Multimodal Technologies and Interaction

Perspective

Representations, Affordances, and Interactive Systems

Robert Rowe

Music Technology Program, Department of Music & Performing Arts Professions, New York University, New York, NY 10012, USA; robert.rowe@nyu.edu

Abstract: The history of algorithmic composition using a digital computer has undergone many representations—data structures that encode some aspects of the outside world, or processes and entities within the program itself. Parallel histories in cognitive science and artificial intelligence have (of necessity) confronted their own notions of representations, including the ecological perception view of J.J. Gibson, who claims that mental representations are redundant to the affordances apparent in the world, its objects, and their relations. This review tracks these parallel histories and how the orientations and designs of multimodal interactive systems give rise to their own affordances: the representations and models used expose parameters and controls to a creator that determine how a system can be used and, thus, what it can mean.

Keywords: representations; interactive systems; algorithmic composition; generative systems

Citation: Rowe, R. Representations, Affordances, and Interactive Systems. *Multimodal Technol. Interact.* **2021**, *5*, 23. https://doi.org/10.3390/mti5050023

Academic Editor: Insook Choi

Received: 19 January 2021
Accepted: 24 April 2021
Published: 1 May 2021

> "Man is a stream whose source is hidden. Our being is descending into us from we know not whence. The most exact calculator has no prescience that somewhat incalculable may not balk the very next moment".
>
> (Ralph Waldo Emerson *The Over-Soul*)

1. Introduction

Music is an inherently multimodal experience. Sight, sound, movement, and community all converge to create the lived experience of music, often with active participation by all. Indeed, many cultures have no word for music that represents an activity distinct from movement, community, and ceremony [1]. However, in every culture, music as it is experienced in time, illustrates the inextricable links between mental and physical responses and understanding [2].

Many artists today use the digital computer as a tool for the generation of musical material, sometimes in the context of a live performance. Widely used tools for developing such processes include Max/MSP [3], Pure Data [4], and OpenFrameworks [5]. An interactive music system is one that changes its behavior in response to musical input [6], typically in real time and as a response to a human musician or other such systems. As their development has evolved, many artists have incorporated other input signal types, yielding multimodal interactive systems. Some works track a visual scene, for example, or sensors attached to dancers to detect movement patterns [7]. There is a flourishing community of artists creating interactive installations, where the input to a music generation system may consist entirely of input from visitors to a museum or gallery, including motion, touch, and sounds [8–10].

In this paper, we examine how the concept of representation threads through the fields of algorithmic composition, cognitive science, and artificial intelligence, and how such representations directly impact the scope and capabilities of generative, interactive art works as well as the tools used to make them. Representation here indicates a set of symbols or patterns of activity that "stand for" or indicate the existence or operation of something else. In the field of cognitive science, including its sub-field music cognition, human thought is presumed to be a form of information processing, where subsequent stages in a processing chain are connected by increasingly abstract representations of the outside world and a subject's relationship to it: "thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures" [11] (p. 10).

This conception of representation and processing maps nicely onto the design of most interactive music systems. For example, Carlos Guedes's work with dance uses information from a video camera that captures the physical movement of a body in time [7]. The function of a digital video camera is to transduce patterns of light into pixel values. The changing pixel values are then a representation of the light patterns—a raw, sampled representation that is a computer's closest approximation to the underlying stimulus. A similar process converts sound waves into sampled amplitude values, and following our understanding of sampled representations, both can be extensively treated using digital signal processing.

Such approaches are multimodal in their combination of sensorial modalities to guide interaction between a human and a machine. However, we may also consider different modalities of thought—in particular, between high-level, symbolic representations and the rule-like computations they afford, as opposed to raw signals feeding into networks that converge on a state able to recognize salient differences. This contrast—between leveraging domain knowledge over a system of symbols and learning classifications and operations through exposure to examples—has been a primary axis of design focus throughout the history of artificial intelligence and has become a critical research question today, particularly as the tools of AI are adapted to the use case of algorithmic composition. Here, we will consider how the development of multimodal interactive systems could contribute to the decades-old question of how best to harmonize symbolic domain knowledge with purely data-driven models.

2. Symbolic and Sub-Symbolic Representations

A traditional musical score is a classic example of representation—musical events are recorded as a sequence of notations that stand for pitch class and height, time signature, bars, beats, sub-divisions, phrasing, and so on. Here we would speak of a *symbolic* representation: the alphabet of the representation encodes high-level concepts that are already a synthesis or abstraction of raw stimuli such as the arm movements or audio waves discussed above. Accordingly, the raw samples coming from primary sensors are referred to as *sub-symbolic* representations. The audio samples of a recording have no bars, beats, or notes explicitly encoded, only raw numbers representing changes in air pressure.

The earliest history of algorithmic composition consists of symbolic music generators. For example, Gottfried Michael Koenig's *Project One* was a set of procedures that generated notes, dynamics, instrumentation, and other symbolic outputs that could be assembled into a notated score for human musicians to perform in the traditional manner [12]. Other symbolic generation systems from Iannis Xenakis and Lejaren Hiller are well known. In writing about the development of his thought, Koenig observed: "The analytical task—given the music, find the rules—is reversed: given the rules, find the music" [13] (p. 10).

We may see the power of a representation to change the course of computational thinking through the example of the Musical Instrument Digital Interface (MIDI), developed in the mid-1980s. MIDI was the dominant protocol for the design and implementation of interactive music systems for a long time, and its use persists today. There were good reasons for this dominance, particularly at the end of the 20th century when these systems were first being developed. MIDI synthesis, sampling, and effects gear produced high-quality digital audio at an affordable price. Moreover, offloading the synthesis duties onto a dedicated piece of hardware freed the CPU of the computer to concentrate on control level analysis and composition.

The wild success and proliferation of the MIDI standard engendered an explosion of applications and systems that are still based today on a 35-year-old conception of music, and what information is sufficient to encode it. This standardized representation is basically that of music as played on a keyboard—MIDI was conceived as a communications protocol between keyboards and synthesizers, so sends out note numbers and velocities (how hard a key was struck, a proxy for loudness).

A more thorough representation of music was layered on by the way applications called *MIDI sequencers* added information to the raw keyboard messages. In particular, sequencers keep track of the time at which MIDI messages are received or sent—information that is nowhere encoded in the standard itself. However, by capturing and structuring these timestamps, sequencers can organize MIDI messages into bars, beats, and tempi, thus fitting them into the armamentarium of the Western rhythmic system. Note that adopting this standard then commits the user to the Western theoretical way of understanding music. While some work has gone toward expanding the common representations in a way that would encompass other musical cultures, much remains to be done.

3. Algorithmic Composition

Artificial intelligence is the combination of strategies and techniques that has aimed to emulate or replicate the activity of human intelligence [14]. Since emerging some sixty years ago, AI has oscillated between using explicit rules for intelligence emulations, or the training of models to do the same thing. Much of the early history in this field involved algorithmic reasoning over explicitly structured representations, for example in *expert systems*.

The algorithmic composition systems developed by Koenig, Xenakis, and Hiller are examples of this kind of thinking—a rule system is formulated that organizes and sometimes reasons over high-level musical constructs adopted from traditional notated scores. These are not, however, expert systems. They are not characterized by reasoning so much as they are exploring permutations of pre-existing repertoires of symbols, as Koenig notes: “Formulating a strategy differs from composing a piece in that not details, but basic conditions, are established—in minute detail, however. Formulating a strategy, after all, means generalizing formal relationships, which is at variance with the common practice of expressing musical ideals as concrete musical forms. Generalizing, unlike specifying, means making sure that everything can occur once somewhere, but that it may not be ‘missing’ either. It is hard to define ‘everything’: as something both present and absent” [12] (p. 176).

In this respect, the early algorithmic composition systems resemble the “generative” processes explored and articulated by Brian Eno: “From now on there are three alternatives: live music, recorded music and generative music. Generative music enjoys some of the benefits of both its ancestors. Like live music, it is always different. Like recorded music, it is free of time-and-place limitations—you can hear it when you want and where you want. And it confers one of the other great advantages of the recorded form: it can be composed empirically. By this I mean that you can hear it as you work it out—it doesn’t suffer from the long feedback loop characteristic of scored-and-performed music” [15] (p. 332).

In this paradigm, interactive systems may be read as a form of *mapping* from input sensations to output actions. We derive features (spectra, motion trajectories) from some input stream (audio samples, camera feed) and map these time-varying features onto parameters of a music (or multimedia) generation process. In a rule-based system, we are establishing a network of meanings: raising an arm “means” sweeping the center frequency of a filter up, for example. This has been the basic model of a large repertoire of interactive multimodal systems, and remains very much a viable construct for the composition of such work today.

Creators, including Joel Chadabe, resist the mapping paradigm: “Mapping describes the way a control is connected to a variable. But as instruments become more complex to include large amounts of data, context sensitivity, and music as well as sound-generating capabilities, the concept of mapping becomes more abstract and does not describe the more complex realities of electronic instruments” [16]. Given the complexities of both input and output spaces, Chadabe favors a navigational approach through the available permutations. He often uses a sailing metaphor, echoing Xenakis: “With the aid of electronic computers the composer becomes a sort of pilot sailing in the space of sound, across sonic constellations and galaxies that could formerly be glimpsed only in a distant dream" [17].

The difference appears as a change in orientation toward the underlying representations: mapping creates point-to-point correspondences between input features, or groups of features, and output behaviors. Navigation (or sailing) suggests the exploration of a high-dimensional space of possibilities whose complex interactions will emerge as we move through them. Pushed too far, mapping could become a rigid limitation of the possible, while navigation might aimlessly circle through undifferentiated choices. It becomes a meta-compositional problem that asks creators to decide where to apply their expertise: in developing a space of opportunities, specifying rules that govern how to invoke different parts of the space in response to unknown inputs, or both?

4. Artificial Neural Networks

The most rapid progress in AI currently is due to the development of Artificial Neural Networks (ANNs), a technique of building models that "learn" to perform classification and other tasks based on exposure to a large dataset of labeled examples. Loosely based on human brain function, units (or "neurons") in ANNs receive some number of inputs, either as inputs to the model as a whole, or from prior layers of units. Each input value is multiplied by a *weight*, and a *bias* is added to the weighted sum of all the neuron inputs. Finally, that result is run through a non-linear *activation function* to arrive at the output of the neuron.

In *supervised learning*, such a network is initialized with a set of random weights on all connections from one neuron to another. During the training phase, features associated with some class—say, pixel values of images of dogs or cats—are fed to the input neurons of the network. The learning is called "supervised" because the network is trained on a set of examples to which the designer has attached labels. In the case of labeling cats vs. dogs, the development of a training set can be straightforward. When working with musical material, labeling examples as, say, rock vs. pop can be more subjective, and the quality of the training labels will have a determinative impact on the accuracy of the resulting network. However, either way, initially, the network would predict the presence of a dog or cat randomly, due to the random weights attached to each neuron. In a process called *back propagation*, the previously attached correct output is compared to the predicted output. Then, weights in the network are adjusted moving backwards from the output to the input, adjusting them to make a correct answer more likely when presented with the same inputs again. The training process iterates through back-propagation passes until, ideally, the model converges on a set of weights that produce the correct outputs on all or most of the training examples. If the set of training examples has captured enough of the regularities of the classification problem being modeled, the trained network can subsequently be used on previously unseen feature inputs and correctly predict their class.

There are many detailed and definitive accounts of neural network design, implementation, and training [14,18,19], so I will not continue to elaborate on those points here. What is important to our further discussion is a consideration of what these networks have learned, what kinds of questions they can answer, and how they can be used in a multimodal, interactive music system. "A first fundamental challenge is learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals" [20] (p. 423). That being said, recent work has shown that training on more than one modality can help train a network to isolate parts of a video: "We show that by working with both auditory and visual information, we can learn in an unsupervised way to recognize objects from their visual appearance or the sound they make, to localize objects in images, and to separate the audio component coming from each object" [21] (p. 1).

In some respects, architectures based on supervised learning models track with the more rule-based systems we have considered so far: in particular, we typically would identify some number of features that would feed the input units of the network. These features have again been both symbolic and sub-symbolic in numerous studies. Supervised learning is most often used for classification tasks, where sets of inputs are categorized into a set of output options based on the training phase.

There are two aspects to note here relative to the discussion so far: first, a classification network constitutes a kind of perception—the trained model takes in raw sub-symbolic data, or features derived from such data, and returns a symbol. The network performs a signal-to-symbol transduction. The second point is to notice the internal representation that the network learns to perform its transduction: the only thing the network "knows" is the set of weights between units onto which it has converged during training. While it is tempting to characterize the knowledge that has been captured by these weights as purely musical, several studies have demonstrated how the network may instead have learned some entirely extrinsic regularity of the training set, e.g., background noise or recording artifacts [22,23]. One of the difficulties in working with neural networks is that they end up with this black-box quality. There is a significant research effort underway to improve the interpretability of ANNs [24,25], but as things stand, inputs go in, and results come out, making it difficult to know what happens in between and what transformations have been made to get from signal to symbol.

5. Ecological Perception

There is a view of perception that diverges from the cognitive science approach, and is similarly indifferent to the structure of underlying representations. The psychologist J.J. Gibson (1904–1979) pioneered the ideas of *ecological perception*. An outline of Gibson's thought, with respect to representations, is given by Eric Clarke: "The nature and existence of these representations is purely conjectural (they are inferred in order to account for behavior), and more fundamentally they suffer from the 'homunculus' problem: a representation only has value or purpose if there is someone or something to perceive or use it, which leads to an infinite regress of homunculi inside the perceiver's mind, each of which 'reads' and in turn generates an internal representation. Rather than making use of the structure that is already out there in the environment, the outside world is needlessly and endlessly internalized and duplicated (literally 're-presented')" [26] (p. 15).

J.J. Gibson said that the senses can "pick up" directly from the environment everything that is needed for survival. "When the constant properties of constant objects are perceived (the shape, size, color, texture, composition, motion, animation, and position relative to other objects), the observer can go on to detect their *affordances*. I have coined this word as a substitute for *values*, a term, which carries an old burden of philosophical meaning. I mean simply what things furnish, for good or ill. What they *afford* the observer, after all, depends on their properties" [27] (p. 285).

The response from cognitive science to Gibson's account emphasizes the impact of prior knowledge on perception: for example, it takes subjects four times as long to identify the suit of playing cards that have the wrong color. "When subjects in this experiment were provided with knowledge of the errors they were making, they became more aware of the types of stimuli that were being presented and made fewer errors, an interesting result that shows how people's immediate experience can affect their perception" [28] (p. 24).

This debate is of interest relative to a discussion of representations in that it reveals different commitments to the intelligibility of the data used to make decisions. Ecological perception discounts a need for internal representations as a human understands the world and takes actions in it. The collection of weights in a neural network are difficult to interpret, and offer a challenging set of parameters for an artist to manipulate. In both cases, we instead encounter a kind of resonance between the external environment and the interpreting agent.

6. Representations and Affordances

As designers of interactive multimodal systems, the choice of architecture determines the kinds of parameters available for manipulation. If we approach creation as mapping, do we map from symbolic or sub-symbolic inputs onto symbolic or sub-symbolic outputs, and where, if anywhere, do we insert rules leveraging our understanding of relevant symbols in the process? If we approach creation as exploration, how do we understand the space to be explored? Is it bounded by intelligible categories that could guide the exploration, or is it an emergent space with indescribable dimensions? In practice, these questions are best understood by looking at how a creator can guide the system toward an output: the exposed control parameters present an explicit set of affordances, in the Gibsonian sense, to the user.

A system's representations uniquely determine the artistic choices available to creators, and demonstrate the commitments of a system's designers. For example, the audiovisual installation *Meandering River*, co-designed by the Berlin studios *onformative* and *kling klang klong*, combines an algorithmically generated landscape of a river developing its characteristic meanders with a musical score composed using artificial intelligence [29]. The music was generated using tools from the Google Magenta project, trained on MIDI data captured from improvising musicians. Magenta's Performance RNN is "an LSTM-based recurrent neural network designed to model polyphonic music with expressive timing and dynamics" [30]. A recurrent neural network is a specialization of the ANNs we have been considering that incorporates an element of time-variance by sending outputs of the network back around to the inputs for processing of the next time step. The network takes MIDI and temporal information at the input and returns a sequence of MIDI events at the output—in both cases a lightly structured symbolic notation capturing performance information including tempered pitches, dynamics, and timing.

A similar model is OpenAI's MuseNet: "We've created MuseNet, a deep neural network that can generate 4-min musical compositions with 10 different instruments, and can combine styles from country to Mozart to the Beatles. MuseNet was not explicitly programmed with our understanding of music, but instead discovered patterns of harmony, rhythm, and style by learning to predict the next token in hundreds of thousands of MIDI files. MuseNet uses the same general-purpose unsupervised technology as GPT-2, a large-scale transformer model trained to predict the next token in a sequence, whether audio or text" [31]. "Unsupervised" here means that the model was trained on unlabeled examples, finding regularities in successions from one note to the next in hundreds of thousands of sequences. Note that there is no reason to believe the model has learned anything at all about "harmony, rhythm, and style"—it has learned how to generate the next token in a sequence, and all we can see about the mechanism by which it does that is a set of weights between units in an artificial neural network.

What representations are used and how they are manipulated by a system determine what controls are available to a creator in using that system. For MuseNet, the available controls in the "Advanced" mode are style (Chopin, Lady Gaga, Video Games, etc.); prompt (initial MIDI sequence, chosen from existing intros or uploaded in a file); instruments; and number of tokens. Whatever MuseNet may have learned about harmony and rhythm, creators have no access to or control over that knowledge. MuseNet is undoubtedly a sophisticated and valuable tool and its designers are to be commended for its development and free distribution. The point here is that this type of model, and the choice of representations used to train and target it, determine the possibilities available to creators when using it—its affordances.

7. Conclusions

In this brief review, we have considered the centrality of representations in shaping the possibilities for multimodal interaction. There is considerable controversy over the nature and function of mental representations in human cognition and great variety in the status and qualities of computational representations as they appear in interactive multimodal systems. In any permutation of beliefs and uses concerning representations, compelling artworks can and have been made—we have here noted multiple examples produced by artists with quite different commitments to the status of their representations, whether they consider the question explicitly or not.

The issue then is not whether one approach or another will lead to compelling works of art—that depends far more on the artist than it does on any technique. However, it does suggest new ways of thinking about representations and their manipulation in art making. One pathway is suggested by Gary Marcus in his recommendations for progress in artificial intelligence: "The right move today may be to integrate deep learning, which excels at perceptual classification, with symbolic systems, which excel at inference and abstraction" [32] (p. 20). Domain knowledge has been developed by musicians over centuries. It may be that networks trained to converge on a set of weights will capture ways of understanding musical structure that surpass that domain knowledge. However, it seems equally likely that finding ways to incorporate those symbols and structures into a hybrid model that leverages both what humans know and what machines can find will produce tools that afford more possibilities for creators to explore. Marvin Minsky made the following call for hybridity thirty years ago: "Our purely numeric connectionist networks are inherently deficient in abilities to reason well; our purely symbolic logical systems are inherently deficient in abilities to represent the all-important heuristic connections between things—the uncertain, approximate, and analogical links that we need for making new hypotheses. The versatility that we need can be found only in larger-scale architectures that can exploit and manage the advantages of several types of representations at the same time. Then, each can be used to overcome the deficiencies of the others" [33] (p. 36).

Artists are in a position to try out such constructs and no cars will drive into a wall. No one's illness will be misdiagnosed and no one will be incarcerated for longer than is just. Artificial intelligence has great utility and has made rapid progress, but still has a long way to go. The issues of representation, abstraction, and computation are now at the forefront of further development in one of the most crucial social, cultural, and scientific movements of our time. Art is always an important voice in epochal movements, but rarely is positioned to contribute as centrally as it can now.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Agawu, K. *The African Imagination in Music*; Oxford University Press: Oxford, UK, 2016.
2. Leman, M. *The Expressive Moment: How Interaction (with Music) Shapes Human Empowerment*; The MIT Press: Cambridge, MA, USA, 2016.
3. Lyon, E. *Designing Audio Objects for Max/MSP and Pd*; AR Editions, Inc.: Middleton, WI, USA, 2012; Volume 25.
4. Puckette, M.S.; Apel, T.; Zicarelli, D. Real-Time Audio Analysis Tools for Pd and MSP. International Computer Music Conference 1998. Available online: https://quod.lib.umich.edu/i/icmc/bbp2372.1998.315/1 (accessed on 15 January 2021).
5. Noble, J. *Programming Interactivity: A Designer's Guide to Processing, Arduino, and OpenFrameworks*; O'Reilly Media, Inc.: Newton, MA, USA, 2009.
6. Rowe, R. *Interactive Music Systems*; The MIT Press: Cambridge, MA, USA, 1993.
7. Guedes, C. Translating dance movement into musical rhythm in real time: New possibilities for computer-mediated collaboration in interactive dance performance. In Proceedings of the International Computer Music Conference 2007, Copenhagen, Denmark, 27–31 August 2017.
8. Savary, M.; Schwarz, D.; Pellerin, D.; Massin, F.; Jacquemin, C.; Cahen, R. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2013; pp. 2991–2994.
9. Camurri, A.; Volpe, G.; De Poli, G.; Leman, M. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimed.* **2005**, *12*, 43–53. [CrossRef]
10. Bohus, D.; Andrist, S.; Jalobeanu, M. Rapid development of multimodal interactive systems: A demonstration of platform for situated intelligence. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 493–494.
11. Thagard, P. *Mind: Introduction to Cognitive Science*; The MIT Press: Cambridge, MA, USA, 2005.
12. Koenig, G.M. Working with 'Project 1' my experiences with computer composition. *J. New Music Res.* **1991**, *20*, 175–180. [CrossRef]

13. Koenig, G.M. Composition Processes. In Proceedings of the Presented at the UNESCO Workshop on Computer Music, Aarhus, Denmark, 30 August 1978.
14. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson: Harlow, UK, 2020.
15. Eno, B. *A Year with Swollen Appendices: The Diary of Brian Eno*; Faber & Faber: London, UK, 1996.
16. Chadabe, J. The limitations of mapping as a structural descriptive in electronic instruments. In Proceedings of the 2002 Conference on New Interfaces for Musical Expression, Dublin, Ireland, 24–26 May 2002.
17. Xenakis, I. *Formalized Music: Thought and Mathematics in Music*, Revised Edition; Pendragon Press: New York, NY, USA, 1992.
18. Neural Networks and Deep Learning. Available online: https://static.latexstudio.net/article/2018/0912/neuralnetworksanddeeplearning.pdf (accessed on 15 January 2021).
19. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2018.
20. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
21. Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; Torralba, A. The sound of pixels. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 570–586.
22. Sturm, B.L. A simple method to determine if a music information retrieval system is a 'horse'. *IEEE Trans. Multimed.* **2014**, *16*, 1636–1644. [CrossRef]
23. Esparza, T.M.; Bello, J.P.; Humphrey, E.J. From genre classification to rhythm similarity: Computational and musicological insights. *J. New Music Res.* **2015**, *44*, 39–57. [CrossRef]
24. Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; Mordvintsev, A. The building blocks of interpretability. *Distill* **2018**, *3*, e10. [CrossRef]
25. Chakraborty, S.; Tomsett, R.; Raghavendra, R.; Harborne, D.; Alzantot, M.; Cerutti, F.; Srivastava, M.; Preece, A.; Julier, S.; Rao, R.M.; et al. Interpretability of deep learning models: A survey of results. In Proceedings of the 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI) 2017, San Francisco, CA, USA, 4–8 August 2017; pp. 1–6.
26. Clarke, E.F. *Ways of Listening: An Ecological Approach to the Perception of Musical Meaning*; Oxford University Press: Oxford, UK, 2005.
27. Gibson, J.J. *The Senses Considered as Perceptual Systems*; Houghton Mifflin: Boston, MA, USA, 1966.
28. Goldstein, E.B. *Cognitive Psychology: Connecting Mind, Research and Everyday Experience*; Nelson Education: Toronto, ON, Canada, 2014.
29. Urdesign. Onformative Turns Climate Changes into an Audiovisual Artwork. 2019. Available online: https://www.urdesignmag.com/art/2019/03/20/meandering-river-onformative-kling-klang-klong/ (accessed on 15 January 2021).
30. Simon, I.; Oore, S. Performance RNN: Generating Music with Expressive Timing and Dynamics. 2017. Available online: https://magenta.tensorflow.org/performance-rnn (accessed on 15 January 2021).
31. Payne, C. MuseNet. 2019. Available online: https://openai.com/blog/musenet/ (accessed on 15 January 2021).
32. Marcus, G. Deep learning: A critical appraisal. *arXiv* **2018**, arXiv:1801.00631.
33. Minsky, M. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *Ai Mag.* **1991**, *12*, 34–51.