Information Theory and a Musical AI:
Toward an integrated information system

Jacob Elliot Reske

YALE UNIVERSITY

DECEMBER 2012

AMENDED JANUARY 2014

ABSTRACT

The purpose of this paper is to apply the basic principles of Information theory to the field of artificial music cognition— its problems, some solutions, and the role of computers as the final stage of a fully informed information system. Some of the earliest writings on Information Theory and aesthetics— specifically, Claude E. Shannon's *A Mathematical Theory for Communication,* as well as Abraham Moles' *Information Theory and Aesthetic Perception*— are just as relevant for computer cognition today as they were when they were published. As the focus of music cognition shifts from the role of the transmitter to that of the receptor (the computer), both Shannon and Moles' original texts can provide valuable insight for solving the biggest issues of creating a generalized music AI today. Three basic problems of making a musical AI will be explored— the need for a representation system that can handle the system's entropy, the necessity of reassembling a transmission contextually, and the role of an aesthetic arbiter outside the system— problems which correlate strongly with those of communication systems decades before. The goal of viewing these problems in this way is to use Information Theory principles to unite various disciplines in music AI under a larger task: creating a machine that can receive sound just as an individual would and, ultimately, extend further the basic human process of distinguishing signal from noise.

TABLE OF FIGURES

1. INTRODUCTION

> *"There is a fundamental semantic gap between low-level audio signal and high-level human perception."* (Yang and Chen 3)

So begins "Music Emotion Recognition," a survey by Taiwan-based AI researchers Yi-Hsuan Yang and Homer Chen that is one of the first of its kind: a systematic survey of techniques for analyzing large music databases with computers. Their frustration is one echoed by musicologists, software developers, and composers of A-life music; when computers are put at the receiving end of a string of communication, their attempts at deciphering it so are often woefully inadequate. The "high-level perception" that humans have requires a set of complicated cognitive processes, all of them designed to perceive the world in a conceptual way. In music's case, the sound's instruments, timbre, harmony, cultural associations, and emotion are part of a series of complex interactions, of which "meaning" is the byproduct. Each of these serves as a kind of semantic-based "signal," transmitting concepts fluidly between the musician and a listener at the other end. When computers fail to receive these concepts, the profundity of human cognition is brought directly to the forefront. Clearly, Yang and Chen know this; having an AI try to "listen" to music properly makes that semantic gap feel even wider.

Yet this semantic gap is nothing new; if anything, AI research merely amplifies a problem that is inherent to any information system. In fact, Yang and Chen's fundamental problem is not so different from another "fundamental problem" of communication: maintaining the integrity of a message, especially when noise is present. This problem was articulated sixty years ago by Claude E. Shannon, whose paper, *A Mathematical Theory for Communication*, single-handedly spawned the field of Information Theory. In his introduction, he writes:

> "The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages." (Shannon 1)

For Shannon, a model was needed to approximate where distortion of the message— noise— was entering the signal, and how best to eliminate it; the integrity of the message itself was most important, rather than what it necessarily meant. And, for most early information theorists, the semantic concepts of the message were only secondary to ensuring the quality of the message from transmitter to receiver. Yet when the receiver is a machine (as with an AI) and the act of receiving a message must approximate how a human would respond, the meaning of the message is crucial to its reception. Information Theory's universal dialectic between "signal" versus "noise" can be powerfully applied to computer cognition: "accurate" versus "noisy" interpretation of musical signs. Perhaps a new focus on approximating human understanding of signals can fill in the final gap in the system— in the words of Information Theory, the "reception" of the message. Moreover, with this knowledge, this closed system could have a powerful purpose: as a processing tool for determining signal and noise, an augmentation of the very foundation of our communication through music. Modeling human processes of signification could not only inform us on the validity of these signs, but also introduce new significance that we have not yet found ourselves. All of this could make AI the final, significant

piece of an almost noise-free system, one that could utterly transform the noisy act of musical perception.

2. INFORMATION THEORY — BASIC PRINCIPLES

Before considering how an AI would be used in such a way, it is important to outline the workings of the system itself and its relevance to a musical AI. Our system will be based heavily on the basics of Information Theory, a field that studies the communication of messages— any concept that can be represented as signal— from a source to a receiver. For this relatively new field, the moment of catalysis was in Claude E. Shannon's publication at Bell Labs, *A Mathematical Theory for Communication*; for the first time, the concept of transmitting information was given a series of concrete, mathematical models. In his first paper— one that would provoke response in dozens of other disciplines— Shannon outlines the basics of what he would call a "general theory of communication," suggesting a system that could reduce the signal-to-noise ratio drastically. (Shannon 2) The core of Shannon's system lies in the way information is defined: a quantity measured by the probability of it occurring in a sequence of events. Unlike a continuous system, each piece of information— the signal that needs to be transmitted, is measured by a *bit,* the smallest unit by which information can be stored. Once the number of bits per piece of information is calculated, the capacity of each channel (number of bits transmitted per second) can be deduced. (3). With information his basic quantity, Shannon can create an entire system that encodes, transfers, and decodes sequences of information. Three parts are essential: a *transmitter* for quantizing, encoding, and preparing the message from its source, a *signal* that (in theory) should be identical to its *received signal*, and a *receiver* that "ordinarily performs the inverse operator as done by the transmitter." (2) His accompanying diagram (*Fig. 1)* would outline the basic schema for a generalized communications system:

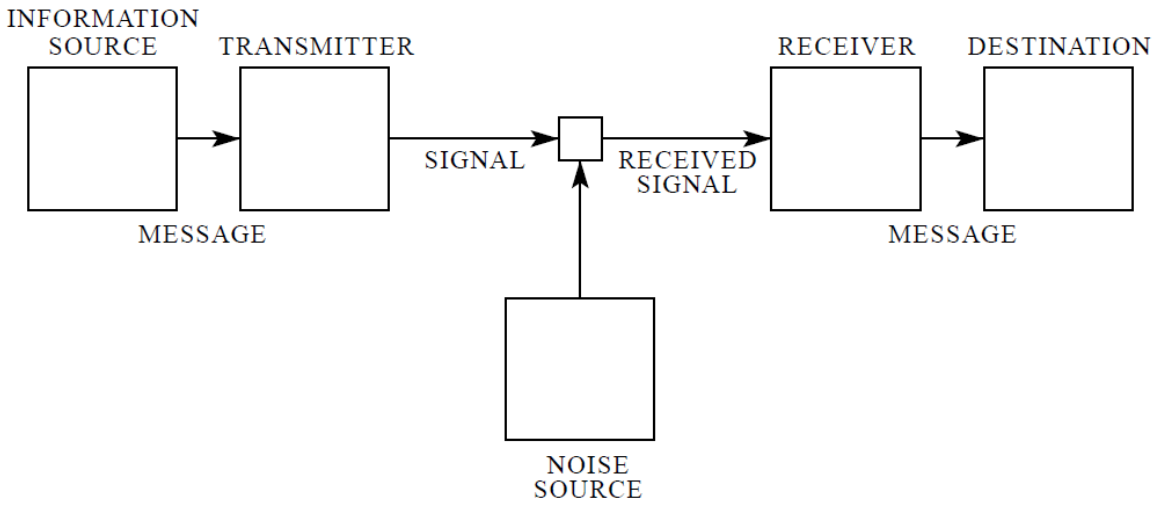


Fig. 1 — "Schematic diagram of a general communication system."[1]

[1] Reprinted from the 1998 edition of Claude E. Shannon's "A Mathematical Theory of Communication," p.2.

As the bit moves from transmitter to receiver, it serves as the smallest unit of information, and its variability (entropy) will determine the complexity of the message received. Shannon is quick to make this distinction: the amount of information (variability in a probabilistic environment) is dependent on the entropy *H,* a continuous function of the probability of each event $[H(p_1, p_2, \ldots\ p_n)\ \forall\ p_i \in \sigma]$. (10) The more equal the probabilities $p_{i,j}$, the more entropic the system is; less determinacy in a system will require that each bit has more information, as per Shannon's definition. As the signal's redundancy increases— its inverse, or lack of information— it becomes easier to predict the next bit of the message via Markov or n-gram models. (6) The beauty of Shannon's theory, however, is in its generality; it constructs hypothetical channels that can transmit all messages efficiency, regardless of the value *H*. Ideally, each channel is message-agnostic, limited only by the *n* possible discrete bits of each piece of information. Hence Shannon's "Fundamental Theorem for a Noiseless Channel:"

> *Theorem 9*: Let a source have entropy H (bits per symbol) and a channel have a capacity $\mathcal{C}$ (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{\mathcal{C}}{H} - \varepsilon$ symbols per second over the channel, where ε is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{\mathcal{C}}{H}$. (Shannon 16)

Furthermore, this system can be applied to both discrete systems (where each piece of information consists of "a sequence of symbols" that are then translated by the receiver) and continuous systems (where continuous functions transfer data, such as from electrical impulses) (Ash 46). Most information theorists are concerned only with the first of these two cases, since any stochastic process with *n* finite combinations can be rendered as discrete, even "continuous information sources that have been rendered discrete by some quantizing process." (Shannon 3). Musical information, for example, is usually a continuous function rendered discrete; a computer waveform, for example, is simply a collection of highly accurate, discrete samples of a continuous wave function. More importantly, discrete systems are highly desirable because they have the possibility of being noiseless— that is, the receiver will always reproduce the exact discrete signal of the sender. The "noise source" of Shannon's diagram, then, is far less dangerous to the integrity of the message. If noise does present itself in a discrete signal, the difference between the output and the intended message will be arbitrarily small, as long as the noise can be observed and corrected with capacity of at least *H*:

> *Theorem 10:* If the correction channel has a capacity equal to $H_y(x)$ it is possible to so encode the correction data as to send it over this channel and correct all but an arbitrarily small fraction ε of the errors. (Shannon 20)

So ends the topic of noise for Shannon's paper, whose conclusions are arguably one of the greatest harbingers of the Digital Revolution. With the advent of discrete processors orders of magnitude more powerful than those at Bell Labs, Information Theory became a representation of preservable, lossless communication, and noise was its great opposition. For music, the days of a "noiseless" communication system have practically arrived; uncompressed digital files can sample sound at rates far beyond our capacity to hear the difference. As quantization errors become nonexistent and the signal-to-noise ratio is reduced, music listeners have come to expect that the signal from a transmitter will reach their ears with little (if any) discernable difference in quality. Compressed formats, such as CD's and .mp3's, have made massive strides toward maximizing signal-to-noise, using Markov techniques for noise correction that were first outlined

in Shannon's paper.[2] But the system is not perfect: it is explicitly designed for messages transmitted to *human musicians*. No attempt is made in Shannon's model to consider a closed system, where a computer can process the information sent as well as transmit it. Yet this is precisely the problem that today's AI specialists are interested in exploring: can computers be designed to interpret the same messages that they transmit?

3. THE COMPUTER AS A RECEPTOR

With this question, one sixty years after Claude E. Shannon launched the field of Information Theory, noise remains the greatest impediment to achieving a fully integrated information system. But as the focus has shifted to developing a cognizant computer, so, too, has the source of the noise. In an AI-integrated system, computer scientists are no longer interested in the noise introduced between the transmitter and receiver, as in *Fig. 1*, but in noise introduced as the receptor is analyzing it. For music, the cause of this noise is that same "semantic gap" between the musician and machine; computers are capable of analyzing audio with a level of detail far beyond human concepts, but they are expected to estimate at the level that the musician does.[3] Thus, to teach a computer to make informed judgments is to construct general concepts that are deduced from larger ones, synthesizing large sets of data into a smaller, more manageable output.

Before considering the problems of creating such an AI, however, the concept of what constitutes "noise" versus "signal" for a computer deserves some clarification. For classical Information Theory, noise is simply an unwanted differential between transmitted and received signal, introduced somewhere before the signal is decoded. (Ash 231) Moles generalized this concept for all sound, noticing that "there is no structural difference between [sonic] noise and signal… they are of the same nature." (Moles 78) Thus, a noise is "a signal that the sender does not want to transmit," – or, if the signal comes from the environment, "a sound that we do not want to hear… a signal we do not want to receive." (79) If an AI program is the receptor, however, the output serves as confirmation that the concepts of the signal was transmitted intact. Essentially, for the system to transmit data correctly, a musician must inform the AI about the basic concepts to be analyzed, and noise is anything outside the intended message. Moles' definition can be generalized further: for artificial intelligence, *a noise is a concept that we do not want the computer to receive.* It is a misinterpretation of the signal, a gap between the intended message and the AI's decoding of it. The burden of success, then, has to be placed on the software that analyzes the selected sound source.

This allows us to discuss the basic problem: Music listening is a highly complicated cognitive process, synthesizing dozens of disparate concepts almost instantly and deriving meaning from all of them. In the words of Information Theory, music information is transmitted on multiple channels— pitch, timbre, rhythm, instrument type— all of which are encoded in low-level audio and have to be decoded. Simply defining a "generalized" form of representation

[2] CD-ROMs, for example, employ Shannon's "Fundamental Theorem for a Discrete Channel with Noise" to correct errors in playback in real time, even when the disc has minor scratches. (Sergio and McLaughlin 502)

[3] In *Music and Artificial Intelligence,* Chris Dobrian points to three different types of musical error: 1. "motor error" (a musical performance is not as accurate as the standards used to measure it); 2. "conceptual error" (the musician makes highly informed conceptual generalizations of a piece of music); 3. "intentional error" (the performer of the music often operates outside the systems that the music is made, such as *rubato* or pitch bending). (Dobrian 7)

has been a major problem for programmers. There are thousands of parameters by which humans perceive music, all of them within the context of the music's genre, year written, location, and cultural practices. As Roger Dannenberg points out, "Music is distinguished by the presence of many relationships that can be treated mathematically, including rhythm and harmony, [but] there are many non-mathematical elements such as tension, expectancy, and emotion." (Dannenberg 20) Furthermore, musical concepts also often apply only to a very small subset of genres; when algorithms are applied to music outside the genre, they often fail immediately. When writing an algorithm, programmers must ask what *type* of music they intend the program to listen to, rather than taking a general approach. "Feature and attribute selection," argues Eleanor Selfridge-Field, "is the single most important aspect of a scheme of representing music… if one attribute for a particular purpose is not present, the representation cannot be used for the intended application." (Selfridge-Field 569)

To make matters worse, some musical concepts remain poorly defined, or described with words that have close to the same meaning. In *Music Emotion Recognition,* researchers Yang and Chen describe this as the "ambiguity versus granularity" issue: affective words such as *"calm/peaceful, carefree, laid-back/mellow, relaxing"* are more or less synonyms, but their similarity to each other varies drastically from subject to subject. (Yang and Chen 6) Describing timbre, on the other hand, raises the granularity problem: human language has a limited set of words for the vast complexity of timbre, few of which are useful for approximating "what amounts to an infinite series." (Selfridge-Field 570) For this reason, most electronic music representations have historically not quantified timbre, due to practical reasons (not enough channels) or lack of a need. Western music notation, for example, has no way of notating timbre beyond technique text or expression markings; if the intended use is for music printing, the limited spectral range of MIDI is sufficient. Some representation programs, such as MIT's *Csound*, have tried to include timbre in instrument definition (Bainbridge 111); however, such representations are either limited in use or too generalized.

Often, the consequence of this is that computer receptors[4] are made only for specific, quantifiable modes of representation— restricted "almost necessarily… to the parametric model of music perception." (Dobrian 6) Thus, some elements of music listening are usually prioritized in music AI, primarily because they are a) easy to quantify, or b) apply to a wide range of music. Pitch and harmony, for example, are easily quantifiable and often figure prominently in levels of analysis. Rhythm usually figures prominently in multi-level models. But meta information— lyric content, methods of recording, etc.— are often conspicuously left out of these models. Yang and Chen's five large parameters, for example, are features in "…energy, rhythm, temporal, spectral, and [harmonic]" domains. (Yang and Chen 35) Research into instrument identification has opened the possibility of doing further work with timbre, adding another level of cognizance to the computer's arsenal. But the great task is constructing an ideal receptor, one that, like a musician, can recognize all of these attributes and process them, either at once or sequentially. An example of such a system is outlined in *Fig. 2*; it adapts concepts from Robert Rowe's analysis diagram in *Machine Musicianship,* while accounting for new methods of analysis, such as advanced spectral analysis:

[4] To stay consistent with Information Theory jargon, the term "receptor" will henceforth refer to an AI receptor— the cognition of music by a computer.

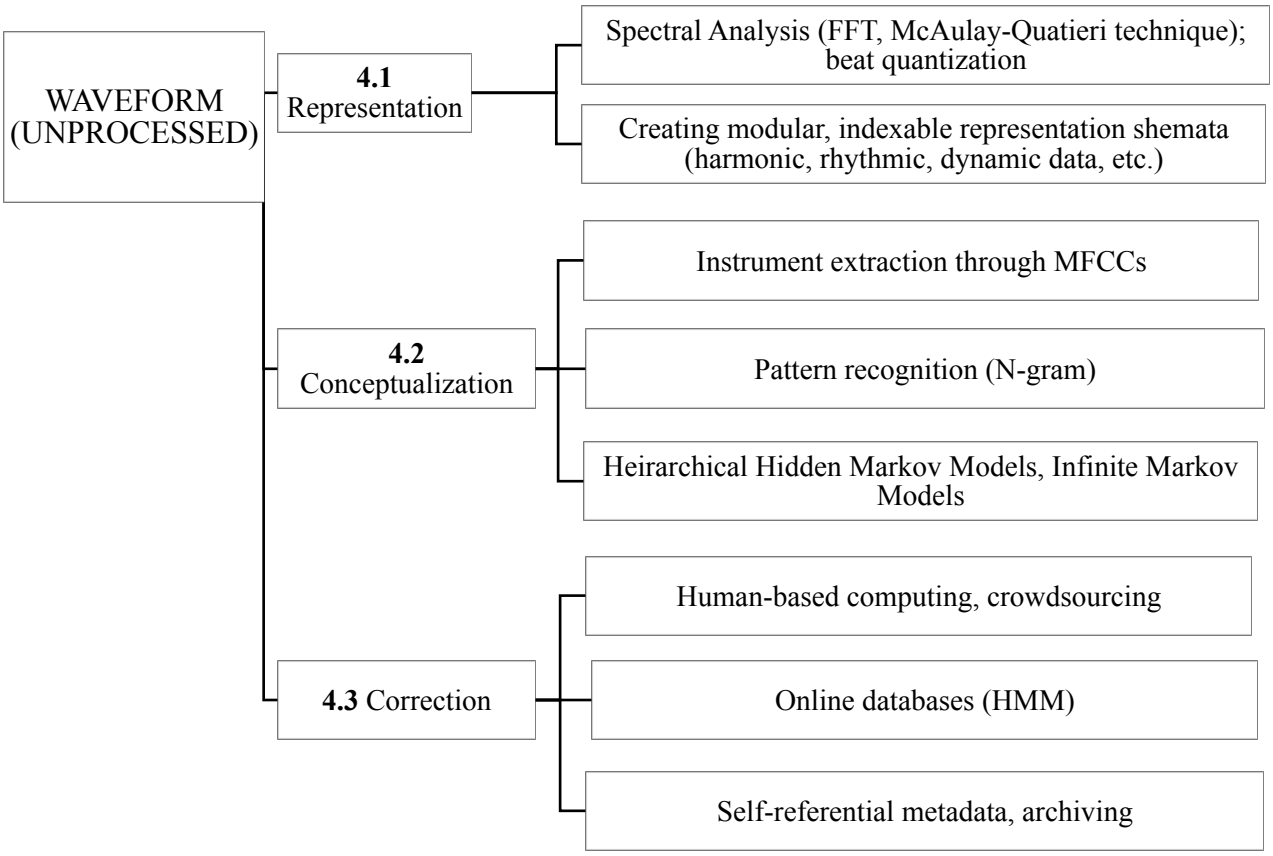


Fig. 2 — Anatomy of an ideal receptor.[5]

Note that the order with which these processes are used is open for debate; as Chris Dobrian points out, "Computer scientists tend to concentrate on a specific model of intelligence: reference to knowledge bases and stylistic behavioral scripts, probabilistic decision-making…. Intelligent musical behavior, [however]… indubitably involves use of more than one process simultaneously or sequentially." (Dobrian 28) But the basic framework for the process is there, integrated into a Shannon-like communication system. Because of this, the initial principles on which Information Theory is built— entropy versus redundancy, intelligible form versus informative output, recovering a signal lost in noise through probability (Moles 76-77)— are almost directly applicable to the biggest problems of making a musical AI. The following section will apply the principles of Information Theory to three major issues in computer-based musical cognition: the issue of a general representation system, the task of assembling larger concepts, and the need for an aesthetician outside the system. The solutions to these problems are by no means exhaustive, nor are they necessarily the most heavily researched issues by musical AI scholars. Each of them, however, are designed demonstrate how Information Theory's most basic principles can predict the problems that such a system will face— and, in some cases, inform computer scientists as to a workable solution.

[5]"Representation" section adapted from "Figure 1.2: Machine Musicianship Process," (Rowe 12)

4.1. REPRESENTATION

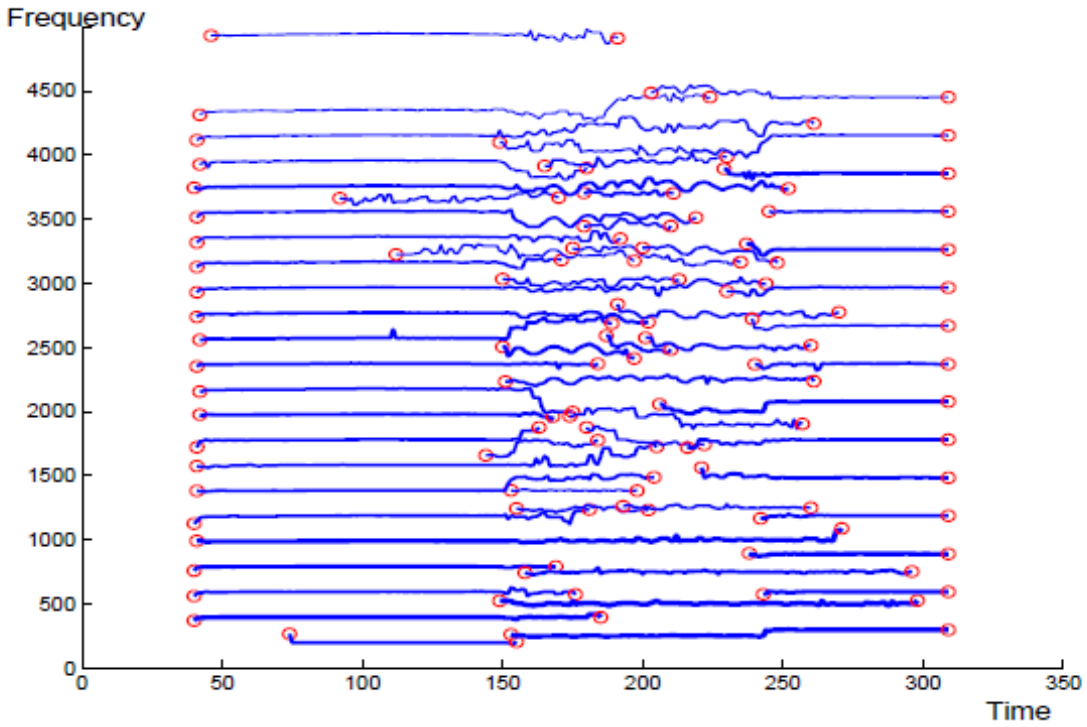
Representation must be the first step of any computerized music receptor. In a musical context, its primary function is the separation of the input— the waveform, or low-level audio signal— into any number of independent data sets. These sets are analogous to Shannon's channels of information: simultaneous ranges of data, each with a limited capacity. Ten years ago, programmers would have found the prospect of inducing rhythm, tempo, or key signature *without* some direct input (e.g. MIDI) daunting. Since then, techniques for spectral analysis have improved drastically, allowing low-level audio to be analyzed and grouped by sound source extremely convincingly. The process is an extremely effective example of analog noise reduction: first, the sound source synthesized using some combination of the McAulay-Quatieri or sinusoidal approach, creating a deterministic portion made up of only sine waves. Next, the stochastic portion of the signal is found by subtracting the sine-based synthesis from the original waveform, which is then put through a noise filter to recover the transients of the original file.[6] At this point, advanced partial tracking can match notes to their partials, and an output can be created that fully analyzable. (Lagrange, Marchandy and Raultz 4) This sine + noise + transients model has only recently become viable for polyphonic music; popular programs such as Melodyne Editor have demonstrated that this method is both stable and highly applicable.[7] For the first time, true representation schemata can be constructed simply from the waveform alone— a powerful step toward a generalized form of representation.

Of course, representing music by its pitch content is just one of any number of representations in which music can be stored. Dynamics, types of instruments, meta information, and beat structure are all valid methods of representation, and all of them require algorithms for problems as complex as the one above. Displaying this information accurately to a computer can also be an issue where program biases are introduced; as mentioned previously, different parameters might be more valuable to one genre of music over another. The rhythmic content of modern recordings of 12th century chants, for example is necessarily less important than its pitch or lyrical content. For this reason, most programmers long ago abandoned the idea of a general music representation system; Robert Rowe, for example, concedes that "trying to develop a comprehensive representation that can mean all things to all projects often leads to a concoction that serves no one." (Rowe 30) Extensibility— the ability to add specific modules to existing systems— is very important for customization, but Rowe concedes that an extensible system often lacks the communication between its parts necessary to draw larger conclusions:

> "Most systems are extensible, but they all become cumbersome when they begin to seem like centipedes— with too little core to support a large array of extensions and too few links between extensions to provide an integrated logical foundation for understanding the music as music." (Rowe 31)

Traditionally, the popular MIDI protocol has done exactly that: provided a simple platform that sacrifices channel capacity for extensibility and some customization. For many applications, such as transfer of controller-based messages to and from instruments, MIDI is sufficient. Yet for

[6] See https://ccrma.stanford.edu/~jos/sasp/Sines_Noise_Modeling.html
[7] "Melodyne" is a registered trademark of Celemony Software; see http://www.celemony.com/



musical analysis, the problems of MIDI— lack of timbral information, fixed channels, little support for extensibility, etc. — far outweigh its benefits. (Dobrian 6)

Fig. 3. Sample partial tracking of a violin tone using MAQ method[8]

Information theorists would argue that the problem of representation is one of channel capacity. If a single protocol (channel) needs to represent every conceivable permutation of music, the signal represented will be far too entropic to recognize patterns. In this case, the "noise" is any information in the protocol that is irrelevant to its function. If MIDI is an analysis tool, for example, any information about the overtone series will only serve as noise, since MIDI has no capacity for information beyond fundamental pitches.[9] Abraham Moles provides an analogy for the problem: to reduce the background noise of an electric channel, it is necessary "…to diminish what is called the "pass band" Δf of the channel, that is, the range of frequencies it transmits." (Moles 84) But to do so limits amount of information that the channel can contain, meaning that the signal in the information contained must be extremely redundant. It is only by limiting channel size that the confidence interval of that channel can increase.[10] This is a conclusion that comes immediately from Shannon's original text, and it applies directly to the problem of AI representation. The spectral analysis case above is a perfect application of this

[8] MAQ: McAulay-Quatieri. Image credit: (Lagrange, Marchandy and Raultz 4))

[9] "Audio-to-MIDI" modules, like those in Max or PD, are extremely imprecise for this reason: as a channel of information, MIDI stores no information about partials above pitches, and every harmonic on the spectrum is dependent on pass band filters to get a proper signal. This is in stark contrast to partial-tracking techniques, as evidenced on p.9 (See also Rowe, 47)

[10] Moles then goes on to propose his own "Uncertainty Principle of signal," a bastardization of the Heisenberg Uncertainty Principle that is both inaccurate and extremely misleading. The evidence, however, remains valid, regardless of Moles' hopelessly wayward conclusions

principle. Partial tracking is essentially a highly sophisticated form of noise removal: preparing the signal by reducing the amount of information, so that later algorithms can expect more redundancy. Simultaneous, redundant information channels, each with their own noise filters and specific parameters, can create a system of compatible databases that are both compatible and equal in importance. Algorithms can be highly specialized and reference other modules, or else run the same input through multiple passes with different levels of covalence. A harmonic algorithm, for example, can assume different types of redundancy (i.e. chord types) based on an instrument identifier, rhythm quantization, or metadata from the file itself. Working with interactive and/or open APIs could create a shareable, modular architecture for music analysis, letting application-specific solutions communicate with ones met for different purposes.

While not yet widely adopted, a "modular" representation system like this is not unheard of. Rowe, for example, mentions Dave Huron's *Humdrum Toolkit*, a system that uses a common protocol, but encourages "breaking up the representation problem into smaller, manageable schemes." Each representation tool can be added onto other, user-created schemes, all of which can realize "a great variety of analyses, including a great many never foreseen by the author of the tools." (Rowe 31-32) Other researchers have explored the idea of a hierarchical data exchange; Lorenç Balsach, for example, proposed another, hierarchical system of exchange. "At the lowest level, only 'note' information would be provided; at the next level information about modifying signs; …at the [next] level, their positions…" (Selfridge-Field 572) While never implemented, this type of system could be useful for making more informed decisions about genre or other higher-order concepts. Perhaps some combination of these two approaches— between hierarchical categories for analysis and an open, modular system— could provide both flexibility and interoperability for different genres of music. Whatever the solution, using a modular system of low-entropy channels would be a good application of Information Theory principles.

4.2. CONCEPTUALIZATION

Using good tools for representation— rhythm/harmonic tracking, instrument separation, etc. — is one thing; constructing large, conceptual categories from those data sets is another entirely. Once the receptor has an adequate representation schema, it can create "concepts," entire categories that are independent of human input, all based on strings upon strings of conditional probabilities. The use of Hidden Markov/n-gram models for finding commonalities is one of the hallmarks of Information Theory; while they had been researched exhaustively before Shannon's paper,[11] his use of Markov Chains to represent discrete sources foreshadowed a resounding adoption in communication fields. With them, Shannon could define entropy *H* in terms of uncertainty, the probability of the next bit occurring given the previous sequence of bits. In terms of our receptor system, Markov models can be used to find connections among separate concepts, or between subsets of concepts, making them ideal for use in a modular representation system. More importantly, they give the computer the ability to make choices, opening the door to conclusions that could be unexpected, noisy, inaccurate, or quite possibly very original. These conclusions can work in tandem with a set "'knowledge base' of aesthetic values determined by the programmer," or they can be entirely different from the programmer's intention. Says Chris

[11] In *A Mathematical Theory of Communication,* Shannon himself references Fréchet's paper on HMMs, published some 10 years before. (Shannon 8)

Dobrian, in his paper on Music and Artificial Intelligence: “By defining and programming *new* functions— as opposed to merely imitating functions which humans already perform— one may enhance the composer's or instrumentalist's operations in ways previously unheard of, actually expanding the number of abilities at that person's disposal.” (Dobrian 2)

However, in the spirit of Information Theory, our receptor has been designed to reach conclusions that are comparable to the ones that the transmitter intended. At the same time, programmers cannot simply program aesthetic judgment into a computer. This would deprive the receiver of choice, besides being impractical and “reflecting the biases of the programmer.” (Dobrian 16) Ideally, this system should be able to approximate *human* conceptualization: making decisions based on the covalence of small, perceptual elements. But if a computer does this, a mediation problem surfaces. A receptor that makes too many statistical parallels without conceptualizing can lead to an overwhelming or irrelevant (noisy) output; one that makes too few can lead to lost concepts.

Early Information theorists— Abraham Moles among them— have already considered this problem. Moles turns to human cognition; he observes that the human ear, though capable of distinguishing hundreds of thousands of elements per second, chooses instead to process only a few. (Moles 90) Other cognitive psychologists confirm this: there is an inherent, hierarchical superstructure to which the senses defer, despite being able to transmit much more information than necessary. Moles continues further:

> “We know that a maximum rate of apprehensible information, an “apperceptual limit,” dominates perception. In reality, *to perceive is to select…* One may consider the whole nervous system as a machine for selecting increasingly sketchy outlines of the richness of elementary stimuli.” (Moles 90)

Perhaps “everyday perception” can provide a clue: it is almost instantaneous and extremely hierarchical, and it operates through “dynamically rising... the difference thresholds.” (92) Since “the signals [of perception] are variable in time, and in fact change rather rapidly, the difference thresholds increase considerably… [This] requires sensations which stand out better as the elementary signals become briefer.” (92) In short, human conceptualization is essentially a form of dynamic noise reduction. Hundreds of noise filters are placed sequentially on the sensual data to construct objects, which are then filtered repeatedly until they approximate a concept. Perhaps, then, teaching a computer to approximate this process requires a highly stratified way of what is inherently an approximation.

Recent improvements on Hidden Markov Models allow computers to do just that, and the extent to which they can improve music AI programs is only now being realized. In fact, this principle of stratified, hierarchical data is a strategy that proved successful for speech recognition programs, and it is only now being applied to pitch structure algorithms. (Yang and Chen 149) Hierarchical Hidden Markov Models (HHMMs) have “multiple levels of states which describe input sequences at different levels of granularity.” (Weiland, Smaill and Nelson 2) This allows the AI to construct groupings that, in turn, can be analyzed as grams in a larger HMM system. Because “music can be seen as a composition of hierarchical structures built in time,” such a system is very useful for analyzing beat structure, pitch classes, phrase structure, and form in larger contexts. (2) A team at the University of Edinburgh, for example, used HHMM analysis to analyze the soprano voice of a series of Bach chorales, hoping to teach the AI the principles of melodic contour. From this data, plenty of trends were found, such as a preference for stepwise

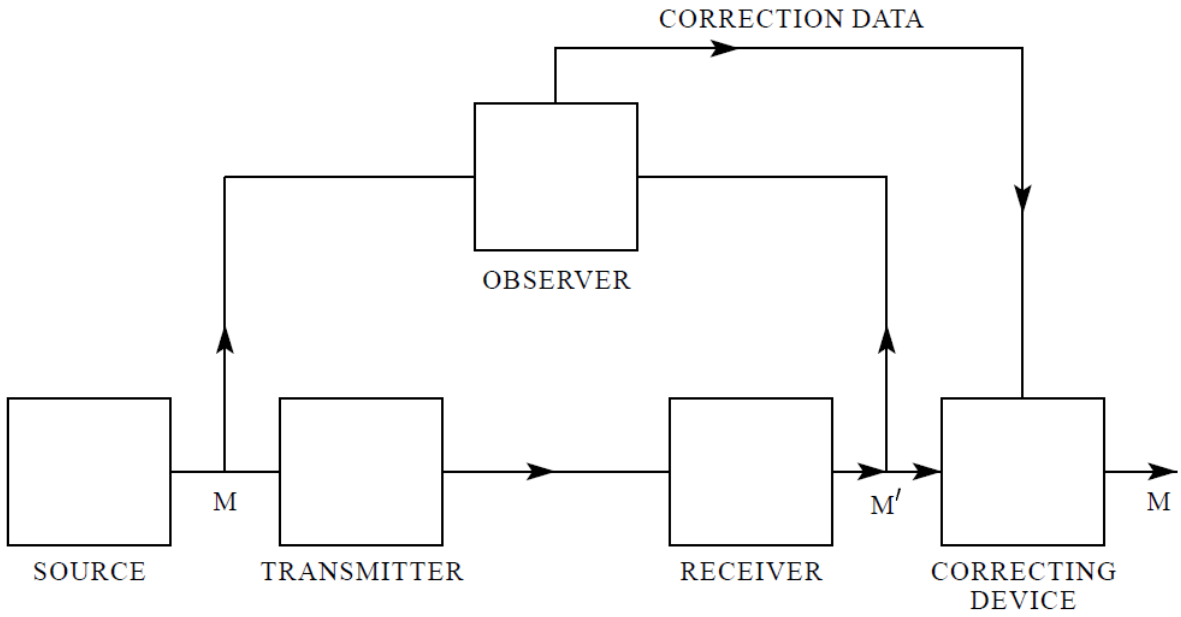
motion in the opposite direction of a preceding leap— a well-known trend in the melody of Bach's chorales. (6) Their conclusion found that HHMMs were extremely useful for stratified pattern recognition, but they conceded that they "are only powerful if their potential is used in the right way, i.e... if appropriate model structures are chosen to represent certain aspects of musical knowledge." (7) Finding representations that are both applicable and highly specific, then, is crucial to constructing hierarchical relationships. Since that publication, more research has been done into so-called "Infinite HHMMs," a generalized HHMM with infinite groupings. Success with speech recognition could pave the way for its use in musical analysis. (Heller)

4.3. CORRECTION

We are now approaching a point where, simply by analyzing a waveform, computers can jump to large structural conclusions that are both sophisticated and highly applicable. Researchers Yang and Chen, for example, use low-level classifications to determine the genre of a piece of music, in the hopes that there is some valence between musical properties and genre. (Yang and Chen 202) Their goal is not so different from the one proposed here; eventually, their team hopes to recognize some relationship between the musical language and its effect on emotion. (22) Along the way, they use algorithms on low-level audio to reduce the margin of error, or else balance errors made through imperfect (noisy) algorithms. Ultimately, however, the results of their analysis have to be checked against *human-made* concepts— emotions attached to music, or words for timbre— that are generally very imprecise. In our model, any noise in the system (where the computer's conclusions differ from human concepts) will naturally be very apparent at this stage. If, for example, a computer is asked to find music that has the same melodic phrase as a Beethoven piano sonata, or songs that are of the same genre and use a Moog synthesizer, there will have to be some human element outside the model to check and see if the set of songs is accurate. The act of *correction*— a member outside the system that adapts to changes— is essential for an accurate and relevant musical AI.

Luckily, the process of signal correction is built directly into Claude Shannon's model for a noisy channel, and his solutions are very applicable to the problem. In some cases, "the [discrete] signal does not always undergo the same change in transmission," unlike channels where "the received signal is a definite function of the transmitted signal." (Shannon 19) Essentially, as long as the receptor is performing the exact inverse function of the transmitter, the signal can be noiseless. Computer algorithms, however, can approximate (but never fully interpret) the intensions of the audio transmitted to them, since the transmission of meaning is naturally asymmetric. If the channel is noisy, then, reconstructing the original message with *certainty*" within the system is not possible. (20) The solution, then, is to have an observer— like the observer in our receptor model— who can see both the transmitted and received signal and make corrections, based on the entropy of the signal. Hence Shannon's Theorem 10:[12] the correction channel simply needs to send the positions of the noisy data to the receptor with a capacity of at least $H_x(x)$— the entropy of the original signal— in order to correct all of its errors. Even if the correction channel has a capacity smaller than $H_x(x)$, then the entropy of the corrected channel, $H_y(x)$, can be as small as $H_x(x) - \mathcal{C}$, where $\mathcal{C}$ is the capacity of the correction channel. (22) This is perhaps the most incredible of Shannon's conclusions about

[12] See page 5.



noise: even in a noisy channel, where the correction channel cannot correct quite as quickly as errors are made, the resulting signal will only have an arbitrarily small amount of noise.

Fig. 4 — "Schematic diagram of a correction system." [13]

What does this mean for the final step of our AI-based receptor? Because the unprocessed channel is naturally noisy (interpretive algorithms are not perfect), there has to be some member outside of the system that can see both the input and output of the signal. A single human being is much too slow; even if the entropy of the main signal is very low, human correction would take enormous amounts of time. Perhaps humans can play a role in data cleansing, though not in the direct sense; methods that are more oblique could provide the necessary speed of correction, while still allowing humans some input. *Music Emotion Recognition*, for example, mentions using user-generated data in the form of online games. Users listen to a song and write down the emotion that best fits it, generating data ripe for HMMs and AI correction. (Yang and Chen 174) Other correction methods use online databases to approximate real human computation, while providing more speed and versatility. Some online music databases, such as Allmusic.com, offer an open source database of songs by musical genre, instrumentation, and emotion— three large concepts that can be cross-referenced with the AI's deduction. (Yang and Lee 1) The approach could be symbiotic— an AI could cleanse the data of user-generated databases, while other databases could be cross-referenced to provide necessary corrections to the original one.

This brings to mind one of the fundamental problems of low-level audio signal, a corollary to the problem outlined by Yang and Chen on p.3: *Musical files encode nothing about the context with which an audio file is received*. Basic metadata— lyrics, date encoded, instrumentation, and recording techniques— are usually absent from the tags of popular formats such as .mp3 and .wav. This is a fundamental fault of the transmitter, not the receiver: for many formats (.mp3, .wav, .aiff), more complex audio metadata is unstandardized and rarely attached. When music is made with computers, however, the potential for metadata is enormous. Meta tags

[13] Reprinted from the 1998 edition of Claude E. Shannon's "A Mathematical Theory of Communication," p.21.

could be added every time that a file's waveform is modified, detailing the process of modification— a kind of continuous changelog for audio processing. Their presence could provide a useful database for correcting the AI dynamically and without actual human processing. Researchers could take advantage of the vast amount of music being made today in digital audio workstations, creating a codec for rendering audio that integrates such a changelog into every "bounced" waveform. This alone could provide a wealth of information, all speedily accessible and ready to be cross-analyzed by an AI correction channel. The corrector, then, is part of the file itself— the noise reduction self-perpetuating, by turning the metadata of an audio file inward on itself.

4. AWARENESS (CONCLUSION)

Furthermore, such a system could help to further two underlying goals of AI research: awareness and reduction. A computer's ability to understand of musical structures can give it tools to make connections that its creators might not have otherwise seen. Our robust receptor can approach music analysis through small, distinct modules; it can analyze those modules as strains of commonality in a larger structure; and it can quickly correct noise when its conclusions are not entirely logical. Once the communication system is robust enough, hundreds of applications could be made, combining the semantic understanding of a human being with an AI's immense computational power. Some of these have been implied here: computers could provide connections between pieces of music that are both logical and exhaustively researched, filtering out the pieces that the listener (or researcher) does not want to hear. Composition tools can use this analysis to work interact directly with the composer, providing its own analytical complements to the musician in real time. Focus on automatic metadata embedding could retain the information lost in reducing a piece of music to a simple waveform, all while archiving it for future analysis. All of these highlight the biggest benefit of a complete information system: an AI can dynamically augment the processes that humans take for granted, and on a far larger scale than they could ever attempt manually.

The most important benefit, perhaps, is that the noise of this system is completely quantifiable. The points where musical concepts break down, or become either inconsistent or difficult to understand, is completely demonstrable with an AI model. Having a computer try to interpret musical concepts can give evidence to those concepts that have meaning, and those whose meaning is ambiguous. All of this can give the AI another important skill, one that can be applied in every aspect of human-machine interaction: the ability to distinguish objectively between subjective signal and noise. Up until now, we have considered the AI as a receptor of a small, Information Theory-based system, where computers attempt to decipher semantic messages in the same way as the (human) transmitter. But consider the greater applications of such a system: if such a receptor can be built, computers could approximate— and possibly predict— new concepts that humans will create. As Abraham Moles defines it, the "sounds that we do not want to hear" can be filtered out entirely, leaving only those concepts which are most important for the individual— the signal. Moles admits that, in the case of aesthetics, "there is *no* absolute structural difference between noise and signal… The only difference which can be logically established between them is based exclusively on the concept of *intent* on the part of the transmitter." (Moles 78-79) But as the "semantic gap" becomes narrower and narrower, that ambiguity of intent could be significantly reduced, simply by parameterizing the idea of meaning

entirely. This entire system could simply be the perfection of an automated “correction channel” in a larger schema. If computers can predict the concepts and reactions of humans, surely they could be involved in a larger movement toward noise reduction.

All of this may seem like distant speculation, but the necessary tools are rapidly falling into place, and some of these applications may be realized much sooner than anticipated. Claude Shannon’s 1948 paper predicted the benefits of a discrete, noiseless communication system, well before processors were fast enough to realize these advantages. The problems of musical AI research— representation, conceptualization, and correction— practically mirror the problems faced by early discrete communication. Significant improvements in representation codecs, efficient hierarchical models, and metadata need to be made first, but the methods of solving these problems are not so different from basic problems in communication. The “inflection point,”— the moment where a sophisticated musical AI can interpret a piece of music with an arbitrarily small level of error— is nearing closer, one not unlike the inflection point foreseen by Information theorists sixty years ago. To get there, the gap left by communication’s largest barrier— noise— must be the primary focus.

WORKS CITED

Ash, Robert. *Information Theory*. Courier Dover Publications, 1965.

Bainbridge, David. "Csound." Selfridge-Field, Editd by Eleanor. *Beyond MIDI*. Center for Computer Assisted Research in the Humanities, 1997.

Dannenberg, Richard. "Music Representation Issues, Techniques, and Systems." *Computer Music Journal 17* (1993).

Dobrian, Chris. "Music and Artificial Intelligence." (1993).

Heller, Katherine. *Infinite Heirarchical Hidden Markov Models*. Cambridge: University of Cambridge, 2009. Document.

Lagrange, Mathieu, Sylvain Marchandy and Jean-Bernard Raultz. *Tracking Partials for the Sinusoidal Modeling of Polyphonic Sounds*. Cedex: LaBRI, Université Bordeaux, n.d.

Moles, Abraham. "Informnation theory and Aesthetic Perception." *Music Educators Journal, vo. 53* (1966).

Rowe, Robert. *Machine Musicianship*. Cambridge: Massachusetts Institute of Technology, 2001.

Selfridge-Field, Elanor. "Beyond Codes: Issues in Musical Representation." Selfridge-Field, Edited by Elanor. *Beyond MIDI*. Center for Computer Assisted Research in the Humanities, 1997.

Sergio, Verdú and Steven W. McLaughlin. *Information theory: 50 years of discovery, Volume I*. IEEE Press, 2000.

Shannon, Claude E. "A Mathematical Theory of Communication." *The Bell System Technical Journal* (1948).

Weiland, Michele, Alan Smaill and Peter Nelson. "Learning Musical Pitch Structures with Heirarchical Hidden Markov Models." *Journées d'Informatique Musicale*. La Maison des Sciences de l'Homme Paris Nord, 2005.

Yang, Dan and Won-Sook Lee. *Music Emotion Identification from Lyrics*. Ottawa: University of Ottawa, n.d. Document.

Yang, Yi-Hsuan and Homer H. Chen. *Music Emotion Recognition*. CRC Press , 2011.