The MIT Press

The Dysfunctions of MIDI
Author(s): F. Richard Moore
Source: *Computer Music Journal,* Vol. 12, No. 1 (Spring, 1988), pp. 19-28
Published by: The MIT Press
Stable URL: https://www.jstor.org/stable/3679834
Accessed: 29-01-2020 17:32 UTC

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/3679834?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

The MIT Press is collaborating with JSTOR to digitize, preserve and extend access to *Computer Music Journal*

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

F. Richard Moore
Computer Audio Research Laboratory
Center for Music Experiment, Q-037
University of California, San Diego
La Jolla, California 92093 USA

The Dysfunctions of MIDI

Introduction

The Musical Instrument Digital Interface (MIDI) is now a *de facto* standard for the digital representation of musical events. Actions of live musical performers are being "MIDIfied," commercial software is being based on MIDI-gated conceptions, and digital synthesizers are being slaved to MIDI masters. MIDI-based hardware and software is also proliferating at a tremendous rate, virtually ensuring that the characteristics of MIDI will play an important role in shaping a significant portion of future music. This paper concentrates on known dysfunctions of MIDI from a purely musical point of view, paying particular attention to performance capture, the digital representation of musical control processes, and synthesizer control.

Aesthetic Motivations

Human beings interact with each other acoustically in three basic ways: speech, music, and "other." While the boundaries among these three types of human interactions are not sharply drawn, all three types of sound interactions share the characteristic of conveying information from the source to the listener in ways that are only beginning to be understood.

Much of the information that passes from source to listener is nonverbal, even in the case of speech. We gain from the intonation pattern and microrhythmic variations of a person's speech information about the speaker's emotional state, place of origin, even the amount of sleep had the night before, as well as the flow of linguistic meaning.

If we—somewhat artificially—separate the *expressive* aspects of speech from its linguistically *communicative* aspects, we begin to identify an important human capacity that makes music possible. The purely expressive aspect of sounds makes music not only possible but desirable and perhaps even necessary as a human activity. While listening to music we exercise, explore, and refine our capacity to apprehend the expressive aspects of sounds. Our ability to do this well is often essential to our ability to survive, for often the truth of a human statement is conveyed far more by its intonation than by its "literal" meaning.

Before the advent of electronic music all matters of musical expression were basically in the hands of performers. Learning to play a traditional musical instrument is largely a matter of dealing with three basic issues:

Operation of the instrument itself
Customary and contemporary performance practice pertinent to the instrument
Literature available for the instrument

In short, we might say that learning a musical instrument requires the acquisition of control, musicality, and perspective. Composers generally write music with a keen regard for these issues, even if they are not necessarily proficient at playing all the instruments for which they write.

In most electronic musicmaking, however, expressive aspects of the sound are handled in a different manner. In non-real-time music assembly—sometimes called "sound sculpting"—all aspects of the sound are determined in advance of its audition by the listener. Great difficulty arises in non-real-time synthesis from the fact that the equivalent of performance nuance is limited according to how well it is understood by the assembler. Even when the assembler is a highly skilled performer, that understanding is likely to be far more limited than the performance skills acquired through years of practice. It is essential that human beings deal effectively with tasks and situations far more complex than those they can understand, since thought processing is far too slow to allow us to consciously

Computer Music Journal, Vol. 12, No. 1, Spring 1988,
© 1988 Massachusetts Institute of Technology.

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

control a walk across a room, let alone play a violin concerto.

A fundamental motivation for achieving a real-time performance capability in computer music, then, is to recapture this level of "visceral control" of musical sound—by which I mean a kind of control that includes both intuition and conscious and unconscious thought. Physical capabilities of human performers are simply too magnificent to be ignored altogether in any form of musicmaking. At least, creating completely prespecified music is a very difficult task. Any musically successful composition achieved in this way is a veritable monument, not only to the musicality but also to the patience and uncommon musical understanding of its creator.

But there is another even more fundamental reason why real-time performance control is desirable in computer music. We are acutely sensitive to the expressive aspects of sounds, whereby a performer is able to make a single note sound urgent or relaxed, eager or reluctant, hesitant or self-assured, perhaps happy, sad, elegant, lonely, joyous, regal, questioning, etc. The more a musical instrument allows such affects to be reflected in the sound spontaneously at the will of the performer, the more musically powerful that instrument will be. Consider the inflections of a human voice. Consider the intimate nuances of a violin. Consider the plaintive saxophone. Now let us consider a MIDI-based synthesizer.

Before proceeding with this analysis, I would like to point out that the very popularity of MIDI-based systems testifies to their utility and widespread acceptance. On the other hand that same popularity is responsible for a proliferation that makes it all the more important to understand any inherent dysfunctions in the MIDI control concept, so that they may be taken into account insofar as possible. MIDI is great. MIDI is good. Now let us examine what's wrong with it.

The Robustness of Expression

In speech synthesis there are at least four levels of synthesis quality that have been recognized for some time. Each level is associated with a range of information rates involved in the transmission of speech from the speaker to the listener.

The highest level of quality is high fidelity reproduction, in which the entire audible spectrum of the sound is transmitted as faithfully as possible. High-fidelity digital speech transmission today approaches passing—if not the Turing test—at least the "Ella Fitzgerald" test [this is a reference to a commercial for audio tape featuring Ella Fitzgerald—Ed.] wherein it would be extremely difficult to tell the difference between a live speaker and digitally encoded speech if both sources were carefully matched and some distance from the listener behind an acoustically transparent curtain. At this level of quality the transducers typically do more damage to the sound than the analysis-synthesis mechanism, which consists of 16-bit linear pulse-code-modulation (PCM) analog-to-digital and digital-to-analog converters running in excess of 40,000 samples per second for a total transmission bandwidth of more than 640,000 bits per second per channel of sound. At this rate, we hear virtually every expressive nuance that a speaker (or musical performer, for that matter) produces.

The second level of speech quality is toll-quality telephone transmission, typically achieved by eliminating frequencies above about 4 KHz and using 8-bit nonlinear quantization such as μ-law at sampling rates around 8 KHz. This yields a total transmission rate of about 64 Kbits per second. Over a good telephone connection we can not only understand the speech, but we also can easily detect affective sound qualities (such as the emotional state of the speaker) as well. In other words, we still know "who the speaker is" as well as what is being said. Of course music loses far more quality than speech over a telephone, but that is at least partly due to the fact that telephones are optimized exclusively around the properties of speech, and we certainly would not want to be in the position of trying to make music with something that is not up to that task. More sophisticated encoding and transmission schemes allow subjective speech quality to remain more or less the same down to transmission rates of about 32 Kbits per second.

Below 32 Kbits per second, more sophisticated

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

analysis-synthesis schemes such as sub-band coding, adaptive differential PCM, and linear predictive coding (LPC) allow speech intelligibility to be preserved down to rates of around 8 Kbits per second. The speech quality at these transmission rates begins to change in qualitative ways, however. We not only begin to lose information about the sound of the speaker's voice, but we find it increasingly difficult to recognize "who" the speaker is. Interestingly, losing knowledge of the identity of the speaker is tantamount to losing the affective qualities of the speaker's voice. We can no longer tell if the speaker is angry or tired—we can only recognize what words are being said.

Below 8 Kbits per second or so (such measurements vary enormously) there seems to be a "forbidden gap" in speech transmission that exists between the lowest transmission rate for intelligible speech and the transmission rate required for representing what the speaker is saying as real-time text on a terminal screen. If a speaker utters about 200 words per minute, an average word contains five letters, and ASCII codes are used to transmit the letters in the text, real-time text transmission requires roughly 133 bits per second (never mind the enormous amount of processing that the analysis procedure would require!). Perhaps a fast talker (an auctioneer) could sustain 1,000 bits per second or so. At this level of speech transmission, virtually all of the expressive information in the sound has been lost—we are simply reading! Of course the speaker might choose words allowing us to learn of a relevant emotional state, but the speaker's problem has now been reduced essentially to that of a writer.

This compressed review of speech synthesis demonstrates how the semantic content of language is relatively robust over changing information rates while the expressive or affective content is quite vulnerable to such changes. Herein lies a plausible explanation for the fact that while data compression techniques have been successfully applied to speech, such attempts have been largely unsuccessful for music. Music addresses a part of human perception and cognition that deals with affective qualities of sound to a greater degree than most speech. We might use a few kilobits per second to transmit a synthesized musical score to a real-time display, but that would make the listener's task one of imagining what it would sound like when realized by live performers, analogous to what we sometimes do when we read a play. We will return to this vulnerability of the expressive aspects of musical sound to low bandwidth transmission rates momentarily.

Control "Intimacy"

Music can evoke subjective reactions for which no words exist. Just as Eskimos allegedly have an extensive vocabulary for various qualities of snow, musicians have a specialized vocabulary for describing music in words. But just as you or I might have difficulty understanding the distinctions among the words in the Eskimo "snow lexicon" without firsthand experience, the vocabulary of musicians is mostly based on shared experience that is largely ineffable. Shorthand notations like *stretto* or *con tutta forza* merely act as reminders to those who already know how to achieve such effects, not as descriptions of how the effects can be achieved. Such issues are the subject of musical performance practice.

For subtle musical control to be possible, an instrument must respond in consistent ways that are well matched to the psychophysiological capabilities of highly practiced performers. The performer must receive both aural and tactile feedback (Cadoz, Luciani, and Florens 1984) from a musical instrument in a consistent way—otherwise the instrumentalist has no hope of learning how to perform on it in a musical way.

The best traditional musical instruments are ones whose control systems exhibit an important quality that I call "intimacy." *Control intimacy* determines the match between the variety of musically desirable sounds produced and the psychophysiological capabilities of a practiced performer. It is based on the performer's subjective impression of the feedback control lag between the moment a sound is heard, a change is made by the performer, and the time when the effect of that control change is heard.

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

The musical instrument with the greatest control intimacy is probably the human voice. A singer's vocalic control is largely innate and highly informed by speech as well as music. The range of musically desirable sounds producible by the human vocal mechanism is enormous—far greater than that commonly used in traditional singing as amply demonstrated in the research work of the UCSD Extended Vocal Techniques Ensemble (Kavasch 1980). We are so adept at apprehending affective qualities in the human voice that many listeners dispense altogether with a requirement to understand words while listening to musical voices.

Other instruments exhibiting large control intimacy include the violin, the sitar, the flute, and many others. With such instruments the microgestural movements of the performer's body are translated into sound in ways that allow the performer to evoke a wide range of affective quality in the musical sound. That is simultaneously what makes such devices good musical instruments, what makes them extremely difficult to play well, and what makes overcoming that difficulty well worthwhile to both the performer and the listener.

MIDI Control Properties

When a synthesist strikes a key on a MIDI keyboard, several bytes of information are transmitted serially over a typical duration of about a millisecond to a synthesis engine. (For good technical descriptions of the details of this process see [IMA 1983] or [Loy 1985].) Such information typically includes the number of the key that was struck and the velocity with which it was struck. If multiple keys are struck simultaneously, information regarding each key is transmitted in sequence over the MIDI connection. The more keys depressed at a given moment, the longer the transmission will take.

One of the fundamental assumptions of the MIDI concept is that these small delays introduced by serial transmission are either imperceptible or—if not exactly imperceptible—that they don't make any difference in a musical context. According to this criterion, we might say that if two musical events are as similar to each other as a live performer could make them then they may be considered to be "musically indistinguishable" even if they are not indistinguishable in a strict perceptual sense (Moore 1977). Music is obviously possible within the tolerances of human performers, so if a performer is incapable of striking a key within a millisecond of a desired time, why should it matter if the starting time of MIDI-generated events are off by no more than a few milliseconds?

While this assumption undoubtedly holds—at least well enough—in many practical musical situations, there are at least three ways in which this assumption also can be shown to be false. These are fundamental dysfunctions of MIDI as it is currently implemented.

Imperceptibility of Millisecond Delays

The first dysfunction results from the fact that in some musical situations, millisecond delays do matter. Human perception is a wonderful thing, and it can make life difficult for those who try to fool it. While people cannot reliably distinguish which of two events comes before or after the other when the time difference between event onsets is small (less than about 30 msec), time delays even smaller than a millisecond between successive events can be readily distinguished in the pitch/timbre domain.

If we listen to a sequence of sounds, each consisting of a pair of clicks separated by 0, 1, 2, 3, etc. msec, we readily hear a recognizable and predictable sequence of pitches. Classical psychoacoustic studies have shown our ability to identify musical instrument sounds to be strongly linked to their attack transients, in both the sense that we find it difficult to recognize the sound of a saxophone or a trumpet if the attack portion of the sound is removed, and in the sense that if the attack portion of one sound is grafted onto the nonattack portion of another then we are very likely to identify the sound according to its attack transient rather than its nonattack portion, even though the latter portion may last hundreds of times longer than the attack. If the click-pairs mentioned previously were used as the attack transients attached to sounds of

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

longer duration, the amount of delay between the note onsets would become an important determinant of the identity of that sound in the case of chords consisting of two or more notes. Such delays produce an effect that is equivalent to the confusion about "who the speaker is" (sound source identification) in the case of speech transmission at low data rates.

Uncertainty in the Amount of Delay

A second problem with the MIDI assumption is that there is an unpredictable amount of delay between the time a performance gesture occurs and the time it is communicated to the synthesizer. While this uncertainty is only on the order of a few milliseconds its "temporal smearing" of attack times means that the timbre associated with any synthesizer key can be context-dependent in ways not controllable by the performer.

Unpredictability in the delay between key depression and sound onset leads to perceptible changes in the character of the sound. This unpredictability thus determines the extent to which the performer lacks control over the precise nature of the musical sound, even with practice, and lessens the "intimacy" of the control. Tiny variations in the performance are not reflected in the sound under such uncertain conditions, they are thwarted—effectively prevented from having any effect on the music that can be controlled either consciously or unconsciously by the performer.

"Event" Orientation

MIDI is designed to report on musical *events* in a timely manner. While it seems fairly intuitive that depressing a key on a keyboard may be well-modeled by a report of its key number and velocity, many musical control paradigms appear to be more consistent with a model of *continuous variation* of some parameter, such as the length of the vibrating portion of a violin string when the performer uses vibrato, or the loudness variations in the sound of a single sustained note when a trumpeter plays a crescendo. Such variations can be handled—at least in theory—by the MIDI channel by transmitting a stream of discrete values that effectively "sample" the continuously changing control parameter.

The *sampling theorem* states that if new events can be transmitted at a continuous rate of about one per millisecond then a continuous parameter with variational frequency components of up to about 500 Hz should be representable without aliasing. The application of the sampling theorem is complicated in this case by the fact that the MIDI channel is shared among many simultaneous event streams. Taken together with transmission time uncertainty, this implies that the sampling rate is not steady, making a theoretical analysis of the effects of such sampling quite difficult. It is known, however, that even small amounts of "sample jitter" can degrade a digital recording significantly (Stockham 1971) since it is essentially equivalent to recording a signal on a tape that does not move at a constant speed (something like a very high frequency "flutter" in the recording process).

In order to understand these three dysfunctions—temporal smearing, temporal uncertainty, and sample jitter—more thoroughly, we must examine some of the underlying processes that represent the problem that MIDI is trying to solve. The first process is that of capturing what the performer is doing in real-time—the "performance capture" problem. The second process is that of controlling a synthesizer in real-time—the "synthesis control" problem. The third is the transmission of information captured from the performer to the synthesizer in real-time—the "control transmission" problem.

Capturing Musical Gestures

How much gestural information can a human performer generate? In other words, measured in bits per second, what is the information rate needed to adequately represent what a human performer actually does during a concert performance with a traditional musical instrument?

At this time, a definitive answer to this question is not available. It is, however, clear that the answer to this question will ultimately depend on the

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

two basic issues of *resolution* and *rapidity*. In digital audio terms, resolution refers essentially to the number of bits needed to adequately represent a single sample of a time-varying quantity, and rapidity refers to the maximum speed of variation (frequency) of the quantity being measured. Increasing the resolution lowers *quantization noise*, while increasing the rapidity increases the frequency response and lowers the susceptibility of a system to aliasing, which is a form of misrepresentation of the quantity being measured.

If we want to make a general system adequate for musical control information, we must at least take into account such matters as the overall value range of parameters that are to be controlled and the *just-noticeable-difference* (jnd) that a performer can control. Simple thought-experiments, while not definitive, can then be used to make at least reasonable estimates of the information rates needed.

For example, how much musical control information can a keyboard performer generate in a second? Assuming that there are 88 notes on a keyboard, we might start by considering how quickly notes can be played by a skilled pianist. I am able to play a glissando across all 88 notes of a piano in about half a second with one hand. This means that during the glissando the equivalent of MIDI events are being generated at an average rate of about 176 events per second. On the other hand, that speed can be quadrupled by using both hands, starting in the center of the keyboard, and playing two simultaneous glissandi from the center of the keyboard outwards in about a quarter second, resulting in an average information rate of about 704 events per second. With practice (and some hand protection!) one might actually be able to generate 1,000 events per second on a piano keyboard in this manner, which is about the maximum speed of MIDI events. (Of course, this example is both rather extreme and perhaps not very typical of most piano playing, but we are trying to understand how to make a general performance capture mechanism.)

A better way of approaching this thought-experiment might be to consider the smallest time interval over which a pianist has either conscious or unconscious control, for example when playing a grace note. Rapid grace notes can be performed at the piano by bringing down two fingers simultaneously with a wrist motion—if one finger is extended slightly more than the other it will strike the keyboard slightly ahead of the other. I can play successive pairs of notes on the piano at rates exceeding 10 pairs per second. To play repeated notes a key must go down and up once per note through a distance of about a centimeter or so, yielding a minimum average key velocity (for each key of the pair) of about 20 cm/s. By extending one finger in each pair by a minimal amount (about 1 mm), it can strike one key sooner than the other by an amount of time equal to the amount of time it takes an object traveling at 20 cm/s to travel 1 mm, or 5 msec. But the average velocity of 20 cm/s describes a back-and-forth motion which involves directional reversals at each end. At the endpoints, when the fingers reverse directions, the velocity must be momentarily zero, indicating that the maximum velocity must be considerably greater than 20 cm/s in order to maintain that average. If a triangular finger motion is assumed, then the peak velocity would have to be about twice the average, or about 40 cm/s. If a more likely sinusoidal motion is assumed, the peak velocity would have to be about $\pi/2$ times the triangular amount, or 62.8 cm/s, requiring a temporal resolution of about 1.6 msec to measure adequately.

Such numbers are of course only rough estimates, but they are reasonable ones. Try as I might, I cannot find a reasonable estimate of piano note timing that would indicate that a temporal resolution much finer than a millisecond would be needed to capture a piano keyboard performance on a per-key basis. It is clear that the MIDI transmission rate is just on the edge of this rough calculation for single notes. If, however, we consider the case of a piano chord in which a dozen or more keys are played simultaneously with two hands, we note that the time needed to transmit the data representing the note events is now about N msec (where N is the number of notes depressed) about an order of magnitude slower than the resolution needed for each key. This is the temporal smearing effect described previously and to which we will return later.

Another thought-experiment using the violin leads to a different result. A violin is a nearly ideal

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

example of an instrument whose control mechanism is not well-modeled by the event paradigm of MIDI. The continuously variable control functions of the violin are much more readily modeled as sampled time functions such as those used in the GROOVE system (Mathews and Moore 1970). To capture what a violinist is doing we need to represent both left hand and right hand activities.

To represent what the left hand is doing we might replace the idea of a key number with a representation of the pitch being produced by each string. We then can meaningfully ask: How many "keys" would be needed to represent violin pitch? Violinists can produce pitches with an accuracy that exceeds the jnd of human pitch perception, since the violinist is capable of tuning one note to another note by a process of zero-beating, which is accurate to within a small fraction of 1 Hz. If we assume that an average resolution of 1 Hz is sufficient, then each time the pitch changes by this amount (or more) a new MIDI note event would have to be generated in order to track the changes in pitch during, say, a vibrato.

A rapid, thick vibrato might be played at about 10 Hz, with a total pitch excursion on the order of a whole tone. Assuming once more for simplicity that the pitch change has a triangular shape, this would mean that pitch event information would have to be generated from −6% to +6% of the fundamental frequency over a duration of 1/20 sec, or 50 msec. Playing a note with a pitch associated with a fundamental frequency of about 1 KHz would then require about 120 pitch measurements every 50 msec, or about 2,400 events per second. To represent a 1 Hz change at the top of the violin range of about 4 KHz would imply a needed precision of about 1 part in 4,000, or about 12 bits per measurement. The measurement of vibrato in this way would then require a total information rate of about 2,400 × 12, or 28,800 bits per second—about the total information bandwidth available on a perfectly encoded MIDI channel. And we have not yet considered double-stops or bow tracking!

While MIDI transmission bandwidth may be large enough to fully represent one real-time performance control parameter, real musical instruments can easily exceed this rate by an order of magnitude or more. On the average, however, MIDI bandwidth is not too far below that required to represent the maximum information rate that can be generated by a human performer. My best guess is that it is probably only about three to four times too slow for single performers playing typical instruments, provided all we are concerned about is representing single events produced by a single performer, such as one-note-at-a-time melodies on a keyboard. Control of the synthesis process, however, requires a much greater bandwidth than this.

Controlling a Synthesizer

A synthesizer cannot accept MIDI information directly, since the meaning of a key number and a velocity value—or any of the other standard codes that MIDI represents music in terms of—must be translated into synthesis parameters such as oscillator increments, modulation indices, peak amplitudes, and so forth. The actual time-varying parameters that are needed to control sound synthesis have a much higher bandwidth than the performance control information generated by the performer.

A worst-case situation for synthesizer control bandwidth arises in the context of additive synthesis, in which the frequencies and amplitudes of a large number of building block components must be varied at or near the audio sampling rate. So far, only the simplest synthesis methods such as frequency modulation (FM) can be implemented in real time, where the simplicity is precisely in terms of the bandwidths of the control parameters. In other words, most of what is known about digital sound synthesis has yet to be made available to performing musicians in the form of practical real-time digital synthesizers. And even in the case of FM and its many variations, the actual shapes of intra-event control variations such as amplitude or modulation index control functions are not determined by the performer in real-time but are generated from prefabricated tables of information that are preprogrammed by the musician or at the factory.

These preprogrammed control functions, triggered in real time during live performance, are the actual determinants of the acoustic signal. They

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

can of course be affected in limited ways by MIDI control data generated by the performer—the entire amplitude envelope of a sound may be scaled by a number proportional to the key velocity, for example—but they are extremely difficult to control subtlely in real time. This once again leads to a musical result that is determined largely in advance of the performance (just as in the case of non-real-time synthesis), and acts as a barrier to the intimacy of the control of the musical sound that the musician can exercise during real-time performance.

Information Bursts Through Sluggish Channels

The reason for this barrier can be seen by considering what happens when we transmit "bursty" data serially through a channel that has a maximum transmission rate considerably lower than the burst peaks. According to our previous considerations, the transmission rate of a MIDI channel is between one and three orders of magnitude lower than the peaks of the information that must flow between a live performer and a real-time synthesizer in order to achieve perfect "control intimacy," wherein the synthesizer could be controlled in the same sense that a violinist directly controls the strings on a violin.

Figure 1 represents the rate of musical control information flow versus time from the performer to a musical instrument (this figure does not show actual measurement data, alas—real measurements of such parameters are sorely needed). This information is modeled as a random function with significant peaks, or "bursts" associated, for example, with the meter or rhythm of traditionally organized music. (Such peaks are partially responsible for the $1/f$ power spectral density measured in music by Voss and Clarke, for example [Voss and Clarke 1978].) The maximum information rate of the MIDI channel is shown as a dotted horizontal line in the picture. The exact relationship of the vertical position of this line relative to the musical control rate function is a matter for further research, but it is clear that there are many realistic musical situations in which the musical control information rate far exceeds the MIDI channel capacity.

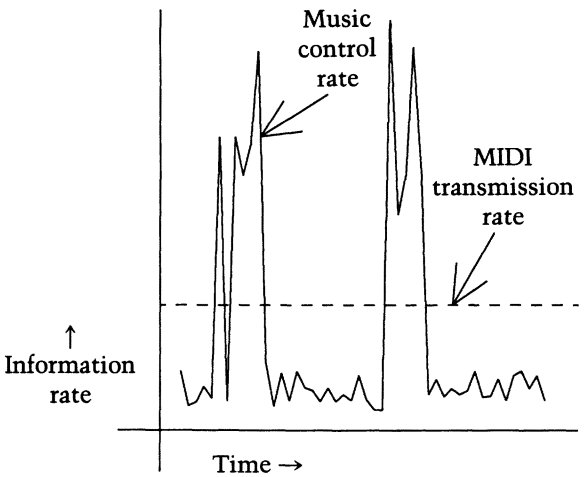

Fig. 1. Information rate of music control process compared with MIDI transmission rate (dotted line). Note the "bursty" behavior of the music control information.

Assuming that the MIDI information rate is fixed (i.e., that it may not be made higher than the burst peaks), then the information flowing out of the MIDI channel cannot accurately reflect the musical control information.

There are basically three ways to deal with the sluggishness of the MIDI channel, clipping, triggering, and smearing, all of which lead to a degradation in the control intimacy as defined previously.

Clipping

The musical information may simply be thrown away if it exceeds the allowable transmission rate for MIDI in the same way that an overdriven amplifier "clips" when its input value exceeds its maximum output value divided by the gain. Control intimacy is then reduced to the level that can be sustained by the MIDI transmission rate at precisely those moments when the most is happening in the music.

Triggering

The information for complex events can be precomputed so that it is stored inside the synthesizer and

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

simply triggered when it is needed. This approach—which is often used in actual MIDI-based synthesizers—lowers the control intimacy to that of trigger rates and values. The "internal liveliness" of the sound is determined in advance of the performance in ways that can be modified only slightly by the musician during live performance. Changing the microgestural quality of an impending musical event during the performance requires downloading of control functions "on the fly" during real-time performance. This process is complicated by the necessity to time-tag the downloaded information so that it will take effect only after the downloading process is complete (otherwise it may put the synthesizer into an illegal or undesirable state). Even if this is done the time it takes precludes it from being a practical way to achieve spontaneously expressive control over the microstructure of the synthesized sound.

Smearing

The triggering solution to the transmission bandwidth problem is noncausal, which means that it cannot be done entirely in real time. Another possibility that is causal is to allow the information rate to saturate at the MIDI transmission rate until the generated amount of control information has been transmitted. This causes information generated at time t to "spill over" into times following time t. We have seen this effect already in the manner in which N notes played simultaneously on a MIDI keyboard are transmitted serially over a duration of about N msec. Musical control intimacy is then degraded in ways discussed previously.

Conclusion

Real-time performance control is so desirable in music that almost any measure of it is welcome. However, one of the chief reasons to have real-time control is to allow performers to manipulate musical sound in ways that are tightly coupled to both what they are hearing and what they are doing. The principal effect of having a sluggish channel between the performer's actions and the synthesized sound is to decrease the sonic identity of each performed note, a process that I have here called a degradation in control intimacy. The result of this process is that triggered synthesized sounds seem rich but repetitive, while smeared synthesized sounds are not intimately controllable by the performer.

There are of course many other things we would like MIDI to do well besides provide a general communication link between performers and musical sound. MIDI can be criticized from the standpoint that it is not a true network, that it provides only one-way communications, and so on (Loy 1985). These are technical problems that have existing solutions besides MIDI. Computer musicians have only to decide which solutions to select from a large technical possibility space. The fundamental musical problem of expressive, intimate real-time control, however, is not addressed by improving the sophistication of MIDI along such lines.

If computers are to realize their potential as an *augmentation* rather than a *limitation* to the expressive means of music, we must not become confused about the extent to which the Musical Instrument Digital Interface solves the actual problems of real-time performance control. Much research is still needed to identify the types of practical solutions that actually exist to the musical control problem. In the meantime, we should not throw away a more general approach to sound synthesis in favor of a highly questionable solution to the real-time control problem.

References

Cadoz, C., A. Luciani, and J. Florens. 1984. "Responsive Input Devices and Sound Synthesis by Simulation of Instrumental Mechanisms: The Cordis System." *Computer Music Journal* 8(3):60–73.

International MIDI Association (IMA). 1983. *MIDI Musical Instrument Digital Interface Specification 1.0.* North Hollywood: International MIDI Association.

Kavasch, D. 1980. "An Introduction to Extended Vocal Techniques: Some Compositional Aspects and Performance Problems." *Reports from the Center* 1(2).

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms

Center for Music Experiment, University of California, San Diego.
Loy, G. 1985. "Musicians Make a Standard: The MIDI Phenomenon." *Computer Music Journal* 9(4):8–26.
Mathews, M. V., and F. R. Moore. 1970. "GROOVE—A Program to Compose, Store and Edit Functions of Time." *Communications of the ACM* 13(12):715–721.
Moore, F. R. 1977. *Realtime Interactive Computer Music Synthesis.* Doctoral dissertation, Department of Electrical Engineering, Stanford University.
Stockham, T. G., Jr. 1971. "A-D and D-A Converters: Their Effect on Digital Audio Fidelity." *Proceedings of the 41st Convention of the Audio Engineering Society.* New York: Audio Engineering Society.
Voss, R. F., and J. Clarke. 1978. "1/*f* Noise in Music: Music from 1/*f* Noise." *Journal of the Acoustical Society of America* 63(1).

This content downloaded from 155.33.16.124 on Wed, 29 Jan 2020 17:32:11 UTC
All use subject to https://about.jstor.org/terms