller melodic interval where the first tone is a simulated $D_5$, the second tone is a minor third higher at $F_5$, and $W$ again equals a minor third. Because this melodic interval ($=W$) is less than $2W$, there is considerable overlap between the two Gaussian distributions. Indeed, the decay of $D_5$ overlaps the rise of $F_5$ to such an extent that the figure may be interpreted as presenting the apparent motion of a single undulating
Fig. 6. The response near the lower center left is to a simulated $D_3$, the subsequent response near the upper center right to a simulated $F_3$. With a value of $W$ equal to a minor third, the melodic interval $D_3-F_3 (= W)$ is less than the critical distance of $2W$. In consequence, there is considerable overlap between the two Gaussian distributions.

phenomenon that shifts to a higher pitch level between successive peaks of activation.

LEVEL FOUR: MOTION ANALYSIS

Level four (and implicitly any higher level) must, in some fashion, analyze the activity at level three. At a minimum, such analysis must involve tracking the peaks of level-three activations as they move from one location to another. Thus level four must somehow reverse the process that occurred from level two to level three, where single points diffused into broad distributions. Lateral inhibition provides the essential level-four process for accomplishing that task. In particular, if broad level-three distributions are transmitted to level four, and if there is strong lateral inhibition between level-four units, then local maxima of level-three activations will translate at level four either into narrower distributions or, in the limiting case, to single points of activation.

Figure 7 traces, with a heavy black line, the moment-by-moment local maxima of the level-three activations shown previously in Figure 6. Notice how, as the first pitch begins to decay and the second begins to rise, the trace of level-three maxima moves smoothly toward the right. Figure 8 eliminates the varying activation levels of this trace in order to show an idealized level-four analysis of level-three activity, which I will henceforth term a motion trace.

Further refinements and complications in both the Grossberg-Rudd model and its predecessors (Burt & Sperling, 1981; Marr & Ullman, 1981; Seibert & Waxman, 1989) found their motivations in problems
Fig. 7. A heavy black line traces the moment-by-moment local maxima of the level-three activations shown previously in Figure 6. As the response to the first tone begins to decay and the response to the second, higher tone begins to rise, the trace of level-three maxima moves smoothly toward the right.

Fig. 8. Eliminating the varying activation levels shown previously in Figure 7 creates an idealized level-four analysis of level-three activity, here termed a motion trace.

raised by special visual phenomena. For instance, whereas the above-described level-two responses to the intensities of flashing lights were sufficient to model the so-called beta or phi phenomena of apparent motion, the problem of simulating the shift from apparent group motion to apparent element motion in the Ternus display (Ternus, 1926/1950) led Grossberg and Rudd to add to their model the agency of detectors sensitive to signal transients. Rather than review these refinements with respect to visual perception, I will attempt instead to show how the same refinements would grow out of problems posed by music perception.
Tracking Individual Voices in a Polyphonic Texture

Figure 9 presents both the musical notation of the final half measure of Bach’s G-minor prelude from *The Well-Tempered Clavier*, Book I (Bach, 1722) and the motion traces that this excerpt generated (note the rotation of axes compared with Figure 8: pitch is now the vertical axis and time the horizontal axis; only the traces for the upper three voices are shown because the trace for the missing bass voice is merely a straight line). Because Bach happened to keep all the voices separated from each other by intervals of a fifth or more (>2W) the basic model described above had no difficulty in producing four separate motion traces (cf. Huron, 1989). Had, however, Bach voiced this passage as shown in Figure 10, with the tenor D3 transposed an octave higher to D4, the motion-tracking model (even with a smaller W = major second) would no longer be able to distinguish between the tenor and alto voices. A musically trivial change would have a catastrophic effect on the model’s analysis, which, as the figure shows, no longer conforms to the contour of the moving alto voice.

![Diagram](image)

**Fig. 9.** The musical notation of the final half measure of J. S. Bach’s G-minor prelude from *The Well-Tempered Clavier*, Book I (1722) and the motion traces generated by a simulation of this excerpt. As compared with Figure 8, the pitch and time axes have rotated: here pitch is the vertical axis and time the horizontal axis. Traces are shown for the upper three voices only; the trace for the bass voice is, like the traces for the similarly stationary soprano and tenor (G4 and D3), merely a straight line.
Fig. 10. Introducing a musically trivial change in Bach’s cadence—transposing the tenor D₃ (cf. Figure 9) an octave higher to D₄—has a catastrophic effect on the model’s analysis. A confounding of the tenor and alto voices (even with a smaller W = major second) leads to a motion trace that bears little resemblance to the contour of the musically important alto voice.

The problem could be resolved in part by switching attention to a still narrower W. Yet a W smaller than a semitone would be required to segregate C₄ and D₄ into distinct streams. More likely the problem lies in the model having responded just as strongly to the held portions of the long tones as it did to the musically more salient tones in the rapidly moving voice. To surmount this problem, the model must have a means of responding preferentially to change rather than to stasis. As mentioned, such a means was introduced by Grossberg and Rudd (1989) in the form of detectors at level two that respond to signal transients. For the computer simulations, the response of these transient detectors is given by approximation to the differential equation

\[ \frac{d}{dt} x_i = -Bx_i + CF_i, \]  

where \( x_i \) is the activation strength of the \( i \)th transient detector at level two, \( B \) is its decay rate (\( B = 0.03125 \) in the simulations), \( F_i \) is the magnitude of the difference between the values of the \( i \)th input signal at times \( t \) and \( t - 1 \), and \( C \) is a constant (\( C = 0.03125 \) in the simulations). The signal then sent on to level three is the \textit{product} of the moment-by-moment out-
puts of equations 1 and 2. As shown in Figure 11, in response to the sudden onset of a steady-state tone, the gradual rise of the level-two sustained response is multiplicatively gated by the level-two transient response in such a way that the time course of the signal received at level three approximates a curvilinear pulse. A strong pulse would also be sent to level three should a tone end with a sudden and sharp offset. Indeed the instantaneous offset of a long sustained tone creates the strongest possible response (the product of a high sustained value and a high transient value).

Returning now to Bach’s cadence, we can see in Figure 12 that, with the addition of transient detectors at level two, the model can successfully track the moving voice even when it remains in close proximity to one or more sustained tones.

**Tracking Linear Processes: Scales and Arpeggios**

A scale is the *locus classicus* of melodic motion. The above-described model, with the addition of transient detectors, can approximate a continuous motion trace for any common type of scale. Figure 13 presents the musical notation of an ascending C-major scale and the motion trace produced by the model as it processed that scale. In what may be more than a fortuitous coincidence, the visual appearance of the motion trace suggests the metaphor inherent in the musical term *scale* (from the Italian *scala*: “stairs”).

The model is also capable of approximating a continuous motion trace for the most common types of arpeggios. Figure 14 presents the musical notation of an ascending C-major arpeggio and the motion trace produced by the model as it processed that arpeggio.
Fig. 12. With the addition of transient detectors at level two, the model can successfully track the moving alto voice even when it remains in close proximity to the sustained D₄ in the transposed tenor.

Fig. 13. The musical notation of an ascending C-major scale and the motion trace produced by the model as it processed that scale. The visual appearance of the motion trace suggests the metaphor inherent in the musical term “scale” (from the Italian scala, “stairs”).
Fig. 14. The musical notation of an ascending C-major arpeggio and the motion trace produced by the model as it processed that arpeggio. Although scales and arpeggios contain a series of unequal musical intervals, the motion traces traverse these unequal intervals in approximately equal times up to a distance of 2W.

The musical notation of the arpeggio hints at a potential problem for a motion-tracking system: the intervals from tone to tone are not of uniform size. Indeed, the scales and arpeggios that music theorists describe as linear “processes” (Meyer, 1956) are objectively nonlinear. A C-major scale contains two interval sizes: whole tone (C₄-D₄, D₄-E₄, F₄-G₄, G₄-A₄, A₄-B₄) and semitone (E₄-F₄, B₄-C₅). And a C-major arpeggio contains three interval sizes: minor third (E₄-G₄), major third (C₄-E₄, C₅-E₅), and perfect fourth (G₄-C₅). Were a motion-tracking system limited to a uniform speed in traversing the pitch axis, the interval of a perfect fourth would take five times as long to track as the interval of a semitone. Yet musical practice and perception show no strong evidence of such distinctions in linear processes. On the contrary, the very notion of “scale” or “arpeggio” assumes isochronous performance as the default case, suggesting that a motion-tracking system must be capable of traversing unequal intervals in approximately equal times.

Grossberg (1977) has shown how, for two sequentially activated Gaussian distributions whose origins are separated by less than 2W (two standard deviations), the maximum of their summed activations will traverse the halfway point of the interval between them in the same amount of time regardless of the exact distance involved. Thus unequal intervals
will be traversed in approximately equal times up to a distance of 2W. But what happens at distances greater than 2W?

Figure 15 presents the musical notation of an ascending linear process of alternating fifths and fourths. The motion-tracking model was able to connect the moves across the fourth ($G_4\rightarrow C_5 < 2W$ where $W =$ minor third) but was unable to connect the moves across the fifths ($C_4\rightarrow G_4$, $C_5\rightarrow G_5; > 2W$). As Miller and Heise (1950) said more than 40 years ago, “It is as if the listener’s ‘melodic tracking’ could not follow a sudden change larger than a certain critical amount.” On the one hand, this inability to connect the larger intervals is in keeping both with Meyer’s characterization of large intervals as “gaps” that imply a subsequent filling in (Meyer, 1973) and with Narmour’s categorical distinction between intervals smaller than a tritone ($=2W$) that imply linear continuation and intervals larger than a tritone that imply reversal of direction (Narmour, 1990). But on the other hand, the inability to connect large intervals is inconsistent with the musical effect of Figure 15. Listeners are quite capable of integrating an ascending sequence of fifths and fourths into a perceived linear process. Narmour himself recognizes this fact with a special category of “retrospective process.” He claims, in effect, that after the first two intervals of Figure 15, a

![Figure 15. The musical notation of an ascending linear process of alternating fifths and fourths. The motion-tracking model was able to connect the moves across the fourth ($G_4\rightarrow C_5 < 2W$ where $W =$ minor third) but was unable to connect the moves across the fifths ($C_4\rightarrow G_4$, $C_5\rightarrow G_5; > 2W$).]
retrospective shift occurs in the listener's sense of how large is large (Narmour, 1990, pp. 273–277).

If given a broader W, the present model could easily connect larger intervals. Yet how should the model determine when to shift the width of W? One obvious method for accommodating such a shift would be to make W a variable controlled by the width of occurring intervals. There are at least two major problems with this approach. First, large intervals are often registral gaps, not the beginnings of linear processes. By readjusting W for every large interval, the model would eliminate all of Meyer's registral gaps and all of Narmour's potential melodic reversals. Second, at an early stage the model would need to tag every tone as belonging to a particular polyphonic voice in order not to confuse the large intervals produced within a single voice with the large intervals produced between different voices. Yet how is the model to "know," at an early stage, which voice is which, especially if all voices have the same timbre? And in cases where a large interval in one voice coincides with smaller intervals in other voices, how would a single W be capable of accommodating the multiple sizes?

A more practical scheme for accommodating multiple interval sizes gives the motion-tracking system multiple copies of levels three and four operating in parallel, each copy with a W of different width. Such a multiplex system—an obvious analog of Marr's description of the visual system with its multiple receptive-field and filter sizes (Marr, 1982)—allows a large interval to be simultaneously connected and disconnected in different parts of the model. Automatic processes of attention allocation and attention shifting, to be described below, then contextually guide the overall interpretation of motion.

To summarize the treatment of linear processes, let me note that the basic model interprets common musical scales and arpeggios as connected motion traces resembling flights of stairs. Although the steps of these stairs are not equal distances apart, the model nevertheless traverses the distances in approximately equal times. For those linear processes with steps more than a tritone apart (>2W), the model can produce a shift of attention to a parallel copy of levels three and four where W is wider. In particular, the motion-tracking model is assumed to be multiplex, with numerous level-three-and-four combinations whose values of W range from the very small to the very large, centering perhaps around a mean of about one critical bandwidth.

**Multiplex Interpretations of an Ambiguous Melodic Figure**

Figure 16 shows a short melody that can be perceived in several different ways. One can focus on the regular succession of eighth notes and
Fig. 16. A short melody with an ambiguous contour. One can attend to either a widely oscillating seven-note series \((G_4-G_5-A_4-E_4-C_5-B_4-C_5)\) or two converging melodic lines (ascending, \(G_4-A_4-C_5-B_4-C_5\); descending, \(G_5-E_4-C_5-B_4-C_5\)). Or perhaps one can hear a little of both. For the motion-tracking model, a single choice of \(W\) necessarily forces a single interpretation of the melody. Setting \(W\) equal to a whole tone, for example, produces the motion analysis shown in Figure 17. While this analysis does a creditable job of following the two converging melodic lines, it shows no hint of the oscillation between the alternating high and low tones. A much wider \(W\) could follow the note-to-note oscillation, but at the expense of the converging outer lines.

As mentioned earlier, a solution to this dilemma is to have multiple copies of levels three and four, each with a different size \(W\) (Grossberg & Rudd, 1989). The outputs of 15 such level-three-and-four combinations, with their \(W_s\) rising exponentially from a minimum width of one quarter tone to a maximum slightly in excess of a major sixth, are overlaid upon one another in Figure 18. At the smallest \(W_s\), only the leading-tone \(B_4\) and \(C_5\) were close enough to form a joint trace; all the other tones formed the separate, straight traces seen most clearly toward the right of the figure. At the medium size \(W_s\), the tones of the two converging melodic lines formed stairlike traces of the type seen in Figure 17. And at the largest \(W_s\), the

Fig. 17. For the melody of Figure 16, a narrow \(W\) (= major second) results in motion traces of the two converging melodic lines but no trace of the oscillation between the alternating high and low tones. A much wider \(W\) could follow the note-to-note oscillation, but at the expense of the converging outer lines. A single value of \(W\) cannot accommodate both aspects of the melody.
Apparent Motion in Music?

Fig. 18. A multiplex version of the model, one with 15 copies of levels three and four, each with a unique size of W ranging from a quarter tone to over a major sixth, can produce motion traces that capture both aspects of the melody shown in Figure 16. For the medium size Ws, the tones of the two converging melodic lines formed stairlike traces of the type seen in Figure 17, while for the larger Ws, the alternating high and low tones formed the sinuous traces seen bouncing up and down toward the left of the figure. The horizontal lines represent the smallest size Ws, where all intervals except B⁴-C⁵ were more than the critical distance of 2W apart.

alternating high and low tones formed the sinuous traces seen bouncing up and down toward the left of the figure.

Many of the traces just shown in Figure 18 are easily affected by subtle shadings in the simulated performance of the melody. Note, for instance, that the traces of ascending motion from A⁴ directly to C⁵ are considerably sparser and less coherent than the traces either from A⁴ to E⁵, or from E⁵ to C⁵. A shading of performance that makes A⁴ more salient and E⁵ less salient will tend to strengthen the direct A⁴-C⁵ traces. Figures 19a–d compare these A⁴-to-C⁵ traces under different conditions of performance. Figure 19a gives the baseline performance of Figure 18, where all durations and amplitudes are uniform. Figure 19b shows the case of A⁴ lengthened by one eighth of its duration, and E⁵ shortened by the same duration. Note that the A⁴-C⁵ traces (the steeply rising short traces indicated by the arrow) are more coherent and filled in than those in Figure 18. Figure 19c gives the case of equal durations but with A⁴ made a simulated 2.5 dB louder than the other tones. Again, the A⁴-C⁵ traces are improved. And Figure 19d gives the combination of both durational shading (as in Figure 19b) and loudness shading (as in Figure 19c). Clearly the combination of both shadings greatly improves the A⁴-C⁵ traces.

A more global aspect of performance is set by the chosen tempo. As tempo increases (and durations decrease), the model begins to have difficulty connecting the traces between the most distant tones. Figure 20 compares the traces connecting the opening G⁴-G⁵ octaves at three different simulated tempos. At the left of the figure is the base tempo simulated in Figures 17–19, a tempo of x beats per minute. When, as at the middle of
Fig. 19. Highlights of four simulated performances of the melody shown previously in Figure 16. Arrows point to the direct A₄-C₅ traces resulting from (a) all tones having uniform duration and amplitude, (b) A₄ being lengthened by 12.5% and E₅ being shortened by the equivalent duration, (c) A₄ having a simulated 2.5-dB increase in amplitude, and (d) the combination of changes (b) and (c). Performance nuances can affect perceived melodic organization.

Fig. 20. Highlights of three simulated performances of the opening octave leap in the melody shown previously in Figure 16. As the tempo increases, first doubling (2X beats/min) and then quadrupling (4X beats/min), the motion traces become progressively more tenuous. The ability of the model to connect wide musical intervals is thus tied to tempo, with slower tempos allowing wider connections.
the figure, the tempo is doubled to 2x beats per minute (thereby halving durations), the G4–G5 traces begin to break up. A further doubling of tempo to 4x beats per minute, as at the right of the figure, increases the effect. And for each individual level-three-and-four combination, the breaking up of the trace is more severe than the figure indicates, since the figure shows a composite of several superimposed traces. Of course, this breaking up of the ability to connect wide intervals at very fast tempos is exactly what has been repeatedly demonstrated in studies of “trill thresholds” (Miller & Heise, 1950) and “temporal coherence boundaries” (van Noorden, 1975).

Figures 17–20 give equal visual prominence to the traces produced by every level-three-and-four combination, no matter how large or small the W. Such equanimity may not, however, conform to normal modes of Western music perception. In tracking the motion of voices in a Mozart string quartet, for instance, listeners might pay more attention to the normal interval sizes of scale steps and arpeggios than to tiny inflections of vibrato or to potential one-and-a-half-octave leaps. Whether a product of nature, nurture, or immediate context, this allocation of attention might profitably be represented in the motion traces.

Figure 21 attempts such a representation by implementing two principles. The first is that the sizes of W should be normally distributed about a moderately sized central value, chosen here to be 2.5 semitones (musically, a so-called neutral third, and psychoacoustically, a value near one critical bandwidth). The exact sizes of the 15 Ws displayed are (in semitones): 0.50, 1.00, 1.45, 1.95, 2.15, 2.30, 2.41, 2.50 (the central value), 2.59, 2.70, 2.85, 3.05, 3.55, 4.50, and 6.65. The second principle is that the heaviest lines should correspond to the central values of W and that progressively lighter lines should correspond to Ws at progressively greater distances from the central value.

Fig. 21. An alternative to Figure 17. Here sizes of W are normally distributed about a moderately sized central value: 2.5 semitones. The heaviest lines correspond to the central values of W and progressively lighter lines correspond to Ws at progressively greater distances from the central value.
distances from the central value. Though this scheme is admittedly clumsy, it does seem to produce in Figure 21 a musically reasonable interpretation of the performance described above for Figure 19d (here the final C₄ is also emphasized with simulated dynamic stress). The heaviest traces highlight the important neighbor-note pattern of C₄-B₄-C₅ on which the two apparent voices converge; the next heaviest traces follow these converging voices, with this particular performance giving a slight edge to the ascending A₄-C₅ over the descending E₅-C₅; light traces show potential connections across the opening octave, with these oscillating traces becoming more coherent as the up-and-down intervals contract; and finally, the entire pattern of traces provides a visual record of the resolution of conflicting primary and secondary melodic patterns—what is initially ambiguous and highly implicative becomes unitary and closed.

The van Noorden Demonstration: Pitch Streaming

Van Noorden (1975) describes a striking interaction between a static and a mobile stream of pitch. His demonstration, schematized in Figure 22, generates its static stream by rapidly repeating a 1000-Hz tone: on, off, on, off, ad infinitum. The demonstration generates its mobile stream by interposing, during the odd-numbered silences in the static stream, a different tone whose pitch rises slightly between every presentation. As a visual image, Figure 22 closely resembles illustrations of the Gestalt law of good continuation: the diagonal line representing the mobile stream appears to pass smoothly through the horizontal line representing the static stream. Listeners, however, do not hear the mobile stream pass smoothly through the static stream. Rather, the static stream seems to capture the mobile stream when it nears 1000 Hz. Moreover, the rhythmic sense of evenness that had prevailed suddenly vanishes, to be replaced by what van Noorden terms a “galloping rhythm” of three-tone groups: dum-di-dum, dum-di-dum. Only after the mobile stream’s pitch has passed through and moved well away from 1000 Hz—even further away than the distance at which capture occurred—does the mobile stream seem to break free and reintegrate as a separate percept.

Fig. 22. A visual representation of the Gestalt law of good continuation also serves to schematize the static stream (white squares) and mobile stream (black squares) of van Noorden (1975).
Though the change in rhythm during capture still awaits comprehensive explanation (and highlights a lacuna in the study of rhythmic perception), the motion-tracking model can simulate both the capture of the mobile stream by the static stream and the hysteresis effect that impedes the mobile stream from breaking free. Figure 23 presents the motion traces from a simulation of van Noorden’s demonstration. The rapid repetition of the static stream generates the horizontal line at mid-figure. The mobile stream generates the stairlike diagonal that approaches, is captured by, and then escapes from the static stream. Note that during capture, the two streams interact to produce a periodic inflection of the static stream, first below and then above its central frequency. This inflection of the central stream would seem to affect, in some unspecified way, the perceived rhythm. That is, the inflections in the static stream correspond to the unstable “di’s” in the dum-di-dum galloping rhythmic groupings.

In this model, the hysteresis effect—the seeming reluctance of the static stream to release the mobile stream—is not the product of any unique property of the static stream, though its higher level-three activation is a contributing factor (higher because sustained responses to the rapidly repeating tones of the static stream decay only a third as long as those for the sparser tones of the mobile stream and thus, when multiplied by transient responses, generate stronger onset and offset responses). Instead, the two prime factors causing the hysteresis effect originate in the mobile stream itself. The first can be metaphorically described as the “trailing edge of the

![Diagram](image)

**Fig. 23.** Motion traces from a simulation of van Noorden’s demonstration (cf. Figure 22). The static stream generates the horizontal line at mid-figure. The mobile stream generates the stairlike diagonal that approaches, is captured by, and then escapes from the static stream. Note that during capture, the two streams interact to produce a periodic inflection of the static stream, first below and then above its central frequency. The inflections in the static stream correspond to the unstable “di’s” in the dum-di-dum galloping rhythmic groupings perceived during the period of temporal coherence.
wave” created by the mobile stream. By “trailing edge of the wave” I mean the residual level-three Gaussian distributions of previous tones in the mobile stream. Though the first tone of the mobile stream always produces a symmetric Gaussian distribution at level three, when the first new (and higher) tone occurs, the combination of the two distributions becomes slightly asymmetric or skewed, with a sharper slope in the direction of the new tone and a more gradual slope trailing off toward the old tone. This asymmetry, which becomes more pronounced as more new and incrementally higher tones sound, is in itself sufficient to explain much of the hysteresis effect. That is, the interactions with the static stream will be asymmetric as well, preserving segregation as the steep leading edge of the “wave” approaches the central frequency and prolonging integration as the long trailing edge of the wave moves across and beyond the central frequency. As shown in Figure 24, where the tones are shorter than those shown in Figure 23 and where a wider W is used, the hysteresis effect can become quite pronounced. As the figure indicates, the distance from the perceived collapse of the mobile stream to the central crossing point (distance m) can be much shorter than the distance from that crossing point to the place where the mobile stream seems to become reconstituted (distance n). Another way to show the hysteresis effect is to overlay upon an ascending scale (such as the one in Figure 24) an analogous descending scale with its time line reversed (see Figure 25). The regions where the one mobile stream is intact but the other is fragmented define the extent of the hysteresis effect.

The second factor that can contribute to a hysteresis effect is the specific dynamics of level two. As described earlier, the output of level two is the product of a sustained response and a transient response. Should either

![Figure 24](image)

Fig. 24. In the van Noorden demonstration, the distance from the breakup of the mobile stream to the crossing point (distance m) is often shorter than the distance from the crossing point to the reintegretion of the mobile stream (distance n). The text suggests how this asymmetry may be caused by the dynamics of the mobile stream.
Apparent Motion in Music?

Fig. 25. If one takes an ascending mobile stream (such as in Figure 24) and overlays upon it an analogous descending mobile stream with its time line reversed, there appear regions where the trace for the one mobile stream is intact but the trace for the other is fragmented. These regions define the extent of the hysteresis effect for the particular parameters chosen.

response be very low, the product will be low. Very brief tones of the type used in van Noorden's demonstration may not allow either the sustained responses to rise very high or the transient responses to decay very low. When these brief tones end, the product of the two responses may not compensate for the sudden loss of input from level one, with the result that at level three there is a collapse of the "leading edge of the wave." As a result of such a collapse, the maximum of the "wave" will move slightly backward toward the still decaying traces of earlier tones, an effect evident in Figures 24 and 25. Notice that the level section of each stair step falls back to a position slightly lower than the left edge of each step. The result is, again, that segregation is preserved as the mobile stream approaches the static stream and that integration is prolonged as it moves through and beyond the static stream.

The just-mentioned tendency for a level-three maximum to shift slightly back toward previous maxima can, when staccato-like tones are involved, result in unexpected side effects. For example, if each pitch in the mobile stream is repeated once before the next higher pitch sounds, the motion traces will indicate that the mobile stream moves up in pitch at each tone. To help describe how this could occur, let me arbitrarily define two successive steps in the mobile stream: step $p$, at 800 Hz, and step $q$, at 900 Hz. When $q$ first sounds after $p$, the level-three maximum will shift momentarily from $p$ to $q$ but then drop back slightly to a point representing about 880 Hz after $q$ falls silent. When $q$ is then repeated, the combined factors of (1) the trace of $p$ having decayed further (making the lower tones less of
an influence) and (2) the recurrence of $q$ producing a somewhat stronger
level-two response both help to solidify a new maximum close to the point
representing 900 Hz. As shown in Figure 26, the simulation suggests that
incrementally ascending pairs of repeated staccato pitches can form a
sequence where, ambiguously, one may hear either repeated pairs of tones
or a linear process where each tone appears at least slightly higher than the
previous one.

Van Noorden’s emphasis was on mapping the boundaries between the
perception of fission (two separate streams) and temporal coherence (one
composite stream). So clear to him was the effect of attentional set that he
reported two boundaries: a fission boundary for subjects attending to sepa-
rate streams and a temporal coherence boundary for subjects attending to a
composite stream. The fission boundary was between one and three seis-
tones and changed very little as the tempo of the tones slowed. The temporal
cohesion boundary, by contrast, began at a value comparable to the fission
boundary but widened considerably as tempo slowed. Put somewhat more
simply, it appears that capture of the mobile stream occurs inevitably at a
narrow pitch interval no matter what the tempo, whereas the ability to
follow the dum-di-dum, up-and-down movement of a composite stream
depends strongly on how fast the tones are presented—slower rates allow
for wider intervals of apparent motion.

I have already discussed how tempo affects the ability of the motion-
tracking model to connect tones widely separated in pitch (see Figure 20).
Thus for any given $W$, there is a threshold of tempo above which temporal
coherence or the sense of one composite stream breaks down. For any
given $W$, there is also a relatively narrow distance from the central pitch at
which capture will take place, and this distance is only slightly affected by
tempo. So the behavior of the motion-tracking model simulates both of
van Noorden’s boundaries.

![Figure 26](image)

**Fig. 26.** A variation on the van Noorden demonstration. Here each pitch in the mobile
stream is repeated once before the next higher pitch sounds. The text suggests how listen-
ers, given staccato tones rapidly presented, might perceive either ascending pairs of tones
or an ascending series where each tone seems slightly higher than the previous tone.
Van Noorden (1975) points out that even beyond the temporal coherence boundary one can still hear a “weak connection” between the alternating tones.

It should not be thought that no interaction at all can be observed between the tones A and B in the region beyond the temporal coherence boundary, i.e. in the region of inevitable fission. In fact, an effect similar to temporal coherence can be observed here. In the tone sequence ABA, for example, a very weak connection can be heard between the tones A and B in the region. However, the tone B is still heard separately in the background, so that this is definitely a case of fission. At very large frequency jumps, this effect is no longer heard. We will not discuss it any further here.

From the perspective of the present motion-tracking model, van Noorden could be interpreted as describing the operation of multiple-scale Ws. For a given W—the W that is the focus of a listener’s attention—there will indeed be a temporal coherence boundary beyond which fission occurs. But for larger Ws, temporal coherence continues to function “in the background.”

Figure 27 shows the superimposed traces of van Noorden’s streaming demonstration as produced by 15 sizes of W ranging from a quarter tone to over a tritone (see above or Figure 21 for the exact sizes). The central line represents, of course, the static stream, and the ascending staircase represents the mobile stream. As the figure suggests, at a distance somewhat less than three semitones from the central pitch, the mobile stream is captured by the static stream even for the narrowest size of W. The figure

![Diagram](image)

**Fig. 27.** Superimposed traces of van Noorden’s streaming demonstration as produced by 15 sizes of W ranging from a quarter tone to over a tritone. At a distance somewhat less than three semitones from the central pitch, the mobile stream collapses into the static stream even for the narrowest size of W. At a distance somewhat larger than a major sixth, the vertical connections between the two streams begin to break up.
also shows that at very large interval sizes—the intervals at the extreme left and right of the figure represent octaves—the vertical connections between the two streams begin to break up. Between these extremes of necessary stream segregation and necessary integration is, when the multiplex system is viewed as a whole, an area of considerable indeterminacy. Presumably, however, a listener focuses attention on a specific size of W, creating thereby an interesting mix of strongly categorical perception at the Ws in the attentional foreground and a more gradual, statistical perception of state transition at the Ws in the background.

As a general prediction of how attention is allocated in a multiplex motion-tracking system, let me suggest that attention flows to the W producing the strongest, most coherent motion traces. Considerations of this principle lead far beyond the confines of this paper, but in the limited context of the van Noorden demonstration, it may be possible to specify some factors that would control such an allocation. First of all, in the case of fission or stream segregation, allocation will be driven by the height of the stairsteps in the mobile stream. If each step rises only a small fraction of an octave, as was apparently the case in van Noorden's work and has definitely been true in replications of his work, then attention will flow to a small W, perhaps to a W of approximately that step size. For example, given a step height of 1/10 octave and a W of 1/8 octave, capture of the mobile stream will occur near a distance 2W from the central pitch or, in this situation, 2.5 semitones from the central pitch, a value largely independent of tempo and consistent with van Noorden's results. By contrast, in the case of temporal coherence or stream integration, attention can sequentially shift to larger and larger Ws as the oscillating interval widens. The maximum possible W varies inversely with tempo, so the temporal coherence boundary will widen as tempo slows. Again this is consistent with van Noorden's results.

Trills

An important point of departure for van Noorden's work was the study of trill perception by Miller and Heise (1950). They demonstrated that stream fission increases when either the interval widens or the tempo increases. Of course the relation between tempo, interval, and the ability of the motion-tracking system (for a particular value of W) to connect two tones with a continuous, integrated trace has already been discussed. Yet the framing of trill perception in terms of stream fission or fusion can tend to create a binary opposition where, as is so often the case, additional factors may be at work.

An important factor that emerges from simulating the motion perception
of trills may be described as “motion contraction” (Grossberg, 1977). The
effect of motion contraction is evident in Figure 28, where the up-and-down
oscillation of the motion trace progressively narrows as the speed of the trill
increases. The extent of motion contraction depends on the particular set
of parameters for equations (1) and (2) and on the amplitude envelopes of the
alternating tones. The cause of motion contraction is tied to the effective
rate of decay for each tone. As the speed of alternation of the two tones
increases, each tone decays less before sounding again. At extreme speeds,
the tones decay very little and the maxima of level-three activations pro-
duced by their joint distributions move toward the midpoint between the
two tones, the position these maxima would occupy if the two tones were
played simultaneously as a chord. Trills, especially long trills of the type
classical soloists use to signal the end of a cadenza, often begin slowly and
then accelerate to a high speed of oscillation. A listener can follow the
alternating up-and-down motion up to a point where the percept becomes
that of something both madly active and securely stationary. This shift in
perception is illustrated in Figure 28 where, near the right of the figure, the
previously gentle up-and-down motion rapidly collapses into a jittery line.
Though stream fission accounts for the stationary aspect of the trill, it does
not explain the aspect of high-speed motion. Motion contraction accounts
for both aspects, since the motion trace continues to oscillate slightly even at
extreme speeds of trilling. Of course, neither explanation excludes the
other. Indeed, within a multiplex motion-tracking system, both fission and
motion contraction can occur in different parts of the system.

**Emergent Broader Streams: Parallel Voices**

Since the Middle Ages, treatises on the art of writing counterpoint have
consistently recommended making voices move in opposite directions, so-
called contrary motion. And just as consistently, they have cautioned against making voices move in the same direction, so-called similar or parallel motion. From a modern perspective, these admonitions speak to the tendency for voices with parallel motion paths to lose their autonomy and become melded into a joint (and perhaps fuzzier) percept (Huron, 1992; Wright, 1986).

Motion traces of voices moving in parallel are subject to the same effect noted when discussing motion contraction—the tendency for maxima of summed level-three activations to be located near the “center of gravity” of the tones involved. For two equally salient voices, that location will be their midpoint. For three or more voices, that location can vary for each size W, but will be somewhere in the middle of the tones involved. Thus, if output from a motion-tracking system is correlated with output from systems sensitive to pitch locations, then voices moving in parallel present a situation where pitch locations and motion traces may not coincide.

Fig. 29. The opening of the “Dance of the Reed Pipes” from Tchaikovsky’s Nutcracker ballet. The motion trace (W = major second) initially shows two streams: the melody and the combination of the two lower voices.
An instance of three-note parallel chords can be found in the opening of the “Dance of the Reed Pipes” from Tchaikovsky’s Nutcracker ballet. The motion trace (W = major second) in Figure 29 initially shows two streams: the melody and a combination of the two lower voices. The tracking of the notated melodic contour is to some extent fortuitous. That is, had Tchaikovsky slightly altered the seventh three-voice chord (G₄-A₄-E₅ in Figure 29) so as to retain the third-then-fourth spacing of the previous chords (A₄ changes to B₄, a modification barely noticeable to casual listeners), the upper motion trace would be drawn “off the track” to follow the descending line of D₅-C♯₅-B₄ rather than the melodic line of D₅-C♯₅-E₅, as Figure 30 illustrates.

The D₅-C♯₅-B₄ descending line suggested by Figure 30 is not easy to hear. A performance designed to bring it out would need to play E₅ pianissimo and B₄ fortissimo in order to overcome the strong tendency of the E₅

Fig. 30. At the beginning of the second measure, a change in the middle tone from A₄ to B₄ (cf. Figure 29) draws the upper motion trace “off the track” to follow the descending line of D₅-C♯₅-B₄ rather than the melodic line of D₅-C♯₅-E₅.
to be perceptually more salient. Explanations of this salience might appeal to special factors of register, masking, harmonic fusion, acculturation, saturation, and so forth. Yet musicians will attest that the highest and lowest voices in polyphonic textures are generally more salient no matter what the register, the harmony, or other factors. Student dictations of Bach chorales, for example, while often quite accurate for the soprano and bass, rarely capture more than bits and pieces of the alto and tenor. The thesis presented here is that this salience of outer voices is, at least in part, an automatic product of lateral inhibition, resulting in what vision studies term edge or boundary enhancement.

Let us take a four-voice, D-major chord, D₃-F♯₄-A₄-D₅, and examine how lateral inhibition at level four will make motion traces for the outer voices more salient than those for the inner ones. The underlying rule is that the mutual inhibition between any two traces varies inversely (here a Gaussian function) with the distance between them. Figure 31 illustrates how this rule would transform the relative salience of four originally equivalenced tones. The outer traces, those representing D₃ and D₅, become the strongest because whereas the inner traces are inhibited from both sides and by a close neighbor (F♯₄-A₄ forms the smallest interval and thus mutual inhibition is strongest within that interval), the outer traces are inhibited from only one side and from larger distances.

The exact spacing of tones and thus traces will determine the exact

![Fig. 31. One result of lateral inhibition between level-four motion traces could be the increased salience of outer voices, analogous to edge enhancement in vision models. The figure shows how lateral inhibition suppresses traces inhibited from both sides (F♯₄, A₄) more strongly than traces inhibited from only one side (D₃, D₅).](image-url)
Fig. 32. The particular voicing of chords can influence the extent of outer-voice edge enhancement. The figure shows how the same traces as in Figure 31 show a reduced effect of edge enhancement when they are more evenly spaced. Here, although the trace for the inner-voice D4 is more salient than that for the outer-voice D5, both outer traces remain more salient than their nearest inner-trace neighbors.

profile of imputed salience. For example, Figure 32 shows what musicians term an “open” voicing of the chord shown previously in Figure 31. Notice that the trace for the tenor voice (D4) now is more salient than in Figure 31 because it has moved away from the trace for the alto voice (A4). The trace for the bass voice (F#3), by contrast, is now somewhat less salient than in Figure 31 because it has moved up toward the other traces. Yet both outer traces remain more salient than their neighboring inner traces.

Adjusting the signal strengths of the tones in the Nutcracker example of Figure 30 in conformity with the edge enhancement predicted by Figure 31 eliminates the tendency for the motion to follow the descending D3-C#3-B4 line (see Figure 33).

As a final example (Figure 34), I present 15 overlaid traces of the Nutcracker excerpt (sizes of W are distributed about a moderately sized central value of 2.5 semitones and displayed as previously detailed for Figure 21; each trace assumes the edge enhancement just discussed). Only the very largest sizes of W can form coherent motion traces for the thirty-second notes that sweep up across the second measure. This sudden ascent overwhelms the system’s ability to track an instrumental part with any accuracy. In place of a well-defined, stairstep-like arpeggio, the system produces a fan of fragmentary traces that unfold across almost two octaves of pitch. The effect on the motion-tracking system seems in keeping with the aesthetic effect of a sudden and exuberant expansion of activity that Tchaikovsky’s music so charmingly evokes.
Fig. 33. A second simulation of the music of Figure 30 incorporating a version of the edge-enhancement effects shown in Figures 31 and 32. The upper motion trace now follows the notated melody at the beginning of the second measure (cf. Figure 30). Outer-voice salience overcomes the lure of between-voice proximity.

Discussion

The above model of apparent motion percepts in music can be profitably applied to a large number of research problems in both the psychology of music and music theory. I have touched on only a few of these. This is not to say, however, that the model is mature or without important limitations. I did not, as mentioned earlier, attempt to integrate this motion model with a model of pitch perception, even though I suspect the two domains are intimately related. And whereas the various motion traces shown in the above diagrams were determined algorithmically by simple calculation of local level-three maxima, the derivation of these traces through processes of lateral inhibition would be far from simple.

From the characteristic dynamics of similar competitive circuits one can
Fig. 34. Fifteen superimposed motion traces from a multiplex simulation of the Tchaikovsky excerpt. Sizes of $W$ range from a quarter tone to over a tritone. Only the very largest sizes of $W$ produce coherent traces for the thirty-second-note arpeggio that concludes the excerpt.

predict at least four important effects of lateral inhibition at level four. (1) There would be a practical limit to the number of surviving maxima for each size of $W$. Thus even if there were, say, 20 melodic streams as input to the model, the system’s output (for a given size $W$) could not produce coherent traces for more than perhaps 3 or 4 of those streams. This would be in keeping with the contrapuntal practice of traditional composers and with Huron’s analysis of the likely number of auditory streams in the works of J. S. Bach (Huron, 1989). (2) Streams that move in parallel motion will sum their effects on level three, with the result that parallel streams will emerge as dominant traces in the context of other, independent streams. Thus a musical shift from two independent voices to voices moving in parallel sixths, for example, would likely produce a shift of attention to the stronger, summed trace at a larger $W$. At the same time, limits to the
contrast-enhancing abilities of real-time lateral inhibition could make the motion traces of wide parallel streams less focused than the streams for independently moving voices. (3) Lateral inhibition creates a type of inertia; that is, feedback loops and other resonances within a neural field can work to delay a change of state. Thus the several abrupt shifts of maxima shown in the above diagrams might, in a more elaborate simulation, be somewhat smoother and more continuous. And (4), the effects of lateral inhibition might work to focus attention on a size of \( W \) just larger than half the mean size of the intervals actually occurring in a particular piece of music. Broader sizes of \( W \), with broader distributions of level-three activation, will, as mentioned above, be more difficult for lateral inhibition to focus into sharply defined maxima, with the result that the more broadly defined maxima will have a lower mean activation level. Conversely, at sizes of \( W \) too small to connect typical melodic intervals, the high number of independent maxima (an octave scale, for example, could produce eight static maxima rather than one mobile maximum) will again result in a lower mean activation level. As a result, if attention within the motion-tracking system flows to those maxima with the highest levels of activation, those maxima should be found at a size of \( W \) just larger than half the size of recently processed musical intervals (the exception being for streams moving in parallel, where their summed activations could shift attention to a larger \( W \)). If these suppositions are correct, then attention would automatically direct itself to “receptive field sizes” optimal for the music being heard.

For the sake of simplicity and replicability, all simulations were made assuming “tones” with instantaneous onsets and offsets, as in Figure 2. The model is, however, capable of taking input from real performances where pitch and intensity vary continuously. As Figures 17–20 suggested, the fine points of musical performance can influence a listener’s interpretation of melodic structure.

The motion-tracking system described above is assumed to be but one subsystem within the auditory system. Output from a motion-tracking subsystem might be used as input by other subsystems. Conversely, output from other subsystems might be used as input by the motion-tracking subsystem. Let me provide a few illustrative examples. First, some interaction seems evident between motion-tracking and rhythm-tracking subsystems in listeners’ perceptions of the van Noorden demonstration. Were this not the case, it would be extremely difficult to explain the radical change in rhythmic perception that accompanies the shift from two perceived streams to one stream. Second, a warping of the pitch continuum of the motion-tracking subsystem by output from some type of tonality subsystem might explain certain effects of common pitch alphabets (Deutsch & Feroe, 1981; Simon & Sumner, 1968) on perceived motion, in particular the sense that the steps of a scale or arpeggio, though objectively of varying
sizes, seem to be equidistantly spaced. And third, one subsystem might, so
to speak, fill in the blanks left by other subsystems. The final flourish in the
Tchaikovsky example (Figure 34) was almost uninterpretable in terms of
the motion-tracking of individual flute parts. Yet because all the tones in
the final flourish are members of a D-major chord, a harmony subsystem
could, with little or no help from motion tracking, give a coherent inter-
pretation of those fleeting thirty-second notes. By the same token, chords of
the “augmented sixth” (e.g., Ab-C-G#4 resolving to G-B-G), which can be
problematic for theories of harmony that emphasize root progressions,
have often been defined instead by their motion. Sessions (1951) noted that
“in some cases . . . the pull of the voices is so much stronger and more
perceptible than that of the root progression that the question of the latter
seems purely academic.” For such chords, as Prout (1903) said, “We must
seek another explanation,” which often takes the form of an appeal to the
motion of the voices. Apparently a clear understanding of what such
chords do compensates for not knowing exactly what they are.

Although the analogy between apparent motion in vision and apparent
motion in music is demonstrably weak when cast in terms of high-level cog-
nition, it seems to gain considerable strength when cast in terms of low-level
neural processes. Visual parameters such as stimulus contrast, size, lum-
nance, duration, color, and figural organization (Kolers, 1972) have natural
analogs in music (respectively, signal-to-noise ratio, bandwidth [or, assum-
ing early pitch determination, chord spacing], amplitude, duration, timbre,
and figural organization). And suitable timings for stimulus onset asynchro-
nies and interstimulus intervals in visual apparent motion are comparable
with timings for successive attack points and between-note articulations in
music. Seventy years ago, Wertheimer (1923) claimed that “quantitative
comparisons can be made regarding the application of the same laws in
regions—form, color, sound—heretofore treated as psychologically sepa-
rate and heterogeneous.” More recently, Bharucha (1987) has argued that
generic forms of neural information processing, rather than specialized, ex-
clusively musical forms, can account for data from a wide variety of psycho-
logical studies of music perception. The ease with which a neural model of
apparent motion in vision can simulate complex examples of auditory
stream segregation in music may lend support to both their contentions.

References

Bharucha, J. J. Music cognition and perceptual facilitation: A connectionist framework. 
Bregman, A. S. *Auditory scene analysis: The perceptual organization of sound*. Cambridge,
Bregman, A. S., & Campbell, J. Primary auditory stream segregation and perception of order


