GrainSynth: A Generative Synthesis Tool Based on Spatial Interpretations of Sound Samples

Archelaos Vasileiou, João André Mafra Tenera, Emmanouil Papageorgiou, and George Palamas(✉)

Aalborg University Copenhagen, Copenhagen, Denmark
{avasil19,jtener19,epapag19}@student.aau.dk, gpa@create.aau.dk

Abstract. This paper proposes a generative design approach for the creative exploration of dynamic soundscapes that can be used to generate compelling and immersive sound environments. A granular synthesis tool is considered based on the perceptual self-organization of sound samples by utilizing the t-Stochastic Neighboring Embedded algorithm (t-SNE) for the spatial mapping of sonic grains into a 2D space. The proposed system was able to relate the visual stimuli with the sonic responses in the context of the generic gestalt principles of visual perception. According to user evaluation, the application operated intuitively and also revealed the potential for creative expressiveness both from the user's perspective and as a standalone, generative synthesizer.

Keywords: Granular synthesis · Gestalt principles · Data visualization · Generative sound · Perlin force field · Machine learning for audio

1 Introduction

Granular synthesis is based on the same principle as sampling. The samples are split into small pieces of around 1 to 50 ms. These small pieces are called grains. Multiple grains may be layered on top of each other, and may play at different speeds, phases, volume and frequency, among other parameters, in order to create what can be thought as "sound clouds". The theory of granular synthesis was initially proposed, in conjunction with a theory of hearing, by the physicist Dennis Gabor [15]. Gabor referred to the grains as acoustical quanta, and he postulated that a granular or quantum representation could be used to describe any sound. Iannis Xenakis, [16], explicated a compositional theory of grains of sound. His theory describes a possible approximation to Gabor's model in the context of an analog synthesis implementation, where he suggested that the grained wave forms could be calculated directly on an appropriately programmed digital computer [10]. Recording a set of sound grains and mapping them as a data visualization could allow the performer to explore the sonic

N. Shaghaghi et al. (Eds.): INTETAIN 2020, LNICST 377, pp. 118–130, 2021.
https://doi.org/10.1007/978-3-030-76426-5_8

space in new ways of musical performance that could integrate a multi-modal, perception based framework, that could fuse both sonic and visual ques, at the same time. In order to make a consistent map of the various samples, these would have to be sorted and organized by a measure of perceptual familiarity. The aim of this study is to explore how the visualization match up with the user's intuitive sensibility of where the various samples should be positioned in space. This sensibility is influenced by the fact that humans associate the listening experience with simultaneous experiences obtained through non-auditory organs. This phenomenon is called synaesthesia [20].

2 Background

2.1 Generative Art

Generative art, as an artistic approach, utilize an autonomous system controlled by a set of predefined properties, balancing between unpredictability and order. This behavior arises out of the dynamics of a complex system. This system can be analysed in individual procedures and can be given a mathematical description which can be modelled and simulated. Thus, the generative system produces artworks by formalizing the uncontrollability of the creative process [18]. According to [19] "Generative art refers to any art practice where the artist uses a system, .., which is set into motion with some degree of autonomy contributing to or resulting in a completed work of art". The use of a generative approach provides new ways of expressing the artistic intent and purpose. Some supporters of generative systems consider that the art is not anymore in the achievement of the formal shape of the work but in the design of a system that explores all possible permutations of a creative solution [17]. Generative Art range from simple probabilistic procedures, to highly complex models that learn from a set of sample examples. Moreno [23] demonstrates a method to generate original bird vocalizations using a Variational Convolutional Autoencoder trained on a dataset of bird songs and call recordings. Training can be autonomous or might include a human in the loop. In their work [22] describe a human motion tracking system, from surveillance cameras on New York Time Square, that was used to feed a generative design algorithm in order to generate emotionally expressive 3D visualizations.

2.2 Audio Visualization

Audio visualizations, based on perceptual similarity of sound, have been used and implemented in various ways, for a variety of applications. The "Bird Sounds" interface [1] created at Google Creative Lab, applied the t-SNE algorithm to self-organize thousands of different bird songs, with a goal to depict their sonic relationships in a two dimensional grid. A similar application, "the infinite drum machine" [2], use a similar topological mapping as a spatial exploration tool of sound similarity. Selected sound samples could then be used to generate drum

loops. The "Audio Explorer" interface by Leon Fedden [3], is a project exploring an audio data-set visualization by mapping a multi-dimensional feature space represented by Mel-frequency cepstral coefficients (MFCCs), into a 2D space. The study considered a variety of dimensionality reduction algorithm such as the UMAP, t-SNE and the PCA. Another project on interactive exploration of musical space [4] build a music-space of 20,000 songs, visually rendered in a way that could enhance music navigation in a way similar to a recommender system.

2.3 Feature Extraction

Dimensionality describes the potential perplexity of a given data-set such as audio samples. A term often used is the "curse of dimensionality" which describes the exponential growth of the space of possible hypotheses as the dimensionality becomes higher [21]. This in effect creates sparse data representations that render the hypothesis statistically insignificant. The procedure of compressing a data-set by crafting new features from the existing ones and afterwards discarding the original set of features, is called feature extraction. The new data-set should be more comprehensive and inclusive in terms of information provided, as a summarized version of the original set.

MFCCs. Mel-frequency cepstral coefficients are commonly used as features in speech recognition systems as well in music information retrieval (MIR) applications such as genre classification and audio similarity measurement for recommender systems. MFCCs are perceptually motivated and spectrally smoothed representations of sound. The mel scale describes the non-linearity of the human ear, where each scale of pitches is perceived as equal in distance from one another [6]. This perceptually meaningful representation could be more comprehensive and inclusive in terms of information provided, as a summarized version of the original set.

2.4 Dimensionality Reduction

Principal component analysis is a popular technique for dimensionality reduction. PCA is essentially a multivariate data analysis method involving transformation of a number of possibly correlated variables into a smaller number of uncorrelated variables known as principal components. However, its effectiveness is limited by its global linearity [7]. Another popular choice that overcome the limitation of PCA is the Stochastic Neighbor Embedding algorithm (t-SNE) [9].

t-SNE. (t-Stochastic Neighbor Embedding) is a manifold learning technique used to visualize high-dimensional data by giving each data-point a location in a two or three dimensional map and it was introduced by van der Maaten and Hinton in 2008 [8]. The t-SNE requires tuning of parameters regarding initialization and visualization. It can be initiated randomly, or even through PCA, and can have its perplexity adjusted. Van der Maaten & Hinton suggests

a perplexity value between 5 and 50 [9]. The distribution of points obtained by t-SNE may be misleading when clusters form but practice may develop an intuition on how to interpret these observations. Looking at a t-SNE map, one of the first things to notice would be arbitrarily shaped clusters of data-points; this usually mean that data-points who are further in distance are considered to be dissimilar, while data-points appearing to be closer in space are considered to be more similar. Depending on their position, in a 2D plane for example, one may be able to recognize the patterns of the distribution and get to know the data-set better and maybe conclude to some solid observations [14] (Fig. 1).

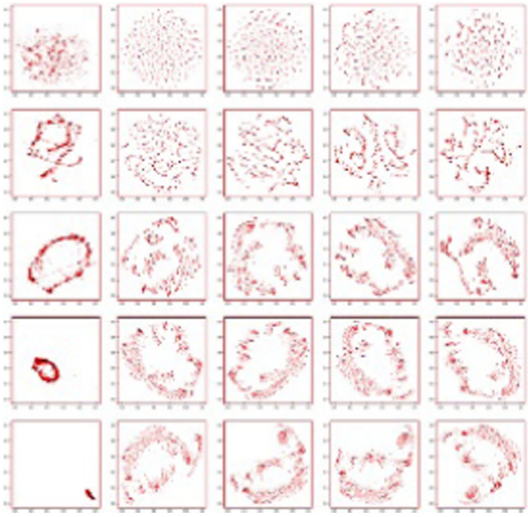

Fig. 1. t-SNE parameters test, with Perplexity (Y-axis) = [2, 5, 30, 50, 100] and Number of Iterations (X-axis) = [250, 500, 1000, 2000, 5000]. The Learning rate was set to 200

2.5 Gestalt Theory of Perception

The Gestalt school of psychology describes how we naturally perceive the world as perceptual groups [13]. Two Gestalt grouping principles, which pose a central role to this study, are the similarity and proximity rules. Ideally, sonic similarities among the sound grains would correspond with their visual proximity. In addition, there are principles describing connectivity and continuation, where points that form lines or other shapes are perceived as a whole. The various Gestalt principles can override each other, which means that in some cases the perceived sonic similarity can outweigh the importance of the visual grouping, and vice versa. Other times the principles are complementary, pulling in the same perceptual direction.

3 Creative Workflow

A modified version of the Audio t-SNE Viewer [5] was used as the basis of this system. The system consists of 4 discrete parts as can be seen in Fig. 2, which are further analysed below.

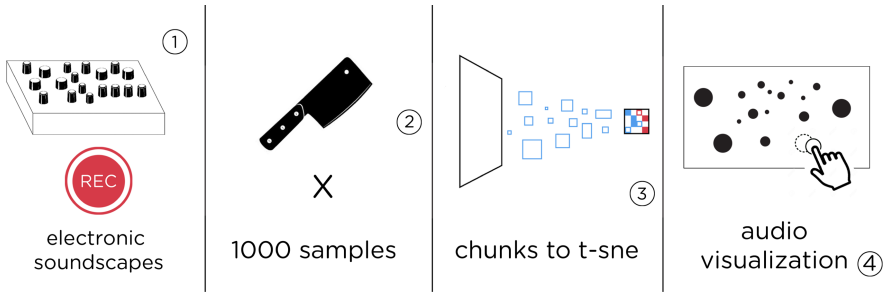

Fig. 2. Workflow of the sound analysis and visualization

3.1 Generating Electronic Soundscapes

Analog electric soundscapes generated and recorded with a synthesizer that has been built from our research team. This single voice semi-modular synthesizer designed as a tool for the exploration of soundscapes through a variety of sound synthesis techniques. The recording lasted for about an hour and as a result more than a thousand samples were generated. Initially, fundamental frequencies were recorded, then periodic frequencies were generated from a Voltage Controlled Oscillator (VCO), controlled by a Low Frequency Oscillator(LFO). The first step was to record every waveform format that could be produced by the oscillator in order to get different timbres. Next step was to repeat the same method and lead the same signal both to the VCO and to a Voltage Controlled Filter (VCF). Moreover, drum timbers were generated by mixing white noise together with a very low frequency and lead them into the filter which was modulated at the same time by a very short Envelope Generator (EG).

3.2 Sample Generation

Next step was to create a set of sound chunks, with 1000ms duration each and a sample rate of 44.1 kHz/24bit. A set of 1063 sound files were generated and an overlap of 1.5ms was used to create a continuous feeling when samples would be played in a sequence.

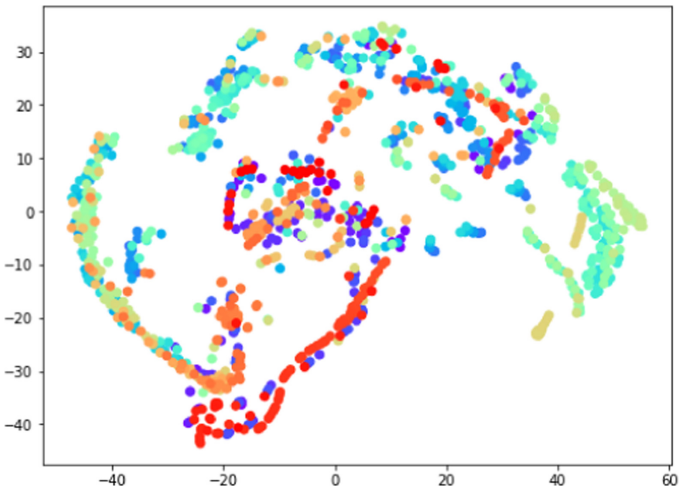

Fig. 3. Spatial mapping of sound samples based on the t-SNE algorithm. Sounds that appear to have a similar color, belong to the same sequence and probably have similar sound content. Different sections (e.g. teal and orange) appearing clustered together as well. This suggests those two sections may have similar audio content.

3.3 Chunks to t-SNE

Mel-frequency cepstral coefficients are a common choice in speech recognition systems as well in music information retrieval (MIR). The mel scale describes a scale of pitches perceived as equal in distance from one another [6]. For each audio chunk the first 13 mel-frequency cepstral coefficients have been calculated along with their first- and second-order derivatives, and concatenated into a single 39-element feature vector which would characterize each clip and is standardized so that each feature had an equal variance.

3.4 Visualizing the Sound Manifold

The t-SNE algorithm, as a manifold learning technique, was used to reduce the dimensionality of the initial $(N \times 39)$ data-set to only two dimensions $(N \times 2)$, where $N = 1063$ is the number of sound grains. Additionally the results were normalized between 0 and 1. The t-SNE requires tuning of two hyper-parameters, the perplexity and the learning rate. The perplexity parameter relates with the number of nearest neighbors; As a rule of thumb, Van der Maaten and Hinton suggest a perplexity value between 5 and 50 [9]. Learning rate was taken into consideration, as if set too high it might cause the data to be hard to analyse due to excessive proximity as well as if too low where most points may get clustered in exaggeration. These considerations take into account the Kullback-Leibler cost function for preventing it to get stuck in a local minimum [11]. In a t-SNE map, clusters might form corresponding to individual sound chunks, with

similar sounds occupying nearby positions and dissimilar sounds positioned far away. Thus, certain clips are placed together in clouds of varying size, while others end up in the periphery of the map (Fig. 3).

4 The Granular Synthesizer

A functionality for looping the samples has been added in the original application. That way, by controlling the duration of the grains the app could work in a similar way as a granular synthesizer. The original code had a minimum clip duration of 100 milliseconds, which is just at the border of what can be defined as granular synthesis [12]. The minimum clip duration was lowered to 1 millisecond and the window length of the grains was determined by the maximum duration parameter. Each triggered clip always starts at its original beginning, and loops at a rate set by adjusting the maximum duration. This means that both the grain size and the grain density are controlled by the same parameter. This density only applies horizontally over time, while the vertical density depends on how many grains the user plays at the same time. Since the grains are all regularly placed over time, this should be characterized as synchronous granular synthesis [10]. However, there is no grain spacing, since they are all played in continuous loops with no spaces in between. A panning gives the application space to breathe in the stereo field. It is used lively and interactively by identifying the spatial position of performative gestures. The technique that has been used identifies the user's relative position in the visualization and multiplies it by 45°. The angle is normalized by $[-1, +1]$ and is converted into radians.

Human Computer Interaction. A common way to generate sounds from a synthesizer is through a MIDI controller, typically by triggering sounds and control parameters during a music performance. A quite different way is to use natural gestures on a touch screen or input from haptic sensors. Sparse distributions and data positions throughout the screen would allow for navigation through these sounds, inviting energetic exploration and improvisation. Moreover, different t-SNE parameter can affect the spatial navigation and perception of sonic stimuli. We have found that distributions with values greater than 30 in perplexity and greater than 500 in number of iterations, could form distributions that visually imposes a good navigational structure for exploring the spatial formations of the samples.

Generative Design Scheme. In order to aid the creative exploration of the synthesis of soundscapes a flow field was used to control a number of particles within it. These moving particles, influenced from this force field could trigger successive grains along their pathways. Flow fields are especially useful for modeling chaotic movement, such as fluid dynamics, for procedurally generated textures and for the control of the movement of autonomous agents. A flow field is an area of usually 2D or 3D space divided up into a grid of cells. Each cell

contains a velocity vector which represents a direction and speed of movement Fig. 5. When a particle enters a cell, its direction is transformed to match that of the vector in the cell. As it moves through the field it will enter other cells containing velocity vectors that change its movement again. The crucial factor in a velocity field is the arrangement of the vectors. We use a special case of a flow field that is based on a perlin noise texture (grid). Each cell of the flow field represents a single, animated, force field vector. The vector is represented by three values, the angle, the magnitude of the vector at the given position and a global time t value that represents the evolution of the process. Thus, two different Perlin textures are needed to describe an animated force field. When a particle is crossing over a sample point, the sound would be triggered by it. That way, a high number of particles could trigger many samples at once, which could generate aesthetically pleasant, polyphonic sequences (Fig. 4). As the sound samples are not evenly distributed to cover all the available space, a lot of empty space forms, surrounding the grains. This neutral space serves as silent "white space" that simplifies the scene, according to the Gestalt approach, into figure and ground.

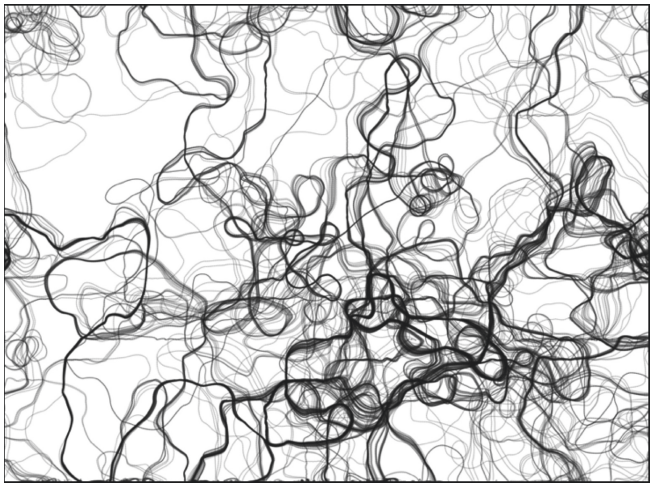

Fig. 4. Ten particles spawned at the center of the canvas, moving under the influence of a perlin based flow field, formed by a grid of size 32×22.

4.1 User Experience Testing

Eleven (11) Participants, with previous experience in synthesizer practices, were asked to explore the interface and customize the available parameters after being briefly instructed on the system's controls. There was no specific time frame in which the user had to complete the evaluation. The users were prompted then

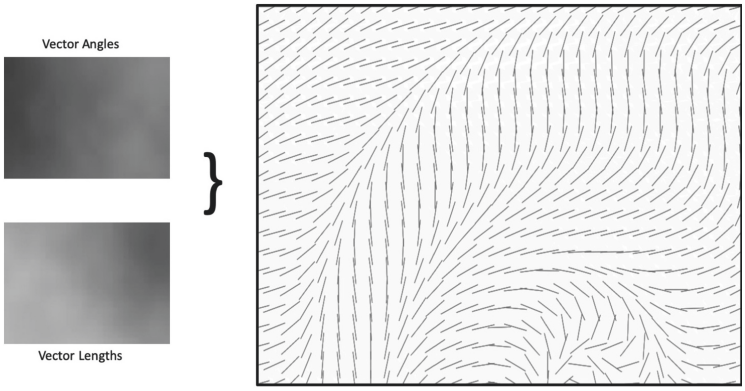

Fig. 5. Two different textures were used to form a perlin noise flow field. The first texture (up) used to store the vector angles of the vectors in radians, while the second texture (down) used to store the magnitude of the force vector. The size of the textures is 32×22 pixels.

to answer a questionnaire in which they were asked to rate some aspects of the interface's performance and their overall user experience, along with some additional commentary. Moreover, a focus group was used as the main method for the evaluation of the generative capabilities of the system.

Gestalt Grouping Principles Testing. A second questionnaire was conducted as a non-parametric test and was addressed to a random sample of twenty (20) people with no cognitive perquisites. The non-parametric test aimed to explore the efficacy of the application to achieve a sense of playing intuition through the scope of the grouping principles of the Gestalt School of thought. Five basic principles were interpreted as questions:

- **perplexity**: "Do neighbouring dots in the visualization represent sonic similarities in sonic space?"
- **Closure**: "Did you notice forming patterns in the distribution and if so, did the sound correspond to sonic landmarks while navigating through them?"
- **Continuity**: "Did the distribution of points suggest a path to navigate through and if so, did the sound correspond to a sonic gesture?"
- **Common Region**: "Did the samples' position display correspond to their position in the sonic space?"
- **Similarity**: "Were similarly sounding samples, visually grouped together in the distribution?"

4.2 Results

Tables 3 and 1 presents the average score on a ten-point scale in regards to each of the system's performance overview and the user experience's aspects

respectively. Most participants reported highly satisfied with the interface control response. Regarding the interface's parameter customization layout, it has been reported that being able to adjust the parameters while producing sound simultaneously, would be a useful and creative feature. As far as the user experience is concerned, none of the participants got the impression of being able to anticipate the audio samples they were triggering, so as a result, they rated the playing intuition and the ability to trigger the desired sounds the lowest. However, after an initial navigation period the participants were able to anticipate the sounds being triggered. This finding might suggest the formation of cognitive spatial maps that could possibly help the user navigate within a previously experienced topological map of sounds. The synth controlling novelty was rated high and five out of six participants with experience in music production, reported that they would use the current interface in a music production project of their own. Moreover, regarding the generative approach the overall aesthetics was rated as good with temporal coherence although some irregularities have been spotted.

Non - parametric Test. The outcome of the second experiment would provide evidence on whether or not the application achieved its goal of providing the user with a novel and intuitive way of playing with a synthesizer. Table 2 demonstrates the average scores of the application's compliance with Gestalt's principles while Fig. 6 provides a histogram of the score of the individual ratings of the grouping principles against the number of occurrences. All collected data underwent the Kolmogorov - Smirnov Test of Normality and found not to differ significantly from that which is normally distributed.

5 Future Work

There are many possibilities of extending the functionality of this application by utilizing the user interface for ultimate expressiveness and better performance. A desirable way would be to use tactile sensors or human motion, such as input from live camera feed, to trigger and manipulate the samples. Within this example, the application would have more potential of being used not only by musicians, but also from dancers or performers. For the moment, our team is focusing on utilizing this system with an optical flow system, based on real time camera feed. Our intention is to integrate a real-time sound generation system, for dance performance, with a motion expressiveness visualization system. (https://vimeo.com/220138824).

Moreover, adding new generative algorithms such as a flocking boid [21] is under development. Flocking boid is similar, in a sense, with a perlin flow field, however within a flocking boid the set of individual agents is capable of interacting with each other. In a similar fashion, when a flocking agent would cross over a sample point a sound would be triggered by it. That way, a population would trigger an arbitrary amount of samples, generating interesting and stochastically rich soundscapes.

Table 1. Average ratings of their user experience on a 10 point scale (sound engineering students).

User experience	AVG score
1. Playing intuition	6.9
2. Desired sounds triggering	5.09
3. Panning correspondence	8.09
4. Sample distribution scheme	7.45
5. Synth controlling novelty	8.18

Table 2. Average ratings of the application's compliance to Gestalt's grouping principles on a hundred point scale.

Gestalt grouping principles	AVG score
1. Perplexity	79
2. Closure	62
3. Continuity	63
4. Common region	69
5. Similarity	75

Table 3. Average ratings of the application's performance efficiency on a 10 point scale (sound engineering students).

Performance overview	AVG score
1. Navigation triggering efficiency	8.18
2. Click-looping efficiency	7.54
3. System latency	9
4. Navigation and loop synergy	7.72
5. Parameter customization	7

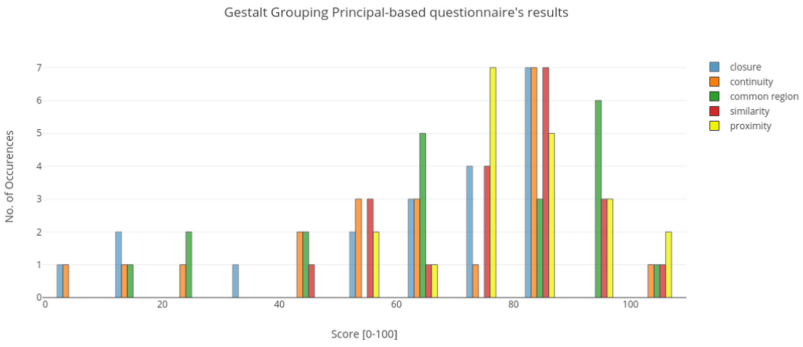

Fig. 6. Histogram of the non-parametric Evaluation Test

6 Conclusion

The proposed system augmented the artistic capabilities of a semi-modular analog synthesizer. A data-set of 1000ms audio clips were self-organized using different visualization techniques, according to their musical content. As has been demonstrated the dimensionality reduction capabilities of t-SNE is a rewarding approach for shaping a "playable" visualization. Perceptually, the application achieved to connect the visual stimuli and aural sound based on the generic gestalt principles of grouping and continuation as well as the figure-ground principle. According to user evaluation, the application operated quite intuitively. The evaluation also revealed the potential for creative expressiveness both from the users perspective and as a standalone, generative synthesizer.

References

1. Tan, M., McDonald, K.: (2006). https://experiments.withgoogle.com/bird-sounds
2. McDonald, K., Tan M., Mann Y.: The infinite drum machine (2018). https://experiments.withgoogle.com/ai/drum-machine/view
3. Fedden, L.: Comparative audio analysis with wavenet, MFCCs, UMAP, t-SNE and PCA (2017). http://bitly.ws/8E26
4. Lionello, M., Pietrogrande, L., Purwins, H., Abou-Zleikha, M.: Interactive exploration of musical space with parametric t-SNE. In 15th Sound and Music Computing Conference (SMC 2018) Sound & Music Computing Conference, pp. 200–208. Sound and Music Computing Network (2018)
5. Kogan, G.: (2018). http://ml4a.github.io
6. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012)
7. Gewers, F.L., et al.: Principal component analysis: a natural approach to data exploration (2018). arXiv preprint arXiv:1804.02502
8. Maaten, L.V., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
9. Wattenberg, M., Viégas, F., Johnson, I.: How to use t-SNE effectively. Distill **1**, 10 (2016)
10. Roads, C.: Microsound. MIT Press, Cambridge (2004)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
12. Park, T.H.: Introduction to Digital Signal Processing: Computer Musically Speaking. World Scientific, Singapore (2009)
13. Wertheimer, M.: Gestalt theory (1938)
14. Lunterova, A., Spetko, O., Palamas, G.: Explorative visualization of food data to raise awareness of nutritional value. In: Stephanidis, C. (ed.) HCII 2019. LNCS, vol. 11786, pp. 180–191. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30033-3_14
15. Gabor, D.: Theory of communication. Part 1: the analysis of information. J. Inst. Electr. Eng.-Part III: Radio Commun. Eng. **93**(26), 429–441 (1946)
16. Xenakis, I.: Formalized Music: Thought and Mathematics in Composition. Pendragon Press, New York (1992)

17. Kaspersen, E.T., Górny, D., Erkut, C., Palamas, G.: Generative choreographies: the performance dramaturgy of the machine. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, vol. 1: Grapp, pp. 319–326. SCITEPRESS Digital Library (2020)
18. McCormack, J., Bown, O., Dorin, A., McCabe, J., Monro, G., Whitelaw, M.: Ten questions concerning generative computer art. Leonardo **47**(2), 135–141 (2014)
19. Galanter, P.: Generative art theory. In: A Companion to Digital Art, pp. 146–180 (2016)
20. Parise, C., Spence, C.: Audiovisual cross-modal correspondences in the general population (2013)
21. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
22. Billeskov, J.A., Møller, T.N., Triantafyllidis, G., Palamas, G.: Using motion expressiveness and human pose estimation for collaborative surveillance art. In: Brooks, A.L., Brooks, E., Sylla, C. (eds.) ArtsIT/DLI -2018. LNICST, vol. 265, pp. 111–120. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-06134-0_12
23. Moreno, J.A., Bigoni, F., Palamas, G.: Latent birds: a bird's-eye view exploration of the latent space. In: Proceedings of the 17th Sound and Music Computing Conference, Torino, June 2020