Narrative-inspired Generation of Ambient Music

Sarah Harmon

Abstract

An author might read other written works to polish their own writing skill, just as a painter might analyze other paintings to hone their own craft. Yet, either might also visit the theatre, listen to a piece of music, or otherwise experience the world outside their particular discipline in search of creative insight. This paper explores one example of how a computational system might rely on what they have learned from analyzing another distinct form of expression to produce creative work. Specifically, the system presented here extracts semantic meaning from an input text and uses this knowledge to generate ambient music. An independent measures experiment was conducted to provide a preliminary assessment of the system and direct future work.

Introduction

Researchers have long established that artificial agents can learn from humans, in addition to each other, to generate creative solutions (Gervás 2001; Boyd, Hushlak, and Jacob 2004; Bisig, Neukom, and Flury 2007; Codognet and Pasquet 2009; al Rifaie, Bishop, and Caines 2012; Bornhofen, Gardeux, and Machizaud 2012; Dubnov and Surges 2014). In contrast, relatively little work has examined how computational systems might analyze artifacts that reside outside their expert domains. While analogical reasoning is well-established as a technique in artificial intelligence, it rarely involves mapping from a source to a genre-distinct target. Even less considered is this type of mapping for the generation of creative works, despite the fact that humans naturally face such tasks daily.

This work provides a step toward what we will term *extra-inspired systems*, i.e., systems that can learn how to usefully and creatively map and/or transform semantic concepts from remote domains to their own, at or beyond the level of human capability (refer to Figure 1). Examples of extra-inspired systems include a poetry generator that can learn from a painting, or a dancing robot that is able to incorporate ideas from the tone or timbre of someone's voice into its choreography. In both cases, the creative system must gain and apply knowledge from outside its primary domain. In so doing, the connection between the source of inspiration and the resulting product should ideally be apparent to the audience. This is not to say that creative products could not result if this connection is not made; rather, we assume extra-inspired systems should possess this feature to promote computer-human storytelling across domains.

Here, we introduce a framework that uses semantic information extracted from natural language texts to produce ambient music. This is a challenging task to pursue, given that music composition is complex even for humans (Delgado, Fajardo, and Molina-Solana 2009) and open information extraction is yet a developing field (Gangemi 2013). In light of these barriers, we will discuss how past work provides a foundation for the design and assessment of such an extra-inspired system.

Related Work

Ambient Music Perception and Composition

There are several well-established heuristics for ensuring that music is pleasurable to the human ear. Many of these are rooted in the concept of consistent and natural musical motion. In other words, large leaps in structure or style that occur rapidly are not considered pleasurable. Conjunct melodic motion and efficient voice leading are two examples, reflecting the notion that the ideal path when harmonies change tends to be the shortest possible. Centricity, a quality in which a tonic note is regarded as prominent and serves as the goal of musical motion, is another example of consistency preservation.

Beyond being able to generate pleasing ambient music, it is important that our system engage in creative behavior. Prior work in computational creativity has suggested that creative composition systems should strive to be fully autonomous (Wiggins et al. 2009). Further, the system should be able to generate music that is *novel* and *valuable* for itself and its audience to be considered creative (Boden 2009). Novelty may be *psychological* (new to the composer) or *historical* (new across society for the first time in history).

We look to a key pioneer in ambient music to suggest what is valuable in a generated ambient composition. Brian Eno, who coined the term *ambient music* and contributed heavily to its development as a genre, points to several crucial features that apply to ambient music specifically. According to Eno, ambient music "is intended to induce calm" and "must be as ignorable as it is interesting" (Eno 1978). We will refer

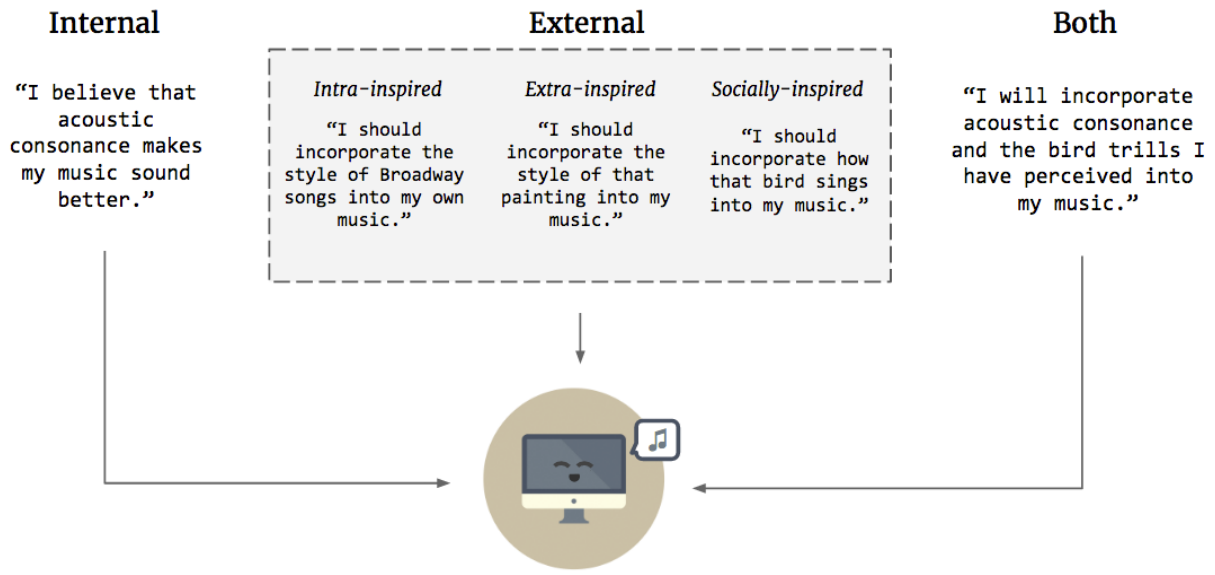

Figure 1: An example of a music-making system is used to illustrate the distinction between systems that are inspired by internal (*internally-inspired*), external (*externally-inspired*), or both internal and external (*ambi-inspired*) cues. These systems may be further categorized by the types of cues to which they attend. For instance, *intra-* and *extra-inspired* systems are defined here as those that apply knowledge from within and outside their expert domains, respectively, whereas *socially-inspired* systems apply knowledge specifically gained through interaction with another being or system. These categories are not necessarily mutually exclusive. Further, note that this categorization does not represent a hierarchy: ambi-inspired systems, for example, do not necessarily generate products of higher or lower value than systems that are purely inspired by internal means. The distinction is simply between underlying methodologies.

to these qualities to guide the design of our ambient music generation framework.

Extra-Inspired Music Generation

A number of systems already exist that use visual cues to generate musical experiences (Granito; Pappas; Walker et al. 2007). Mapping text to music is a natural next step, especially considering that homologous brain regions appear to support functions at the level of semantic and temporal structures for music and language in humans (Brown, Martinez, and Parsons 2006). Unlike motion and other visual cues, however, text does not as reliably map to specific changes in sound. For this reason, prior text-to-music systems tend to rely on surface features of text to direct generation rather than semantic cues.

Rangarajan, for instance, recently proposed three strategies for mapping text to music: (1) mapping letters to notes, and their frequencies of occurrence to note duration, (2) mapping only vowels of the words to notes and note duration, and (3) mapping vowels and respective part-of-speech category for each word to notes (2015). This method, however, was acknowledged to be limited because only surface linguistic features were used, in addition to the fact that the resulting music was not necessarily desirable. The support of heuristics such as centricity and efficient voice leading could not be guaranteed.

Davis and Mohammad took text-to-music generation a step further with their proposed *TransPose* system (2014). TransPose was designed to generate music that captures emotion dynamics in literature ("the change in the distribution of emotion words"). Novels with a more positive emotion profile were represented by assigning a major key to the piece, while novels with more negative emotion words were assigned a minor key. A direct positive correlation was assigned between the frequency of emotion words and tempo of the resulting musical piece.

While these systems are unquestionably important contributions toward meaningful text-to-music generation, much of the substance in the original text does not tend to be preserved in the final product. We chose to couple semantic concept mining with ambient music generation to determine if the interpreted meaning of a text could be clearly and creatively mapped to a novel soundscape. The benefits of pursuing this work include improved responsive audio for interactive playable media, in addition to new, useful, and interesting sound generation for the visually-impaired, as in (Walker et al. 2007).

Perhaps the closest system to the present work is Audio Metaphor (Thorogood, Pasquier, and Eigenfeldt 2012; Thorogood and Pasquier 2013). As in the system described here, Audio Metaphor transforms a natural language query into either (1) a set of audio file recommendations for a soundscape composer or (2) a generated soundscape. The key distinction is the processing of the input text. Audio Metaphor preprocesses the input text by removing common words, and groups the remaining words into a list. This list

serves as the first search query. If the first search query is unsuccessful, more search queries are derived by rearranging or removing some of the words. The input phrase "On a rainy autumn day in Vancouver" would thus result in search queries of *rainy autumn day vancouver*, *rainy autumn day*, *autumn day vancouver*, *rainy autumn*, and so on.

In contrast, the present system performs semantic concept mining on the input text, and can perform transformations on the words or phrases themselves based on its knowledge of the overall story. To illustrate, consider the "on a rainy autumn day in Vancouver" example. In this case, certain properties of the narrative setting are extracted; namely, the fact that it is rainy, autumn, during the day, and in Vancouver. When deriving the search query, the lexical or grammatical categories can be automatically changed (as in *rainy* to *rain*). Generalizations can also be made to make the query more abstract (such as *Vancouver* to *Canada*, *Vancouver* to *city*, or *Vancouver* to *ocean city*). Metaphorical connections may be discovered as well, such as the relation between *rainy* and a particular atmospheric mood. Overall, this method enables the system to represent input as a set of structured concepts rather than a list of unfamiliar words, and to thereby make connections about what is being described. As a result, the system has more techniques at its disposal to make expressive decisions and ensure search query quantity and relevance.

Approach

The Rensa knowledge representation framework was used to encode, extract, and transform semantic information (Harmon 2017).[1] To generate ambient music from text-based narratives, narrative concepts and information regarding how they are related are first gleaned from a provided natural text input (*reading phase*). This knowledge is then automatically translated into search queries for web-based sound libraries (*interpretation phase*), which means that novel sounds may be gathered as the libraries continue to expand. Any returned results are analyzed based on properties specific to ambient music, and combined to generate a set of possible final compositions (*brainstorming phase*). This set is then evaluated by the system based on its understanding of the ambient music domain (*critique phase*). If no results or insufficient results are found during the *interpretation* phase, the system will return to the original text and attempt to extract additional concepts and create new queries. If insufficient concepts are identified in the source text to create a piece of music, the system will still inform the user of any knowledge gained and will suggest sound files based on this knowledge. This enables one to potentially use the system as a computational creativity support tool. The following subsections will explain each major phase of the procedure in more detail.

Read and Interpret

Because state-of-the-art information extraction tools are currently not accurate enough to infer the complete meaning of a text (Gangemi 2013), simple concept relation information was extracted using a set of text pattern rules. These rules permit the extraction of facts such as actions and object properties (examples provided in Table 1). When any information is extracted, our system stores where it found the information temporally in the story. It also checks if this information was already known to be true or false.

The Rensa framework also provided several functions for extracting story characters and predicting their gender identity. Figure 2 demonstrates an example of how the present system uses this information when translating extracted semantic information into more abstract search queries.

The search queries were not enclosed in quotes, and only tags and file names were scanned to increase relevancy. The term "-voice" was appended to queries that described properties of entities (such as "cool wind") to ensure that no distracting speech would be retrieved as part of the search. This term was not appended for queries describing actions, however, as in Figure 2. All query results with at least a 4-star rating were retrieved.

Brainstorm

During the brainstorming phase, the generator selects sounds to be used within each piece. Not all retrieved sounds nor extracted concepts need be invoked when generating a new composition. A primary background sound is chosen among the subset of sounds which are greater than or equal to thirty seconds in length, lending to musical consistency. If no sounds are retrieved with this length, the system still arranges the piece using retrieved sounds, but makes a note of the deficiency. Any other sounds chosen are positioned temporally as in the original source.

To support the generation of a calm, natural-sounding, and ignorable piece, all sounds are analyzed for changes in volume. If it appears that rapid changes in volume occur (*local distraction*), or that a sound is much louder than other sounds in the piece (*global distraction*), the distracting sound or portion of the sound is masked. Small pitch adjustments may also be made depending on the results of using the Krumhansl-Schmuckler key-finding algorithm (Temperley 1999; Shmulevich and Yli-Harja 2000; Zhu and Kankanhalli 2006; Sapp et al. 2011) to support consonance and harmonic consistency.

Critique

The system has several feasible measures for critiquing its own work. Psychological novelty is assessed by examining both the extracted concepts and retrieved sound files used. If the semantic similarity of the extracted concepts is proportionally high to concepts that have been previously extracted (which are stored in memory), then the musical piece is considered less novel. Similarly, if the retrieved sound files already exist in the case library, the system will rank the musical piece as less novel. These criteria enable the system to compare the novelty of its generations relative to each other.

Ignorability and pleasantness are both considered to be qualities that point to how useful the resulting piece will be for the listener. These qualities are assessed by examining drastic changes in dynamics, pitch, and tempo. Changes in

[1] https://github.com/RensaProject

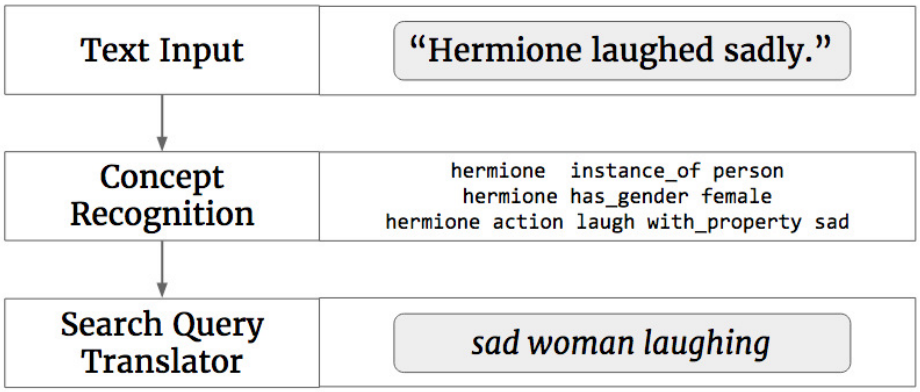

Figure 2: An example that demonstrates how a sentence from an input text is transformed into a search query for a sound file. Semantic concepts are extracted from the text and may be used to make specific terms more abstract. In this case, the word *Hermione* is recognized as an animate actor with a girl's name, and is abstracted to become the term *woman*. The search query translator can also change the part-of-speech or tense of a word to form a more useful query (e.g., *sadly* to *sad*, or *laughed* to *laughing*).

Rule	Example	Extracted Information
a_1 (RB) VB .	*Timothy dreamed.*	actor: **Timothy**, action: **dream**, tense: **past**
a_1 (RB) VB (DT) (JJ) NN .	*Timothy (sadly) hated (the) (autumn) leaves.*	actor: **Timothy**, action: **hate**, with_property: **sad**, action_object: **leaves**, action_object_property: **autumn**, tense: **past**
a_1 (RB) VB a_2	*Timothy (really) loved Koko.*	actor: **Timothy**, action: **love**, with_property: **really**, action_object: **Koko**, tense: **past**
a_1 (RB) VB (PRP$) (JJ) NN TO a_2	*Koko (kindly) gave (her) (helpful) advice to Timothy.*	actor: **Koko**, action: **give**, with_property: **kind**, action_object: **advice**, action_object_property: **helpful**, action_object_owner: **Koko**, action_object_recipient: **Timothy**, tense: **past**

Table 1: A subset of pattern matching rules used by the present system to extract the occurrence of actions performed by narrative actors (story characters). Here, $a_1, a_2 \in A$, where A is the set of all known actors automatically recognized in the text. The '.' symbol refers to the existence of a full stop or period. All other terms are part-of-speech tags (RB=adverb, VB=verb base form, DT=determiner, JJ=adjective, NN=noun, PRP$=possessive pronoun, TO=the word 'to'). Any term that is enclosed within parentheses is considered optional.

dynamics are estimated via a root mean square approach, while tempo is assessed using beats per minute (BPM) detection. Pitch detection is achieved via cepstrum analysis, i.e., fast Fourier transforms (FFTs) coupled with low-pass filtering (Roche 2012; Gerhard 2003). Using this method, the fundamental frequency is estimated as:

$$\hat{f}_1 = \frac{1}{\tau_{max}}, C(\tau_{max}) = \max_{\tau} C(\tau), \tau > 0 \qquad (1)$$

...given that the power cepstrum $C(\tau)$ is obtained by transforming the signal $x(t)$ using a FFT algorithm, con-

verting to the logarithmic scale, and transforming via FFT again:

$$C(\tau) = F^{-1}(\log|F(x(t))|^2) \quad (2)$$

Other qualities to be assessed for a given generated piece include additional measures of consistency, evocativeness of the source text, and atmospheric mood. The number of extracted concepts used helps to assess how evocative a generated musical piece is relative to other compositions. The key of the piece and its components can also be analyzed to assess consistency and predict the emotional mood, similar to (Davis and Mohammad 2014).

A complete evaluation of the automated critique phase is ongoing and will be presented in future work. First, however, we seek to determine whether our system's procedure is able to generate new and interesting musical pieces while preserving semantic meaning. This will serve to strengthen our understanding of meaningful ambient music generation and build toward improved automated modules for creative reading, interpretation, brainstorming, and critique.

Evaluation

Study 1

Method 140 participants on Mechanical Turk were recruited to read a short (<100 words) passage from *The Wonderful Wizard of Oz* that we will call *Emerald City*. They were then asked to listen to an audio track, and to specify whether the audio reminded them of the passage. Participants were required to provide at least one justification for their answer. The experimental group listened to an audio track which actually corresponded to the passage, while the control listened to a track generated from distinct semantic concepts. Specifically, this track was generated from processing another text passage, which we will refer to as *Dungeon*. Each audio track was limited to 30 seconds, and a fade out effect was applied during the last second of each piece.

Results 45.8% of the experimental group and 11.8% of the control described the audio track as pairing well with the text passage. A two-proportion hypothesis test suggested that the difference between the two groups was statistically significant ($p<0.05$). Spearman analyses further suggested no statistically significant correlation existed between the text-audio matching decision and (1) whether the participant had a certain gender identity (ρ=0.0610), (2) read the *Emerald City* passage before (ρ=0.127), or (3) was familiar with concepts presented in the passage, such as Dorothy and her friends (ρ=-0.112).

Approximately 17% of the experimental group (compared to no participants in the control) explicitly remarked in their explanation that the audio would be suitable as background music for a film adaptation of the passage. Other perceived concepts, such as the emotions (e.g., serenity, awe) that the music conveyed could not reliably predict a participant's decision.

Study 2

Method To further explore how participants might have arrived at their justifications, personal interviews were conducted as part of a small but more in-depth case study. Six music tracks were generated from sound files that had been obtained as a result of the present system interpreting six excerpts from narrative fiction. We will refer to these six excerpts and their corresponding tracks as *Emerald City*, *Dungeon*, *Tower*, *Elven Realm*, *Cave*, and *Entering Fantasy World*. Four participants (2M, 2F) listened to the tracks and read the original source texts. All six excerpts were each less than 800 words in length, and all participants were at least moderately familiar with the fictional environments described in the input texts. The tracks themselves ranged from 2-10 minutes in length (2:07 to 9:07).

Participants were then asked to rank the tracks in terms of how well they matched each text, from closest to least closest. They could choose to listen to the tracks again until they were satisfied with their choices. They were also asked to provide a brief explanation for each of their rankings. Afterwards, they were interviewed about the experience. Once the interviews were complete, participants were debriefed and informed of which text each track was meant to convey.

Results Participants justified their rankings by describing the abstract concepts that each musical track brought to mind for them personally. For three of the six text passages (*Tower*, *Dungeon*, *Emerald City*), all concepts identified by participants were exact matches of concepts extracted by the system. For instance, in the *Tower* passage, participants indicated that they were listening for a "roaring fire", "wind", and some indication of "stairs". A subset of the system's extracted concepts included "roaring fire", "windy nights", and "winding mahogany staircase". In the remaining three tracks, the system identified either only some of the concepts identified by participants, or concepts that were semantically related but not perfect equivalents. As an example, the system extracted "fragrant grass", "cool wind" and "bright day" from the *Elven Realm* passage rather than the participant's suggestion of "springtime".

Five out of six tracks were highly ranked (first or second choice) as evocative of their original source text by at least one participant. However, the majority of participants agreed about which track best represented a passage for only four of the six (*Tower*, *Emerald City*, *Cave*, *Entering Fantasy World*). By far the most difficult track for participants to match correctly to its input text was the *Entering Fantasy World* passage. Interestingly, participants also refrained from naming concrete evocative concepts for this passage, with several choosing instead to listen for an overall ominous sound.

Discussion

In this paper, we presented a system that interprets the semantic content of input texts, and uses this knowledge to manipulate and organize retrieved sound files into novel pieces of music. Preliminary interviews and experimental assessment suggest the system's ability to successfully identify relevant explicit information in texts is promising. Generally, however, not being able to understand and act on implicit information (such as narrative themes, emotions, and beliefs) is a fault of the current implementation. Another failing is that of disambiguating lexical meaning. As one example,

"house" is a musical genre, although it is likely not to be meant as such in the source texts used here. Similarly, the *Dungeon* was described as an "underground room", which resulted in the retrieval of a sound clip related to the London Underground. Future implementations should perfect the art of interpreting text, including that of the informal metadata (title, description, tags, etc.) associated with a certain sound file. As research in open information extraction continues to improve, it is expected that our system will demonstrate better performance in reading and interpreting inputs.

In Study 2, participants indicated that the original texts and generated musical pieces were interesting, but also that the greater lengths of each potentially contributed to cognitive load and boredom. Features such as passage length and writing style may thus have influenced how the texts were perceived. *Entering Fantasy World*, for example, was the longest passage (778 words), and this factor may have caused participants to not attend to explicit descriptions in the original text as closely. Overall, participants appeared to describe the passages in more abstract terms when they were longer, which is likely due to memory abstraction (Bransford and Franks 1971).

Convincing sound design, too, is a complex field. Experts recognize that sometimes the most appropriate sound for a target experience is an exaggerated imitation or otherwise mere representation of the real thing (Sonnenschein 2001). We encountered this obstacle with the present system, as some participants in Study 2 unknowingly interpreted certain sounds as distinct from their intended target. For instance, the retrieved sound effect of snowflakes falling in *Entering Fantasy World* was mistaken for a "crackling fire in the background". Future research should seek to identify factors which influence reader and listener perception within similar contexts.

More broadly, future work should thoroughly categorize the space of extra-inspired systems and investigate how, and to what degree, state-of-the-art creative systems might automatically learn from each other. This learning process should take into account their diverse range of underlying philosophies, key principles (of creative process, system design, etc.), and final products. It is the author's hope that these next steps will build toward creative systems capable of meaningfully inspiring each other, as well as the rest of the world.

References

al Rifaie, M. M.; Bishop, J. M.; and Caines, S. 2012. Creativity and autonomy in swarm intelligence systems. *Cognitive Computation* 4(3):320–331.

Bisig, D.; Neukom, M.; and Flury, J. 2007. Interactive swarm orchestra. In *Proceedings of the Generative Art Conference. Milano, Italy*.

Boden, M. A. 2009. Computer models of creativity. *AI Magazine* 30(3):23.

Bornhofen, S.; Gardeux, V.; and Machizaud, A. 2012. From swarm art toward ecosystem art. *International Journal of Swarm Intelligence Research (IJSIR)* 3(3):1–18.

Boyd, J. E.; Hushlak, G.; and Jacob, C. J. 2004. Swarmart: interactive art from swarm intelligence. In *Proceedings of the 12th annual ACM international conference on Multimedia*, 628–635. ACM.

Bransford, J. D., and Franks, J. J. 1971. The abstraction of linguistic ideas. *Cognitive Psychology* 2(4):331–350.

Brown, S.; Martinez, M. J.; and Parsons, L. M. 2006. Music and language side by side in the brain: a PET study of the generation of melodies and sentences. *European Journal of Neuroscience* 23(10):2791–2803.

Codognet, P., and Pasquet, O. 2009. Swarm intelligence for generative music. In *11th IEEE International Symposium on Multimedia (ISM'09)*, 1–8. IEEE.

Davis, H., and Mohammad, S. M. 2014. Generating music from literature. *arXiv preprint arXiv:1403.2124*.

Delgado, M.; Fajardo, W.; and Molina-Solana, M. 2009. Inmamusys: Intelligent multiagent music system. *Expert Systems with Applications* 36(3):4574–4580.

Dubnov, S., and Surges, G. 2014. Delegating creativity: Use of musical algorithms in machine listening and composition. In *Digital Da Vinci*. Springer. 127–158.

Eno, B. 1978. Music for airports liner notes. *Music for Airports/Ambient* 1.

Gangemi, A. 2013. A comparison of knowledge extraction tools for the semantic web. In *Extended Semantic Web Conference*, 351–366. Springer.

Gerhard, D. 2003. *Pitch extraction and fundamental frequency: History and current techniques*. Regina: Department of Computer Science, University of Regina.

Gervás, P. 2001. An expert system for the composition of formal Spanish poetry. *Knowledge-Based Systems* 14(3):181–188.

Granito, G. Generating music through image analysis. *Scientia Review*.

Harmon, S. 2017. *Narrative Encoding for Computational Reasoning and Adaptation*. Ph.D. Dissertation, University of California, Santa Cruz.

Pappas, C. Generation of music through images. *Scientia Review*.

Rangarajan, R. 2015. Generating music from natural language text. In *Digital Information Management (ICDIM), 2015 Tenth International Conference on*, 85–88. IEEE.

Roche, B. 2012. Frequency detection using the FFT (aka pitch tracking) with source code.

Sapp, C. S.; Smith, J. O.; Chafe, C.; and Selfridge-Field, E. 2011. *Computational methods for the analysis of musical structure*. Stanford University.

Shmulevich, I., and Yli-Harja, O. 2000. Localized key finding: Algorithms and applications. *Music Perception: An Interdisciplinary Journal* 17(4):531–544.

Sonnenschein, D. 2001. *Sound design*. Michael Wiese Productions.

Temperley, D. 1999. What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception: An Interdisciplinary Journal* 17(1):65–100.

Thorogood, M., and Pasquier, P. 2013. Computationally created soundscapes with audio metaphor. In *Proceedings of the Fourth International Conference on Computational Creativity*, 1.

Thorogood, M.; Pasquier, P.; and Eigenfeldt, A. 2012. Audio metaphor: Audio information retrieval for soundscape composition. *Proc. of the Sound and Music Computing Cong.(SMC)*.

Walker, B. N.; Kim, J.; Pendse, A.; et al. 2007. Musical soundscapes for an accessible aquarium: Bringing dynamic exhibits to the visually impaired. In *ICMC*.

Wiggins, G. A.; Pearce, M. T.; Müllensiefen, D.; et al. 2009. *Computational modeling of music cognition and musical creativity*.

Zhu, Y., and Kankanhalli, M. S. 2006. Precise pitch profile feature extraction from musical audio for key detection. *IEEE Transactions on Multimedia* 8(3):575–584.