2011

Computer Music Composition using Crowdsourcing and Genetic Algorithms

Jessica Faith Keup
Nova Southeastern University, jessicakeup@gmail.com

Computer Music Composition using Crowdsourcing and Genetic Algorithms

by

Jessica F. Keup

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in
Computer Information Systems

Graduate School of Computer and Information Sciences
Nova Southeastern University

2011

We hereby certify that this dissertation, submitted by Jessica F. Keup, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.

Maxine Cohen, Ph.D. Date
Chairperson of Dissertation Committee

Sumitra Mukherjee, Ph.D. Date
Dissertation Committee Member

Maria Niederberger, Ph.D. Date
Dissertation Committee Member

Approved:

Amon Seagull, Ph.D. Date
Interim Dean

Graduate School of Computer and Information Sciences
Nova Southeastern University

2011

An Abstract of a Dissertation Submitted to Nova Southeastern University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Information Systems

Computer Music Composition using Crowdsourcing and Genetic Algorithms

by

Jessica F. Keup

September 2011

When genetic algorithms (GA) are used to produce music, the results are limited by a fitness bottleneck problem. To create effective music, the GA needs to be thoroughly trained by humans, but this takes extensive time and effort. Applying online collective intelligence or “crowdsourcing” to train a musical GA is one approach to solve the fitness bottleneck problem. The hypothesis was that when music was created by a GA trained by a crowdsourced group and music was created by a GA trained by a small group, the crowdsourced music would be more effective and musically sound. When a group of reviewers and composers evaluated the music, the crowdsourced songs scored slightly higher overall than the songs from the small-group songs, but with the small number of evaluators, the difference was not statistically significant.

Acknowledgements

I owe a debt of gratitude to Drs. Maxine Cohen, Maria Niederberger, and Sumitra Mukherjee for their advice, insight, expertise, and help as my committee. Thanks to Dr. Terry Countermine, Adam Ogle, Carolyn Novak, Jeff Roach, Kellie Price, Mike Lehrfeld and the rest of the Computer and Information Sciences Department at East Tennessee State University for encouraging me on this path and providing technical advice. The encouragement and camaraderie of fellow Nova GCIS students was very valuable to me.

Thank you to Dr. Edith Seier for answering statistics questions and to Zach Smith for proofreading. I appreciate the help from each of the research participants. Finally, I cannot thank Erik enough for his support and patience.

Table of Contents

Appendixes

List of Tables

Tables

List of Figures

Figures

Chapter 1

Introduction

Problem Statement and Goal

Artificial intelligence genetic algorithms (GA) can be used to produce music, and the fitness function that guides the generational evolution may be either pre-programmed rules or ratings of human preference. Computer-created music that is measured against existing music compositions or adherence to rules regarding voice leading, harmonic progressions, and so forth tend to be homogeneous and non-interesting (Biles, 2007; Roads, 1985). It is lacking because "[music] distinguishes itself by the focus on human emotions and aesthetics – qualities that are not fully understood and which are difficult to describe mathematically," (Jensen & Haddow, 2011, p. 41), and "music requires aesthetic judgments which are not easy to model and implement in the form of an algorithm" (de Freitas & Guimarães, 2011, p. 419).

The alternative technique, known as an interactive GA (Unehara & Onisawa, 2003), is to allow human listeners to gauge the quality of a composition; "fitness boils down to deciding the merit of a piece of music, and this is inherently subjective" (Biles, 2007, p. 41). Existing methods of GA music creation using human fitness functions can produce more effective music than GAs without human input; however, they are limited in scope and quality due to a fitness bottleneck. (Biles 2007; Chen, 2007; de Freitas & Guimarães, 2011; Fu, Wu, Chen, Wu & Chen, 2009; Gartland-Jones & Copley, 2003; Jensen & Haddow, 2011; Khalifa, Khan, Begovic, Wisdom, & Wheeler, 2007; McDermott, O'Neill, & Griffith, 2010; Oliwa, 2008; Unehara & Onisawa, 2003).

The fitness bottleneck occurs because humans must attentively listen and precisely rate a substantial amount of audio information to train a musical GA; they may take too long or be overwhelmed in doing so. The fitness bottleneck problem has repeatedly shown up as a limiting factor in musical GA research (Biles, 2007; de Freitas & Guimarães, 2011; Tokui & Iba, 2000). By training GAs that were successful (other than fitness bottleneck shortcomings) for music composition with online collective intelligence, or "crowdsourcing", it may be possible to supply the algorithm with adequate training data without requiring much input from any one evaluator.

This research was intended to show whether applying crowdsourcing to the human review fitness function of musical GAs yields more effective music, as compared to the small groups typically chosen to provide feedback to a compositional interactive GA. Those small groups are limited by a fitness bottleneck, because it takes a great deal of time and effort to fully train the GA. This was accomplished by establishing an interactive GA that creates music in a two-part chorale-like style. The music was intended to be electronically generated and not performed live. A small group of reviewers worked on one instance of the musical interactive GA, as a control group, and those results were compared to a crowdsourced instance of the same musical interactive GA. The resulting music was subjectively rated by another set of reviewers, as well as composers, to show which music the listeners considered more effective.

The hypothesis was that when comparing music created by an interactive GA trained by a crowdsourced group and music created by an interactive GA trained by a small group, as in previous research, the crowdsourced music would be more effective and musically superior. It was tested by gathering feedback from composers and from non-musically trained reviewers.

Relevance and Significance

Music production with computer assistance is important because the music and the creative process used to compose the music can be enjoyable. By giving subjective feedback on chromosomes of a musical interactive GA, the average person may be granted a creative opportunity that they would not have had otherwise (Unehara & Onisawa, 2003) because "[t]he experience of creating music is another side of musical entertainment that is a demanding aspect for a novice population," (Ning & Zhou, 2010, p. 13:1). They are able to contribute to the production of a piece of music, and indirectly collaborate with others, without needing any background knowledge in music theory and composition (Chen, 2007; Yee-King, 2000). Other current computer music research also shares the goal of making creative input accessible for amateurs and non-musicians, either for groups (Miletto, Flores, Pimenta, Rutily, & Santagada, 2007; Tanaka, Tokui, & Momeni, 2005) or individuals (Nichols, Morris, & Basu, 2009), though these do not make use of evolutionary computing.

This research also served to investigate another use for crowdsourcing, which appears frequently in current literature. Others have applied crowdsourcing to problems in the musical domain, such as rating music popularity (Bhagwan, Grandison, & Gruhl , 2009; Xia, Huang, Duan, and Whinston, 2009) and music similarity (Urbano, Morato, Marrero, & Martin, 2010), but only in one project has it been applied to music creation (DarwinTunes, 2010b; DarwinTunes, 2010c; DarwinTunes, 2010d).

Barriers and Issues

In the current literature, music production using interactive GAs has been limited by the fitness bottleneck. It takes too much time and effort to review a sufficient number of musical evolutions. Due to human capabilities and cognitive limitations, there are a number of contributing sub-problems. First, since auditory information is presented serially, it takes more time to listen to each music sample than it would for visual tasks such as viewing a photograph. One can glance over an image quickly to get the gist of it, but a second of a song does not reveal as much information.

Also, human short term memory is quite limited, usually only holding about seven pieces of information (Miller, 1956). Since it takes so much longer to listen to a song than to view an image, it also becomes much more difficult for a person to make comparisons between songs. For instance, four images can be viewed on the same screen and quickly compared, whereas four song excerpts cannot be heard simultaneously, and by the time the fourth one is heard, the first one is mostly forgotten (Biles, 2007).

This limitation could be mitigated by having trainers listen to the songs many times and committing them to long term memory (Sharp, Rogers, & Preece, 2007; Shneiderman, Plaisant, Cohen, & Jacobs, 2009). Again, though, for GAs, a great deal of music must be rated throughout the many generations, so this would not be an effective solution to the problem.

Finally, concentration is a limiting factor. “Listening carefully and critically to music requires a level of concentration that most people seldom demonstrate” (Biles, 2007, p. 43). Therefore, prolonged listening and rating sessions would be ineffective for GA training.

Research Questions

The primary research question was as follows: When music that is created by a GA trained by a crowdsourced group is compared to music created by a GA trained by a small group, is the crowdsourced music more effective? Other information was gathered as well, namely, qualitative data regarding peoples' opinions about music stemming from a crowd-trained compositional interactive GA.

Limitations

There was a potentially problematic issue in dealing with music created with input from a large group of people. Because a crowdsourced group has varying opinions and preferences about music, the results may therefore turn out to be averaged and homogenous: reflecting a bit of input from everyone but satisfying the requirements and preferences of no one. In other words, "composing music that is loved by everyone is an extremely difficult task, if not impossible" (Chen, 2007, p. 9). It was hoped that narrowing down the genre to two-part chorale-like music would lessen this problem, because at least the listeners approached the music with a similar mindset and the interactive GA was not trained with input from every possible type of music.

After the music was created, the reviewers (and perhaps even more strongly for the composers) had individual tastes that may not have match the crowdsourced average.

No research specifically addressing this issue has been found. There were not enough reviewers or composers to make the results statistically significant.

Definition of Terms

- CAPTCHA – A program to recognize real human users and tell them apart from bots. Most CAPTCHAs involve reading distorted letters and entering them into a text box. It is an acronym that stands for "Completely Automated Public Turing Test To Tell Computers and Humans Apart" (Carnegie Mellon University, 2010).
- Chorale – "The congregational hymn of the Lutheran church", a style of music usually sung in four parts and set to sacred text (Sadie, 1988, p. 152). The chorale form, and variations thereof, appears frequently in music, particularly from Baroque composers in the 17th and 18th centuries.
- Chord - multiple notes sounding together (Sadie, 1988, p. 152).
- Drift – A phenomenon in genetics where changes occur in a small population due to random chance, rather than by natural selection (Encyclopedia Britannica, 2011).
- Genre – "a category of artistic, musical, or literary composition characterized by a particular style, form, or content" (Merriam-Webster, 2011).
- Harmonic progression – a coherent series of chords (Sadie, 1988, p. 598).
- HIT – a Human Intelligence Task on mTurk (Amazon.com, Inc., 2010a).
- mTurk – Amazon Mechanical Turk, a crowdsourcing marketplace (Amazon.com, Inc., 2010a).

- Music theory – “the study of the theoretical elements of music including sound and pitch, rhythm, melody, harmony, and notation” (dictionary.com, 2011).
- Octave – a range of notes wherein the highest one has twice the frequency of the lowest. Alphabetical note names (A – G) repeat once per octave (Sadie, 1988, p. 534).
- Pitch – the frequency of a sound. Musically, a pitch can be identified by an octave and note name (A – G) (Sadie, 1988, p. 581).
- Rest – A mark indicating silence and the absence of a note (Sadie, 1988, p. 623).
- Turker – a worker on mTurk who completes HITs in return for payment (Nowak & Rüger, 2010; Wikipedia, 2011).
- Voice leading – a strategy for constructing polyphony – that is, multiple, independent voices or parts (Sadie, 1988, p. 813). It is also known as part-writing (Sadie, 1988, p. 559).

Summary

Chapter 1 described the research problem, significance, a brief summary of the experimental setup, barriers, issues, limitations, and definitions of terms. When GAs are used to create music, the fitness function may be programmatic or guided by human input. The fitness bottleneck has been a limitation so far in musical genetic algorithms that do not rely solely on programmatic fitness functions. This research was intended to evaluate crowdsourcing as a possible solution to the fitness bottleneck. Two musical GA instances, one crowdsourced and one trained by a small group, were run; the resulting music was rated by a group of general reviewers and a group of composers to see which was more effective.

Chapter 2 addresses the background literature regarding musical GAs (both programmatic and interactive) and the fitness bottleneck problem in compositional GAs. The benefits and disadvantages of crowdsourcing are discussed, as well as examples of where it has been used successfully. The proposed contribution of this research is described, followed by a more detailed description of the research plan with methods, procedures, and formatting of results, in Chapter 3. Chapter 4 provides the results of the experiment, summarizing the ratings and comments from reviewers and composers. Then, Chapter 5 explains the significance of the results and suggests questions for future research.

Chapter 2

Review of the Literature

Computer Music

Research in computer music began as early as the 1960’s (Gill, 1963; Mathews, 1963; Seay, 1964). Computers were found to be useful tools for sound generation and as aids to human composers (Shneiderman et al., 2009). As for music creation by computers, it was believed that the compositional algorithms must be based on the voice leading and harmonic progression rules from music theory. However, it was not possible to fully notate all rules, so such results were limited to the rules and methods of a single composer (Gill, 1963) or augmented with elements of randomness (Mathews, 1963).

As computers advanced and artificial intelligence was further developed, researchers realized AI had several musical applications – composition, performance, music theory, analysis, and digital sound processing (Meehan, 1979; Roads, 1985). GAs in particular have been successful for musical computer learning. Referring to transcription, Reis and Vega (2007, p. 1965) said “genetic algorithms are perfect candidates for solving this problem.”

In fact, music composition can be considered a type of search problem with no optimal solution, for which GAs are rather well suited (Biles, 1994; Gartland-Jones & Copley, 2003). “. . . [A] typical musician ‘knows what she likes’, and the aesthetic sense guides the search through the various problem spaces of notes, chords, or voicings” (Biles, 1994, p. 131). In a sense, the evolution of music in GAs mimics the natural order of cultural development, since many people work on compositions, building off what has been successful in the past and adding novel “mutations” (DarwinTunes, 2010). Those ideas, themes, and techniques that are successful, effective, popular, or artistically appreciated survive to be tried in new configurations in new compositions.

In spite of the pioneers’ research, the difficulties of music notation, theory codification, and modeling remained, and it was not possible to digitally duplicate the human composition or performance processes (Kirk & Miranda, 2009; Meehan, 1979; Roads, 1985). Presently, some researchers are still working on the programmatic fitness function approach to composition, where rules of music theory are notated as clearly and completely as possible and AI (usually GA) applications are made to follow them (Birchfield, 2003; de Freitas & Guimarães, 2011; De Prisco, Zaccagnino, & Zaccagnino, 2010; Jensen & Haddow, 2011; Khalifa & Al-Mourad, 2006; Khalifa et al., 2007; Nelson 2003; Nelson, 2005; Oliwa, 2008).

Some argue that the results of compositional GAs that use programmatic fitness functions are non-musical, because artistic, interesting music often does not strictly follow all the rules of voice leading and chord progressions (Biles, 2007; Roads, 1985). Creativity and creative solutions often come from knowing when to “break the rules”. Even deciding which rules to include and how to encode them in the first place – “creating an aesthetically conscious measure of fitness” (de Freitas & Guimarães, 2011, p. 419) – is a difficult research problem (Miranda, 2004).

Bottleneck in Compositional Interactive GA Fitness Functions

An alternative research path has developed, where humans provide insight but computers do the majority of the work with interactive GAs. Musical information is written in digital chromosomes, which are mated and mutated over successive generations, with the best surviving. However, when human insight is used as the fitness function for the interactive GA, there is a major bottleneck.

Tokui and Iba (2000, p. 229) described the problem as follows: “[T]he common difficulty in the practical use of IEC [interactive evolutionary computation] is the human fatigue. Since a user must work with a tireless computer to evaluate each individual in every generation, he/she may well feel pain. It is the biggest remaining problem to reduce the psychological burden on users.” A number of compositional interactive GAs are described below, along with the fitness bottleneck’s impact on them, if applicable.

GenJam, a genetic algorithm which produced jazz solos, was an early musical interactive GA project that mentioned the fitness bottleneck (Biles, 1994). It was only designed for use with one evaluator. The author suggested future work that would use neural networks or an initial seed population to cut down on the amount of fitness listening required for the evaluator (Biles, 1994).

Yee-King’s (2000) Audioserve project was intended to create sound rather than music, though it still has a great deal in common with other compositional interactive GAs and was accessed via a web interface. Some notable differences are a user-adjustable mutation rate, the ability to swap chromosomes with GA instances being trained by other people, and the ability to go back in history to use parents from the past. Perhaps this freedom is due to the domain of FM/AM circuits being more flexible and forgiving than music because it has no rules and fewer user expectations. In any event, Yee-King (2000, Client Program section, paragraph 2) states: “. . . the user can only be expected to effectively audition a small number of candidates at each iteration of the GA."

Unemi (2002) made SBEAT3, an interactive GA intended to help musical novices with composition. It gives users a comparatively large amount of control over the music (e.g. changing key, changing tempo). They note that the time taken for user ratings is burdensome and limits population sizes, since music must be heard in series, unlike other tasks such as image viewing which may be done in parallel.

Legaspi, Hashimoto, Moriyama, Kurihara, and Numao (2007) created a prototype with the Constructive Adaptive User Interface (CAUI) and incorporated the Diverse Density (DD) weighting metric and First Order Inductive Learning (FOIL) heuristic function for multi-part learning. For the tests, evaluators classified training data using six sets of labels (favorable-unfavorable, bright-dark, happy-sad, heartrending-not heartrending, stable-unstable, and beautiful-ugly).

Based on the training data, the system generated new music that matched the labels with 80.6% accuracy for four of the six label pairs, but it was only tested with 11 participants listening to 75 musical songs. Each participant's ratings and preferences were kept separately, so there were essentially 11 instances of the same interactive GA with only 75 songs rated, instead of 1 interactive GA with 825 songs rated. The training, or fitness function creation, was done upfront with existing music, rather than generationally during interactive GA operation (Legaspi et al., 2007).

The authors attributed the 80.6% accuracy to the setup of the music theory constraints. However, effects of tiring or discrimination difficulty in the evaluators, or the small sample of evaluators and small number of songs may also have been factors.

Perhaps if a larger group of evaluators had been used, the accuracy of the training and subsequently created music may have been improved (Legaspi et al., 2007).

There are, additionally, other interactive GA projects that did not specifically mention the fitness bottleneck, either because the sampling was intentionally and admittedly small or because there was no user testing, only the development of a proof-of-concept. For example, Numao, Takagi, and Nakamura (2002) created the Constructive Adaptive User Interface, on which the Legaspi et al., (2007) work is built. It focuses on learning relationships between certain musical constructs and human emotions.

Fitness Bottleneck Workarounds

The research above discussed the fitness bottleneck as a limitation, and there are projects in this domain that have attempted to work around it. For example, Gartland-Jones and Copley (2003) created a musical-building-block application that recomposes itself as blocks are added or removed by a user. They stated that users do not need to explicitly evaluate the music at each step and the bottleneck is removed. However, there is a still a human giving indirect subjective feedback and the application is thereby limited to a small scale.

Unehara and Onisawa’s (2003) interactive GA is a hybrid approach, where a programmatic fitness function is applied first to eliminate clearly poor chromosomes and lessen the work required of users. After the poor chromosomes of a generation are removed, user input is required to select the best chromosomes. Their testing was limited in scope, with six subjects listening to 15 generations of 16-measure chromosomes. They confirmed that the music had subjectively increased in quality by the last generation, but were aware that the fitness bottleneck was a limitation: “The system design using the interactive GA has the problem that users have to repeat simple evaluation. The more users repeat evaluation of musical works, the more users feel fatigue” (Unehara & Onisawa, 2003, p. 86).

Fu et al. (2009) developed the CFE (Composition, Feedback, and Evolution) framework for musical interactive GAs, with the hope of minimizing user input by continuing to create music autonomously after several rounds of user fitness input. They state “it is still hard, if not impossible, to create pleasant music for unskilled people . . . we should make the grading runs as few as possible . . . making a lot of tests by using manpower is not efficient” (Fu et al., 2009, p. 1863 - 1864). They suggest user testing of their system as future work.

Existing Compositional GAs

Several compositional GAs with freely available source code were reviewed. The VARIATIONS algorithm in Perl (Jacob, 2009) and a master’s thesis in C++ from a student at the College of William and Mary (Schoenberger, 2009) were examined, but found to be incomplete or musically primitive.

The work of Numao et al., (2002), Legaspi, Hashimoto, and Numao (2006), Legaspi et al., (2007), and Sugimoto, Legaspi, Ota, Moriyama, Kurihara, and Numao (2008) was explored and the interactive GA developed by this group of researchers seemed relevant to this proposed research, due to the refined iterations of the interactive GA and their related research goals. However, they were contacted as a group with the form on the Architecture for Intelligence Numao Lab website and individually by email, and no one responded.

Spieldose was one complete musical GA for which the source code was freely available. It ran in Matlab and featured several up-front customization parameters, including number of measures, harmonic progressions, number of “invaders” (mutations), size of generations, and type of crossover. It used an interactive fitness function to create melodies with harmonic accompaniment using a MIDI synthesizer sound (GAVAB Research Group, 2007; Sánchez, Pantrigo, Virseda, & Pérez, 2007).

Melodycomposition (MC) was another complete musical GA with source code posted online. It was written in Java with the Genetic Algorithms Package (JGAP) and used a programmatic fitness function (Craane, 2009a; Craane, 2009b; Meffert & Rotstan, 2009). It created melodies (without harmonic accompaniment) with a MIDI synthesizer sound. MC featured even more up-front customization parameters in the user interface: maximum duplicate rests, maximum duplicate notes, proportion of notes to rests, number of major intervals, number of perfect intervals, number of parallel intervals, maximum range, pitch distribution, number of notes, and number of evolutions. All of these options helped customize the music toward a chorale-like style.

Other Applications of AI in Music

Finally, it should be noted that (a) different AI methods besides GAs can be used in music creation and (b) GAs have other musical applications besides music creation. For example, BeatBender uses autonomous percussion agents to create rhythmic patterns (Levisohn & Pasquier, 2008). In a way, it avoids the fitness bottleneck since there is no human input during training. By the same token, it has the same problem as GAs programmed to use rules of voice leading and harmonic progressions for fitness; the result is not musically interesting (Biles, 2007; Roads, 1985).

Of the six GAs for expressive performance that are discussed in Kirke and Miranda's survey (2009), all of them have programmatic fitness functions, rather than human. Another example of a programmatic fitness function is Reis and Vega's (2007) GA for musical transcription where the fitness function is a sum of differences of expected and actual frequencies. McDermott et al. used an interactive GA for sound synthesis (2010). In a limited capacity, GAs have been applied to music processing and listening, too (Biles, 2007). Gabrani, Bhargava, Bhawana and Gill (2008) developed a GA for remixing Indian music; it relies on a programmatic, not interactive, fitness function.

Crowdsourcing

Crowdsourcing is a potential solution to the musical interactive GA fitness bottleneck. First recognized in the mid-2000s, crowdsourcing is a method of solving problems with, and outsourcing work, to the collective online intelligence (Howe, 2006). Su, Pavlov, Chow, and Baker (2007, p.231) recommend it "for training and monitoring machine learning-based applications", as did Lease, Carvalho, and Yilmaz (2011).

Though there are pitfalls such as poor data quality from malicious or uninformed participants, group thoughts and contributions have several benefits. For example, the workload is distributed so that tasks can be completed faster and around-the-clock. Under the right conditions, the group's answers are usually more effective than those of experts

(Surowiecki, 2005). Surowiecki would classify this as a Cognition (as opposed to Cooperation or Coordination) problem because it deals with the averaging or combining of opinions to find a group consensus.

Crowdsourcing is potentially a good fit for this problem because of its success in solving similar types of problems. It is recommended for human review of large amounts of data and collective decision-making and under certain circumstances, the results are moderately high quality, inexpensive, and fast (Alonso, Rose, & Stewart, 2008; Carvalho, Lease, & Yilmaz, 2010; Hintikka, 2008; Ledlie, Odero, Minkov, Kiss, & Polifroni, 2010; Mannes, 2009; Mason & Watts, 2009; Stewart, Huerta, & Sader, 2009; Su et al., 2007). Crowdsourcing has appeared in recent research for purposes such as these:

- labeling articles' search relevance (Ganjisaffar, Javanmardi, & Lopes, 2009)
- collecting creative drawings (Koblin, 2009)
- rating audio and video sample quality (Chen, Wu, Chang, & Lei, 2009)
- tagging location-sensitive queries and points of interest with mobile devices (Yan, Marzilli, Homes, Ganesan, & Corner, 2009)
- describing and organizing geometric shapes (Jagadeesan et al, 2009)
- tracking popular music (Xia et al., 2009)
- rating search result relevance (Alonso et al., 2008)
- collecting user feedback (Kittur, Chi, & Suh, 2008)
- extracting product and brand information (Su et al., 2007)
- tagging specific objects in an image (Von Ahn, Liu, & Blum, 2006)
- remote massage (Chung, Chiu, Xiao, & Chi, 2009)
- assessing visualization design (Heer & Bostock, 2010)

- task delegation in Wikipedia (Krieger, Stark, & Klemmer, 2009)
- collecting volunteer expertise to help the homeless (Li, Buyuktur, Hutchful, Sant, & Nainwal, 2008)
- calculating music popularity (Bhagwan et al., 2009)
- determining the clearest presentation of mashup code for software engineers (Stolee & Elbaum, 2010)
- annotating political campaign ads (Hsueh, Melville, & Sindhwani, 2009)
- training speech recognition software (Ledlie et al., 2010)
- translating text in images (Liu et al., 2010)
- annotating an image corpus with multiple labels (Nowak & Rüger, 2010)
- constructing philosophical concept hierarchies (Eckert et al., 2010)
- indexing films and television shows (Geisler, Willard, & Whitworth, 2010)
- voluntary translating by IBM employees (Stewart, Lubensky, & Huerta, 2010)
- evaluating musical similarity (Urbano et al., 2010)
- iteratively transcribing and editing copy (Little, Chilton, Miller & Goldman, 2009).

Risks of Crowdsourcing

There is a risk of users entering accidental or malicious bad data, but strategies such as discarding outliers, requiring qualification exams, enabling voting schemes, or asking a user repeat questions to see if they give the same answer improve the accuracy of data collection (Chen et al., 2009; Harper, Raban, Rafaeli, & Konstan, 2008; Heer & Bostock, 2010; Ledlie et al., 2010; Su et al., 2007). Qualification tests can also be given to eliminate obviously unqualified users, but one known problem of crowdsourcing is that while it can be used to gather popular opinion, it cannot be counted on to provide expert opinions (Roman, 2009). Since the GA training step of the proposed research does not require informed choices or specialized knowledge of music theory, qualification tests will be unnecessary.

In many cases, users must be motivated extrinsically to participate. It is difficult to get a new crowdsourcing community up-and-running, so existing sites like Amazon Mechanical Turk (mTurk), InnoCentive, CrowdFlower, Wilogo, fellowforce, BootB, CrowdSPRING and Cloud Crowd can be used instead, where a large workforce is already in place and workers will complete given tasks in exchange for money (Amazon.com Inc., 2010a; BootBe, Inc., 2011; CloudCrowd, 2009; CrowdFlower, 2011; CrowdSPRING, 2011; Eckert et al., 2010; fellowforce, 2007; Innocentive, 2011; Stewart et al., 2009; Wilogo.com, 2011). If the topic is considered to be interesting enough, users are willing to participate and add their input without monetary reward; they may still desire other non-tangible rewards such as status or entertainment (Von Ahn & Dabbish, 2004; Yang, Adamic, & Ackerman, 2008).

The user environment was an unknown variable that could not be controlled in crowdsourcing. The users were not in the same place, nor were they using the same hardware, browser, or operating system. They could be working under less-than-ideal conditions, such as having a slow internet connection, having a low resolution monitor, or being in a loud room.

Amazon Mechanical Turk

mTurk is known as a “micro-task market” since it was made for, and is primarily used for, small tasks that pay a few cents and take seconds to complete (Heer & Bostock, 2010). It has been used as a source of participants in many of the recent crowdsourcing research projects listed previously:

- mCrowd for tagging location-sensitive queries and points of interest with mobile devices (Yan et al., 2009)
- Multimedia QoE evaluation, where audio and video sample quality is rated (Chen et al., 2009)
- Sheep Market where creative drawings are collected (Koblin, 2009)
- Geometric reasoning tasks where shapes are described and organized (Jagadeesan et al., 2009)
- TERC where search results are rated for relevance (Alonso et al., 2008)
- User studies where feedback is gathered (Kittur et al., 2008)
- Tests of graphical perception to evaluate visualizations (Heer & Bostock, 2010)
- Ad annotation for political campaigns (Hsueh et al., 2009)
- Software engineering research, where coding strategies for mashups are evaluated (Stolee & Elbaum, 2010)
- Philosophy ontology construction (Eckert et al., 2010)
- Ratings of music similarity (Urbano et al., 2010)
- Transcriptions and edits of text (Little et al., 2009).

Fitness Bottlenecks with Crowdsourcing

In the field of geometric reasoning, the work of Jagadeesan et al. (2009) serves as a parallel example to this proposed research. They state that humans are good at geometric reasoning tasks such as fitting irregular shapes into the smallest possible space. These tasks can be done programmatically, like music creation, but fall short of the results that can be achieved with human input (Jagadeesan et al., 2009).

In their study, they tested three types of geometric reasoning work: canonical viewpoints, shape similarity, and strip packing. They recruited users on mTurk to complete tasks in each of the three areas. The goal was to set benchmarks of good performance with human input with which to judge programmatic methods. In the end though, they came to a surprising conclusion: "[c]rowdsourcing has proved so effective that in many cases the authors have questions if automated solutions are really required" (Jagadeesan et al, 2009, p. 313).

Music Crowdsourcing

Participation in crowdsourcing, music or otherwise, is further encouraged when the costs of contributing are lower; if users only have to vote, or choose a pre-arranged response, it is easier and requires less commitment than writing comments or contributing new content. Xia et al., (2009) call this Ballot Box Communication (BBC) as opposed to Computer Mediated Communication (CMC). They studied the logs of Internet Relay Chat (IRC) music sharing groups to observe that even without written communication, the users were collaborating in a sense. The music the users had chosen to upload and download showed aggregate trends in music popularity (Xia et al., 2009).

In contrast to users’ active involvement in crowdsourcing, they can also become unknowing participants as part of the crowd. In the music-related example of Sound Index, data on user music-listening behavior is collected from many online sources (social networks, online radio, downloads, sales) to more accurately reflect music popularity than traditional Billboard rankings which rely on music sales (Bhagwan et al., 2009).

Three commercial websites, Last.fm, Pandora Radio, and Spotify use crowdsourcing to gather and make music recommendations (Celma & Lamere, 2008, Last.fm, 2009; Pandora Radio, 2009, Spotify, 2010). On Pandora Radio, the music labeling and classification is done by their own employees, not the general Internet population used in many instances of crowdsourcing. However, ordinary users of the site may create, share, and build upon each others’ “stations” (Pandora Radio, 2009). Likewise on Last.fm, users share playlists and recommendations, join fan groups, and listen to music chosen by those with similar tastes (Last.fm, 2009). Spotify provides similar functionality, where users can share their own playlists and collaborate with others to create new ones (Spotify, 2010).

The success of these sites demonstrates that users are willing to listen to music and rate it in exchange for music that they enjoy and that they have an interest in using sites that direct them to new music based on their preferences and crowdsourcing data. A major difference between these sites and the proposed research is that Last.fm, Pandora Radio, and Spotify provide access to existing recorded music, rather than newly created computer music. DarwinTunes and the Music Information Retrieval Evaluation eXchange (MIREX) Evaluations, described below, are the two music crowdsourcing projects most closely related to this research.

DarwinTunes

Darwin Tunes, developed by MacCallum and Leroi at Imperial College, London, is the only existing large-scale crowdsourced compositional interactive GA (DarwinTunes, 2010b). It consists of an interactive GA developed in Perl and accessible via a web interface. A song is four measures of four beats, and it is presented as a loop (DarwinTunes, 2010d). Songs are rated on a 5-point scale, with the labels “I love it!”, “I like it”, “It’s ok . . .”, “I don’t like it”, and “I can’t stand it!” (DarwinTunes, 2010a). After 20 songs have been rated, the top 10 are mated with crossover and a 1/1500 chance per node of mutation (DarwinTunes, 2010c).

There were multiple concurrently developing population groups, and site visitors were automatically assigned to one. This procedure was in place to prevent drift; they suspected that certain changes might take place regardless of the music ratings, and they will be able to compare the independently evolving populations to compare. Any significant similarities will be attributed to drift (DarwinTunes, 2010c).

The DarwinTunes work is the closest to the research proposed here. There are several key differences, however. DarwinTunes has not been presented in any publications; it appears to be a more informal trial of what happens when a compositional interactive GA is offered to the general public – “What, exactly, will we be looking for in our evolving populations? Frankly, we’re not sure – nor do we have to be” (DarwinTunes, 2010c, paragraph 7).

Consequently, there is no side-by-side comparison of the output from a small group and from a crowdsourced group, and no formal evaluation of the music that has been produced. Their research questions were "[H]ow important is human creative input compared to audience selection? Is progress smooth and continuous or step-like?" (DarwinTunes, 2010b, paragraph 4). They are currently evaluating the results and analyzing the data and have not yet published any findings.

MacCallum ran a smaller-scale site called Evolectronica in 2009, off which DarwinTunes is based. It is no longer active (Evolectronica 2011a, Evolectronica 2011b). The latest official statement on progress in December 2010 said there have been 641 generations of evolution on DarwinTunes (Twitter, 2010). With 100 chromosomes per generation, that means that 64,100 songs have been rated, which is larger than the size of the training group in this research. There are discrepancies in generation numbering, though; the Audio Snapshots page indicates that there are at least 3,060 generations, but it may be that the individual populations they were evolving concurrently are counted separately (DarwinTunes, 2010e). While there is a CAPTCHA in place to prevent automated responses (DarwinTunes, 2010a), there is no assurance that this is truly crowdsourced and not trained by a small group of people rating many songs.

MIREX Evaluations

MIREX is an information retrieval system in which music is rated by similarity to other music.The cost of expert time to rate the similarity between songs was unsustainable, so Urbano et al. (2010) attempted to crowdsource the ratings to non-experts on mTurk. They believed their experiment to be the first music-related task research on mTurk.

Urbano et al. (2010) used preference judgments for ranking, where participants are presented with two choices and they choose the better one. In the manner of a sorting algorithm, the list items (songs, in this case) are sorted one pair at a time, and it prevents listener fatigue from hearing too much at once. On mTurk, they posted 281 HITs with 10 assignments each; there were 70 unique workers, an average of 22 seconds per assignment, and a total time of 37 hours and 40 minutes. As safeguards against malicious or lazy answers, they restricted recruitment to a 95% or higher HIT acceptance rate, gathered multiple ratings on the same pairs of songs from different Turkers, and discarded responses that showed bot-like behavior.

The results of the MIREX similarity judgments showed that while agreement between non-experts on mTurk was considerably lower than agreement between experts, the averaged rankings of the mTurk non-experts were similar to the rankings of the experts. The total cost for the mTurk experiment was $70.25; otherwise, the cost would have been the equivalent of 70 hours of expert time. The commonality between their experiment and this research was the collection of non-expert opinions about short songs on mTurk; the differences were that they asked about song similarity and used that data for a sorting algorithm, while this research asked about song quality and used that data in a GA.

Contribution

This research contributes to both the computer music and crowdsourcing literature. For computer music, it demonstrates an alternative technique for compositional interactive GAs that avoids the fitness bottleneck. It allows people without musical training or composition experience to contribute to the creation of new music.

For crowdsourcing, it shows another possible application where the wisdom of crowds and small bits of input from lots of people can solve problems effectively. Crowdsourcing has been shown to solve a wide variety of problems, and by demonstrating its effects on this particular problem, the research will show the efficacy of crowdsourcing for creative collaboration and music creation. Others have applied it to problems in the musical domain, such as rating music popularity (Bhagwan et al., 2009; Xia et al., 2009) and music similarity (Urbano et al., 2010).

Crowdsourcing has only been applied to an interactive GA for music creation in one other project (DarwinTunes, 2010b; DarwinTunes, 2010c; DarwinTunes, 2010d). This research differs from DarwinTunes in that DarwinTunes was an exploratory venture posted on a custom crowdsourcing site, rather than mTurk or another marketplace. The researchers wanted to see what would happen when a large number of people gave input to a musical GA online; their setup had less management and planning, and their results were not compared to a small-group control condition.

Summary

Computers have been used as an aid for musical endeavors in several ways. They can help with sound synthesis, sound processing, music theory analysis, composition, and performance. AI has been applied to creation and composition. There are two tactics, sometimes used in conjunction, for creating music with GAs. The fitness function may be programmatic, based on music theory (e.g. rules regarding chord progressions and voice leading). Alternatively, it may be subjective, relying on user feedback, which is known as an interactive GA. Interactive GAs are subject to a fitness bottleneck because it takes too much time and energy for humans to rate and review many generations of chromosomes, thus limiting the effectiveness of the music created by the interactive GA.

Crowdsourcing is the outsourcing of work to the collective online intelligence. A wide variety of work types have been crowdsourced, and it is most effective for use in tasks that are easy for humans but difficult for computers, such as image tagging and relevance rating. It is subject to misuse by careless or malicious users, but there are techniques to check responses for validity and mitigate that risk. Amazon mTurk is a crowdsourcing community where requestors can post tasks and Turkers will do the work in exchange for small payments.

Crowdsourcing has been applied to the GA fitness bottleneck in other domains and it has been applied to music recommendation systems. In DarwinTunes, it was tried with a compositional interactive GA to see what type of music could be created. Their music may be found at http://darwintunes.org/audio-snapshots, but their work was only exploratory and not to answer any particular research questions. In another study, music similarity-rating tasks were posted on mTurk.

This chapter has covered the relevant existing literature in the intersections of the music, GA, and crowdsourcing domains. The next chapter describes the research methodology in detail. Chapter 4 summarizes the results of the research, and Chapter 5 explains the significance of the results and recommendations for future work.

Chapter 3

Methodology

Research Methods

The hypothesis was that crowdsourcing would alleviate the fitness bottleneck problem for compositional interactive GAs. In other words, music that is created by a crowdsourced compositional interactive GA will be more effective than past music that was generated from compositional interactive GAs, which had been trained by individuals or small groups and were thereby limited in quality. To test the hypothesis, an experimental study was conducted.

In the study, the control was a compositional interactive GA trained by a small group, while the experimental condition was a compositional interactive GA trained by a very large group (i.e. crowdsourcing). Crowdsourcing is the independent variable, the effects of which were tested on compositional interactive GAs. The primary output from the study was two sets of music. However, the music could not be objectively programmatically compared to determine which are best. Musical effectiveness is subjective and must be rated by humans.

Therefore, two more groups of people were recruited to rate the music in a blind study. A more detailed description of the research steps is explained below, including the genetic algorithm choice, fitness functions, musical genre choice, prototype creation, task setup, precautions, genetic algorithm training, and recruitment and instructions for trainers, reviewers, and composers. Many design decisions are based on work by Legaspi et al. (2007) in “Music Compositional Intelligence with an Affective Flavor”, due to similarities between the studies and the fact that they provided a helpful level of detail for replication. The Nova Southeastern University (NSU) and East Tennessee State University (ETSU) Institutional Review Board (IRB) approval letters may be found in Appendices A and B, respectively.

Procedures

Genetic Algorithm Choice

As discussed in the literature review section, several pre-existing musical interactive GA potentially met the requirements for this study. Melodycomposition (MC) was selected for several reasons, though Spieldose was also a promising alternative. MC has more customization built-in for setting up the starting point. Neither was perfectly suited to the planned chorale-like style, because Spieldose produces a single melody with harmonic accompaniment and MC produces a single melody. This research requires two simultaneous melodies, so either GA would have needed modification in that respect. Both GAs create export files in the MIDI format.

Since MC was implemented in Java, it could be more effectively integrated with mTurk's Java API. Spieldose was implemented in the MATLAB programming language, which placed more restrictions on the way it must be run on the server (a potential licensing issue) and would have been much more difficult to integrate with mTurk. The biggest disadvantage of modifying MC is that Spieldose uses a human fitness function that MC lacks.

Melodycomposition (MC)

Below is a representation of a gene from a sample chromosome from MC. This run was setup to contain 24 notes, so the chromosome contains 24 notes/genes. Each note is represented by the pitch (any of the 12 chromatic notes represented by letter name or a silent rest), the octave (an integer in the range of 0 – 8), and the relative duration (sixteenth, eighth, quarter, half, whole).

```
[F#:7:QUARTER]
[A#:4:QUARTER]
[F#:6:EIGTH]
[A#:4:QUARTER]
[F#:5:EIGTH]
[G#:6:EIGTH]
[B:4:QUARTER]
[E:5:QUARTER]
[D#:5:QUARTER]
[A:6:QUARTER]
[D#:5:EIGTH]
[C#:5:QUARTER]
[A#:6:EIGTH]
[C#:5:QUARTER]
[C#:6:QUARTER]
[D#:6:EIGTH]
[D#:5:QUARTER]
[REST:5:QUARTER]
[A#:7:EIGTH]
[REST:4:QUARTER]
[F#:4:QUARTER]
[A#:4:EIGTH]
[G#:6:EIGTH]
[REST:6:EIGTH]
```

In MC, the GA is setup as follows. Some constants are set, such as octave range (4 – 7 by default), possible note values (only quarter and eighth, by default), and population size (40 by default). The initial population is filled with chromosomes generated at random within those constraints.

When the GA was modified, it was made into an interactive GA, where the most fit chromosomes were the ones marked best by the users; the programmatic fitness function was left in place with relatively lower weight to supplement the user ratings. This was important since users only provide ratings of best, middle, and worst out of three, and there would have been three large groups of equally rated music songs without supplemental programmatic fitness. Additional programmatic fitness rules were added to the default rules that came with MC; the final guidelines are described in *Fitness Functions*.

Out of each generation, the best chromosome was preserved for the next generation, and the rest of the next generation was bred from the other chromosomes. The genetic operators used were crossover and mutation. Crossover occurred at a random point between two chromosomes at a given rate, and the default is 35%. Mutation occurred at a given rate and the default was 1 in 12 genes being mutated. If a gene was selected at random, each atomic element – note, octave, and duration, in this case – had a 50% chance of being changed to something else at random. These operations were left with their default settings.

The selection and replacement strategies are written in the JGAP package, not MC. The termination condition was reached when a specific number of evolutions have been completed, which were 11 and 200, for the small group training and crowdsourced training, respectively. The population size and termination condition were modified to meet the crowdsourcing requirements detailed in *Genetic Algorithm Training* below.

Fitness Functions

In order to avoid a population of 75 having 25 best, 25 middle, and 25 worst chromosomes, lower-weighted programmatic fitness functions were put in place. They were also intended to guide the music toward an instrumental chorale-like style with only two parts

- After Large Skip Strategy – After a skip of a 5th or larger in a melody, the next motion should be stepwise in the other direction. An augmented fourth should never occur in ascending motion; its inversion, the descending diminished fifth, should be followed by a step in ascending motion. The leading tone (*ti*) should be followed by the tonic (*do*) one half-step above it.
- Consecutive Skips Strategy – Two consecutive skips should not add up to more than an octave (within a melody line).
- Global Pitch Distribution Strategy – Within a melody line, 98% of the intervals should be less than 9 half steps.
- Human Review Strategy –Melodies are more likely to be bred if they are rated favorably by users on Amazon Mechanical Turk.
- Interval Strategy –Within a part, at least 50% of the notes should follow stepwise motion, rather than skips.
- Parallel Motion Strategy -There should not be parallel motion in octaves or perfect 5ths. The two parts should not arrive at a 5th from the same direction nor skip at the same time. The parts should not cross.
- Proportion Notes/Rests Strategy – A melody should be at least 90% notes and not more than 10% rests.

- Range Strategy – The bass part should be in the two octaves below middle C and the soprano part should be in the two octaves above middle C. There was hard limit, not a fitness function, of absolutely nothing lower than octave two octaves below middle C and absolutely nothing higher than two octaves above middle C. It was possible, though unlikely, for bass to appear above middle C or soprano below it due to the Parallel Motion Strategy.
- Repeating Notes Strategy –More than one rest in a row or more than two repeating notes in a row are discouraged.
- Scale Strategy –All melodies are in C Major, and notes outside the scale (accidentals) should not appear.
- Strong Beats Strategy – Beat one in each measure should not be an eighth note. The last note should not be an eighth note or quarter note.

Musical Genre Choice

Limiting the music to one genre was intended focus the evolutions into a better end result. In other words, some algorithmic fitness was left in to guide the general direction of the music. Two-voiced chorale-like music has been chosen for this purpose, and this decision was made for several reasons.

First, MC is not set up for varied instrumentation; adding it would significantly increase the GA complexity. Because the instrumentation is not varied, some popular genres (i.e. rock, country) would not have been appropriate because they depend on the instrumental timbre and percussive instruments to create the overall sound, while the melodic content is less important.

On the other hand, many types of Baroque, Classical, Romantic, and 20th-century music would have intricate rules for fundamental algorithmic fitness and would have been less accessible to non-musically-trained listeners. Chorales (with some conditions) seemed to fit all the required criteria. They were suitable and intended for homogenous instrumentation. Their overlap with church hymns and folk music made them familiar-sounding to the general audience, but their numerous occurrences in art music (e.g. Baroque and Classical) mean that they were taken seriously by, and were of interest to, professional musicians. Finally, they had a well-defined set of guidelines with which to initially set up the GA, which have been appealing to and used by other compositional GA researchers (De Prisco et al., 2010).

A few aspects of the traditional chorale style were modified for use in this study. It was two-voiced, with only soprano and bass, so as not to make the GA rule encoding overwhelming; the alto and tenor were omitted. The soprano and bass are considered the most aurally salient of the four parts (Huron, 2001). As previously stated, voice-leading and harmonic rules are relatively well-defined, but also quite extensive. This reduction in parts kept the GA setup workload from becoming a barrier to progress, while still allowing harmony and implied chords.

A second difference was the absence of text and therefore religious context. Though traditional chorales were set to text, words will not be relevant to the focus of this work and will therefore be omitted. The parts were kept approximately in the soprano and bass ranges of a choral work. Thus, the genre resembled two-part instrumental music (similar to a four-part chorale in range and voice-leading, with the omission of text and the alto and tenor lines missing).

Prototype Creation

Naturally, the interface was a webpage; this is standard for crowdsourcing and helpful for contacting a wide variety of users. Web pages were constructed for communication between the end user and the musical interactive GA. The user interface was written in XHTML, CSS, PHP, and JavaScript, while the processing and connection between mTurk and the interactive GA were accomplished with Java, using the Eclipse IDE. The storage of users' answers and music from the interactive GA required a database, for which MySQL was used. Since MC creates MIDI files, the MIDI files first had to be converted to .wavs as an intermediary step, then to .mp3s for ease of playing in a browser and for sound consistency on participant computers.

As with other multimedia HITs such as CastingWords audio transcriptions and the Flash application used by Heer and Bostock (2010), neither the text nor the audio files of the HIT were stored by Amazon (Amazon.com, Inc, 2010; CastingWords, 2010, Urbano et al., 2010). While the HITs appeared on mTurk, the questions and links to mp3 files were shown within an iframe on the mTurk page and hosted on www.jessicakeup.com through Rackspace. The audio files were linked to and served from that same server.

mTurk provides several ways to post HITS. They may be posted through a GUI on mturk.com, through a command line interface, or using one of the four SDKs in Java, Ruby, Perl, or .NET (Amazon, Inc., 2010d). Due to the need for communication between the GA written in Java and mTurk, the Java SDK was selected. It proved to be a minor problem (explained in Experimental Setup, below) that the tasks could not be posted as multiple assignments in the same HIT, but had to be posted as separate HITs. They could not be posted as one large HIT with many assignments because they were posted generation-by-generation, instead of all at once.

There were slight differences in the source code of the two interactive GA instances. The biggest difference was that the test GA posted HITs to the live version of mTurk, and the control GA posted to the sandbox version. By posting the control HITs to the sandbox, the trainers from both groups were able to see the same screens (for consistency), and a second interface did not have to be created. Other differences included the database to which they wrote and the number of generations they were to complete. The source code for the small control group program may be found at http://www.jessicakeup.com/research/controlCode.zip; the source code for the large test group program is located at http://www.jessicakeup.com/research/testCode.zip. Screenshots of the user interface may be found in Appendix C.

Task Setup

Each task consisted of an interactive GA trainer listening to a set of three musical excerpts (chromosomes) and ranking them by preference, so that the interactive GA instances could be run through many generations of improvements – each time breeding the preferred chromosomes selected by the trainers. Participants in the large group would listen to three eight-measure songs by clicking a play button for each, and participants in the small group would listen to 25 sets of three eight-measure songs.

The interaction options shown onscreen were to play or replay each of the songs and to select the best, middle, and worst. The time required go through a cycle of listening and ranking took approximately 60 seconds, rather than several minutes or even hours or days, so it was not a large amount of work for any one person. After all, that was the purpose of applying crowdsourcing to the human fitness function of a musical interactive GA – it is too time-consuming and tiring for a small group of people to rate a large set of songs (Biles 2007; Fu et al., 2009; Gartland-Jones & Copley, 2003; Khalifa et al., 2007; Oliwa, 2008; Tokui & Iba, 2000, Unehara and Onisawa, 2003).

The number of songs per task and the ranking method were chosen for the following reasons. Chen et al. (2009) discuss rating and ranking schemes for use in crowdsourcing for their experiment with gathering feedback on the quality of audio and video files. There are some similarities in that participants are choosing the best sounding choice, but for Chen et al. (2009) “best” means clearest audio recording with least loss and noise, and for this research “best” means most effective music. They point out problems with the commonly used MOS (mean opinion score) test. When given a rating scale, participants will interpret it differently or may not understand it. Their answers are ordinal, which indicates that averaging them does not accurately reflect the responses (Chen et al., 2009).

Unehara and Onisawa (2003) used such a rating scale, but it was more appropriate for their research than in this case, because the work was not shared among multiple users; it is less of a problem to have the users interpret the scales differently if they are working independently and their responses are not being averaged or added. The DarwinTunes project (DarwinTunes 2010a) used a five-point rating scale, even with collaborative work. Instead of this, Chen et al. (2009) recommend comparisons with prioritization instead for this type of research.

It also would have been possible to offer the option of “keep” or “good” and “don’t keep” or “bad” for each individual chromosome, as was done by Biles (1994) and Unemi (2002). Again, this works for independent users, but with collaboration, it is undesirable because of user interpretation. Some could choose to approve liberally and bad chromosomes could be bred, while some could choose to disapprove liberally and good chromosomes could be lost. Thus, for this research, prioritized comparisons will allow trainers to keep track of the songs in their minds while still narrowing down the “best” music to 1/3 of the original group size and avoiding effects of inconsistent rating styles from different users.

The length of the songs was chosen for three reasons. First, Legaspi et al. (2007) used 8 and 16 measure phrases in their study. Unemi (2002) had songs of 16 beats, Unehara and Onisawa (2003) used 4-measure songs that were combined into 16-measure songs, and DarwinTunes’ (2010b) songs are 4 measures long with 4 beats per measure. This indicates that eight measures is a reasonable and common size for compositional GAs. Secondly, that length corresponds to a complete musical thought in traditional music. Thirdly, at approximately 96 beats per minute with a 4/4 time signature, an 8 measure sample will last approximately 20 seconds.

Listening to all three songs in a task took approximately 60 seconds. Therefore, it was a relatively short and easy task suitable for crowdsourcing on mTurk. While participants could not retain 60 seconds of audio in short-term memory, they should have been able to remember a brief impression of it (e.g. “liked it”, “hated it”, “boring”, “beautiful”, etc.).

Furthermore, since it was a short listening task, it could be remunerated with a trivial payment on mTurk. Su et al. (2009) researched a crowdsourcing work community they called “System M”, where they found hourly rates between $0.78/hour minimum and $6.53/hour maximum. Jagadeesen et al. (2009) paid $1.00 for HITs that average 18 minutes 57 seconds, or about $3.00/hour. Eckert et al. (2010) paid the equivalent of $3.25 an hour. In their experiment to find Turker’s “reservation wages” (i.e. the lowest amount of money for which they are willing to continue working), Horton and Chilton (2010) found an average of $1.38/hour and that workers felt disproportionally motivated to work on HITs with modulo five payments, like $0.05 and $0.10.

Since participants in this study were paid $0.10 for HITs that last approximately 60 seconds, that is equivalent to earning $6.00/hour, which is a typical, or even above average, payment for mTurk workers, per Eckert et al. (2010), Horton and Chilton (2010), Jagadeesen et al. (2009) and Su et al.’s (2007) findings. Heer and Bostock reported that increases in pay led to faster, but not more accurate or thoughtful, responses (2010). The instructions that were shown to both groups of interactive GA trainers may be found in Appendix D.

Precautions

Responses where the choice was made in less time than was required to listen to all songs were be discarded, as the participant did not follow instructions and listen to everything. Both client-side (JavaScript) and server-side (Java) checks were put into place to ensure that they listened for as long as necessary, selected a rating for each sample, and did not select the same answer for different songs(e.g. did not mark two as "best").

Additionally, there are safeguards in place on mTurk to avoid getting maliciously incorrect or random data in return. Certain requirement thresholds, such as ratio of completed HITs and ratio of good completed HITs (those found acceptable by other requesters), can be set to filter out unscrupulous Turkers. For best results, Amazon recommends allowing only those with 95%+ HIT approval ratings to participate (Amazon.com, Inc., 2010b). Another safeguard is that work requesters do not have to pay Turkers if their work is unsatisfactory, and in turn, that a Turker's HIT approval rating will decrease.

mTurk makes it easy to limit assignments within a HIT to one-per-Turker, so researchers who wanted that restriction usually post multiple assignments in a single HIT, rather than multiple HITs (Alonso et al. 2008; Amazon.com, Inc., 2010b; Kittur et al,. 2008). As stated in the previous section, the Java SDK and asynchronicity of task postings required that the listening tasks be posted as separate HITs. Stolee and Elbaum (2010) reported that it is not possible to limit users to only a single HIT, but Little et al. (2009) were able to keep their own records of who had completed what in a database, and handle the HIT-limiting themselves. That functionality was built into this prototype, with the plan to have 5,000 unique Turkers reviewing and rating only one set of three songs each.

Recruitment of and Instructions for Trainers

There was a small (11 participant) locally recruited control group to demonstrate what is possible with a few people providing fitness function input. They were recruited by flyers at ETSU, with no requirement for musical experience or knowledge (Appendix E) and signed an informed consent document (Appendix F).

The size of the group was chosen based on the work of Unehara and Onisawa (2003) and Legaspi et al. (2007), which, of all the articles, books, and proceedings referenced in the Literature Review section of Chapter 2, were the only ones who gave a detailed description of the number of participants and number of songs involved in interactive GA training. The others made no mention of the numbers, because they used a general term like “some”, or only had one trainer. With 11 trainers listening to 25 sets of 3 songs each and a total of 825 songs, this experimental setup was equal to or more thorough than these conditions:

- Six trainers listening to 15 songs each, for 90 songs total (Unehara & Onisawa, 2003).
- Eleven trainers listening to 75 songs each, for 825 total songs (Legaspi et al., 2007)

In contrast to the control group, the test group was a large, crowdsourced group which also had no requirements regarding musical experience. The experimental crowdsourcing group users were found using mTurk, a “marketplace for work that requires human intelligence” where individuals are paid very small amounts of money to complete tasks; interestingly, it is also known as “artificial artificial intelligence” (Amazon.com, Inc., 2010a). mTurk has been shown to be a rich resource of user work, with a reasonable percentage of accurate, reliable, and dedicated workers (Kittur et al., 2008; Mason & Watts, 2009; Milne & Witten, 2008; Nowak & Rüger, 2010; Su et al., 2007).

Before the test group HITs were posted, six pilot participants were recruited from mTurk. This was only to verify that the HITs were set up correctly and the instructions were clear. After the test, their responses were discarded. Small technical problems and clarifications in the title of the HITs were corrected before beginning studies with the large crowdsourcing group.

The experimental group users were recruited on mTurk as the experiment progressed, with a 95% HIT approval rating as a qualification, as recommended by mTurk. Unfortunately, the other answer verification techniques suggested by crowdsourcing researchers are not applicable for these tasks. There is no prerequisite knowledge to test as a qualification and since there are no "right" or "wrong" answers, it does not make sense to compare answers from different trainers. It also means that using "gold standard" questions with known correct answers cannot be used to evaluate the quality of a trainer's answers (Eckert et al., 2010, Heer & Bostock, 2010).

Access to a pool of users was accomplished by placing HITs for Turkers at https://requester.mturk.com/mturk/resources. Their consent was gathered by way of a screen shown to them before continuing to the HIT. It had two buttons for them to choose between: "I do NOT agree to participate" and "I agree. Continue to the HIT" (Appendix C). The crowdsourced interactive GA training cost approximately $550 and it cost approximately $350 more to pay for the time of one small in-person interactive GA training group, one small group of reviewers, and one small group of composers.

The interactive GA training was not expected to take more than a few weeks at most, per these examples:

- Payment of $0.01 each for 2,500 tasks on mTurk which were completed in a “couple of days” (Alonso et al., 2008, p. 14)
- All 210 tasks were completed in 48 hours (Kittur et al., 2008)
- Payment of $0.00 to $0.10 per task for 611 participants to complete 36,425 tasks (Mason and Watts, 2009)
- Payment of $0.15 each for 80 tasks completed in about three hours and $4 each for fifteen 81-minute tasks in about 2 hours (Jagadeesan et al., 2009)
- Payment of $0.04 each for 1,000 tasks that were completed overnight (Hsueh et al., 2009)
- Payment of $0.05 each for 891 tasks that were completed in approximately 9 hours (Nowak & Rüger, 2010)

Genetic Algorithm Training

When the software was ready and control group users had been recruited, the training of the two interactive GA instances began. The two instances started with the same 75 random initial songs, which may be found at http://jessicakeup.com/research/originalRandomMelodies.zip. For the small group, 11 trainers listened to 25 sets of three songs each in March and April 2011. The participants completed their tasks in-person in a consistent environment - the mTurk Sandbox using Google Chrome on an Acer AS5742Z laptop with a 2GHz processor, 15.6" screen, and Windows 7. The tasks took approximately 45 – 60 minutes per in-person trainer.

For the large group in May 2011, this was 200 generations of 75 chromosomes, where each generation of 75 was heard by 25 trainers, with each person listening to three songs. In other words, there were 5,000 listening and ranking tasks and a total of 15,000 distinct songs listened to.

At the beginning of the experiment, there was a restriction in place that prevented any one Turker from completing more than one HIT. Only 125 HITs were completed in the first few days, and it became evident that it would take several months for the experiment to finish, if at all (if there were enough unique Turkers interested in participating). Since almost all of the mTurk researchers explicitly allowed individual workers to complete multiple HITs, with Little et al. (2009) being exceptions, the restriction was removed. The remaining 4,875 HITs took about another three-and-a-half days, with a mean time per HIT of 266 seconds. Heer and Bostock (2010) commented that it was difficult to predict how long HITs would take and what issues such as distractions and connection speed would cause delays. The statistics regarding number of HITS per Turker were as follows:

Number of unique Turkers: 154

Maximum # of HITs completed by a single Turker: 442

Minimum # of HITs completed by a single Turker: 2

Mean # of HITS completed per Turker: 31

Median # of HITS completed per Turker: 8

Mode # of HITs completed per Turker: 2

As the numbers indicate, there were a great deal fewer Turkers involved than if the HITs had been limited to one-per-person. The Median and Mode show that many of the Turkers completed two, or another small number, of HITs. However, there were a few outlying individuals who completed a large number of HITs and skewed the mean upward. There were, in fact, 13 Turkers who completed more than 100 of these listening tasks each and heard more than 300 short songs. When the interactive GAs were finished, the top five songs from each instance were exported as mp3 files for use in the evaluation phase.

Recruitment of and Instructions for Reviewers and Composers

To test the research hypothesis, the resulting music was rated and reviewed by another set of general-audience participants and by composers to determine whether crowdsourced compositional interactive GAs produced higher quality music than interactive GAs with a small group of reviewers as the fitness factor.

The top five musical creations produced by each interactive GA setup (small training group and crowdsourced training group, respectively) were presented to the reviewers in random order to prevent bias. They were only be told that they are evaluating ten computer music compositions and were not be informed of the experimental and control conditions or the methodology used to create the music. The instructions that were shown to the general audience reviewers may be found in Appendix G, while the instructions shown to composers may be found in Appendix H.

While the two earlier training groups - the small control group and large crowdsourced test group, who provided fitness to the interactive GAs – were only asked to rank three songs (or several sets of songs) in order of preference, the general audience reviewers and the composers were asked which they liked best and why.

The general audience reviewers were recruited locally with flyers (Appendix I) posted on the ETSU campus. They signed the informed consent document in Appendix J. They listened to the songs on an iPad2 were asked to rate their agreement/disagreement with the following statements: “I like this music”, “This music is artistically effective”, and “This music sounds similar to things I’ve heard before”; they were also given optional, open ended questions reading “What, if any, emotion(s) does it evoke?”, “What, if anything, was memorable about it?”, and “What, if anything, were its shortcomings?” It took approximately 20 – 30 minutes per reviewer to listen to all 10 songs and give ratings and comments.

The composers were recruited online via email (Appendix K) and most of their contact information was found on www.composersforum.org (Composers Forum, 2011). After email recruitment, they signed and returned the informed consent document in Appendix L by mail. They were then asked to rate their agreement/disagreement with these statements: “This music is interesting”, “This music is creative”, “This music is artistically effective”, and “This music is chorale-like.” They were given optional, open-ended questions reading “What, if anything, was memorable about it?” and “What, if anything, were its shortcomings?” With these statements and questions, helpful information was gathered about the ways the music was effective, or how the quality might be improved.

The research hypothesis would prove true if each set of reviewers rates the experimental crowdsourced interactive GA music more favorably than the control small group interactive GA music. The general audience reviewers provided more insight into popular appeal, while the composers provided more insight into artistic merit.

Results Format

The ratings from the two reviewing groups are reported in the next chapter, so as to show the effectiveness of the crowd-trained interactive GA compared to the small-group-trained interactive GA. The mean ratings of the two groups were used to determine whether the null hypothesis or alternate hypothesis must be rejected with p-value < 0.05, where:

H_0 = small-group-trained and crowdsourcing-trained compositional interactive GAs produce music that is equally effective

H_1 = crowdsourcing-trained compositional interactive GAs produce music that is more effective than those trained by a small group

Graphs showing the average Likert scale ratings are provided in Chapter 4, along with qualitative comments from reviewers and composers and links to the final songs that were created by both GA instances. The significance and contributions of the results and unanswered questions for future research may be found in Chapter 5.

Summary

When compositional interactive GAs have been used to create music, the results have been limited due to the extensive effort required for training. An experimental study was conducted to test the following hypothesis: When the training of a compositional interactive GA is crowdsourced, as opposed to being delegated to an individual or a small group, the fitness bottleneck is overcome and the resulting music is more effective.

MC, a compositional GA, was modified to use programmatic fitness rules guided toward a two-part chorale-like style and to collect human input as the most heavily weighted fitness function. Two instances of the modified version of MC were run: one with a very large crowdsourced group recruited from mTurk, and one with a small group of 11 recruited through traditional means. While the GAs were running, participants were given short songs in sets of three and asked to rank them best, middle, and worst.

After the music was created, another small group of eight reviewers was recruited, along with eight composers, to subjectively rate and give feedback on the music. This was done to determine whether the input from the small group or the crowdsourced group was more effective in training a compositional interactive GA.

Chapter 4

Results

Findings

In this section, links are provided to the five best songs created by each of the two interactive GA training conditions. Aggregate, summarized data is reported from the reviewers' and composers' ratings and comments. Statistics were analyzed using Microsoft Excel, Minitab, and The R Project for Statistical Computing. Complete data collected from the reviewers and composers may be found in Appendixes M, N, O, P, Q, and R.

Music

For the remainder of this report, the songs are numbered 1 - 5 for the small control group and 6 - 10 for the large crowdsourced group. The five short songs created by the small in-person control training group may be found at the following URLs:

- http://www.jessicakeup.com/research/1.mp3
- http://www.jessicakeup.com/research/2.mp3
- http://www.jessicakeup.com/research/3.mp3
- http://www.jessicakeup.com/research/4.mp3
- http://www.jessicakeup.com/research/5.mp3

The five songs created by the large, crowdsourced test training group may be found at the following URLs:

- http://www.jessicakeup.com/research/6.mp3

- http://www.jessicakeup.com/research/7.mp3
- http://www.jessicakeup.com/research/8.mp3
- http://www.jessicakeup.com/research/9.mp3
- http://www.jessicakeup.com/research/10.mp3

Reviewers' Feedback on Small Control Group Music

Figure 1 contains the combined non-musically trained reviewers' ratings of all five small-group-trained songs. It shows the agreement/disagreement percentages for the statements "I like this music", "This music is artistically effective", and "This music sounds similar to things I've heard before." It shows that all answers appeared for all statements, though Like was evenly distributed and Artistically Effective and Similar received more agreement and neutral answers.

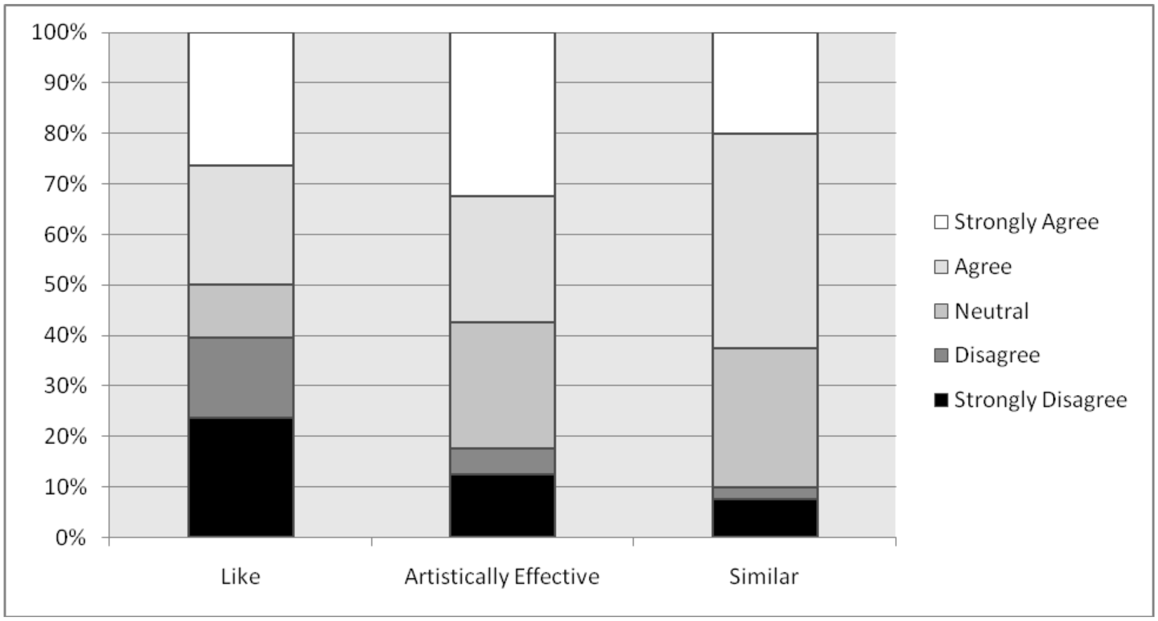


Figure 1. Combined reviewer ratings of control condition music.

The comments from the reviewers regarding the five songs they heard from the control group repeatedly mentioned these words and ideas: curiosity, suspense, dissonance, ballet, storytelling, syncopation, mystery, anxiety, awkward rhythms, and too much distance between the bass and soprano. The non-musically trained reviewers' per-song ratings and comments regarding the control, small-group-trained interactive GA may be found in Appendix M.

Reviewers' Feedback on Large Test Group Music

Figure 2 contains the combined non-musically trained reviewers' ratings of all five crowdsource-trained songs. It shows the agreement/disagreement percentages for the statements “I like this music”, “This music is artistically effective”, and “This music sounds similar to things I've heard before.” Again, all statements received at least one of each answer. The reviewers seemed less polarized by the test condition music, as there were more Agrees and Neutrals and fewer Strongly Agree and Strongly Disagree responses as compared to the control condition music.

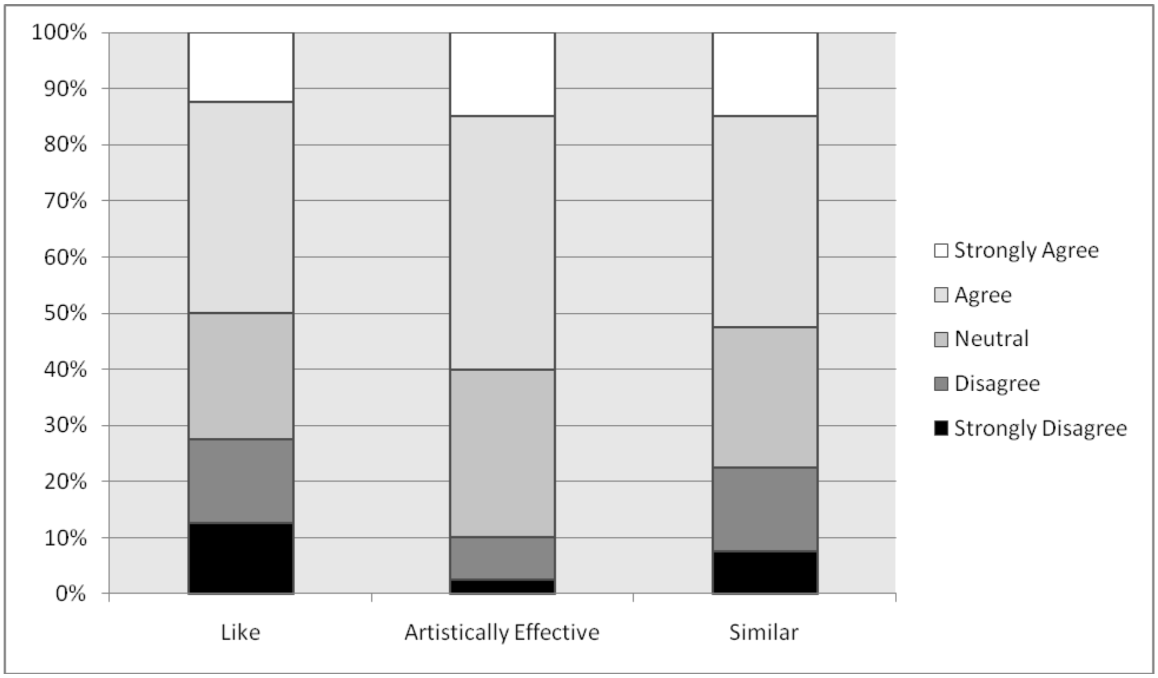


Figure 2. Combined reviewer ratings of test condition music.

The comments from the reviewers regarding the five songs they heard from the test group repeatedly mentioned these words and ideas: darkness, lack of flow, mystery, curiosity, happiness, ballads, major 3rds, and the need for tempo variance. It was very similar to the reviewers' descriptions of the small group music, though a few songs in each group had unique comments. The non-musically trained reviewers' per-song ratings and comments regarding the test crowdsource-trained interactive GA may be found in Appendix N.

Composers' Feedback on Small Control Group Music

Figure 3 contains the averages of the composers' ratings of all five small-group-trained songs. It shows the agreement/disagreement percentages for the statements "This music is interesting", "This music is creative", "This music is artistically effective", and

"This music is chorale-like." It shows that all statement received all answers, except for Chorale-like, which mostly received Strongly Disagree answers.

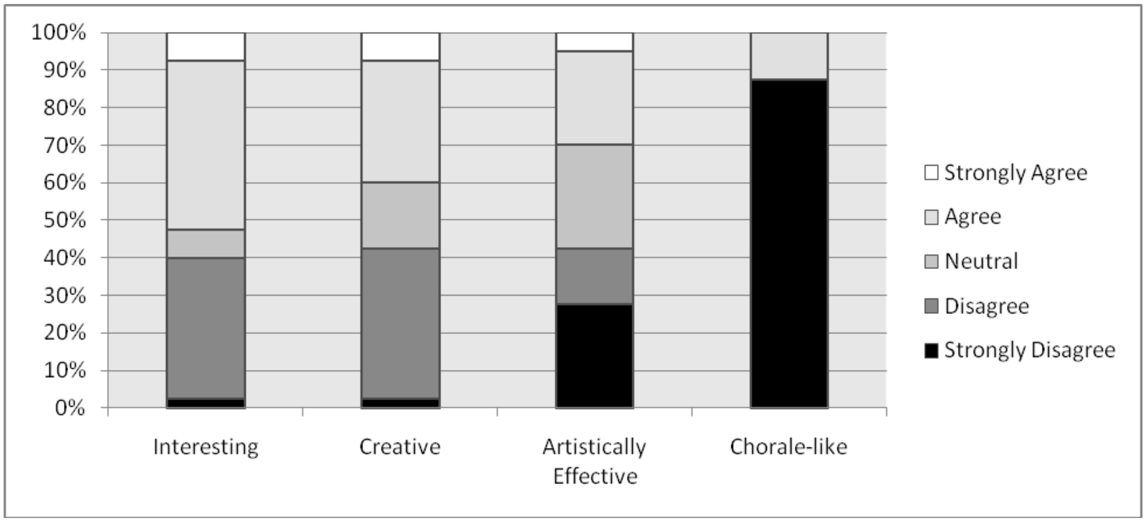


Figure 3. Combined composer ratings of control condition music.

The comments from the composers regarding the five songs they heard from the control group repeatedly mentioned these words and ideas: randomness, dissonance, lack of coherence, lack of shape, and atonality. The composers' per-song ratings and comments regarding the control small-group-trained interactive GA may be found in Appendix O.

Composers' Feedback on Large Test Group Music

Figure 4 contains the averages of the composers' ratings of all five crowdsource-trained songs. It shows the agreement/disagreement percentages for the statements "This music is interesting", "This music is creative", "This music is artistically effective", and "This music is chorale-like." It is nearly identical to the graph of composers' feedback on the small control group music in Figure 3, as answers were mixed on all questions except for Chorale-like, where the answers were in strong disagreement.

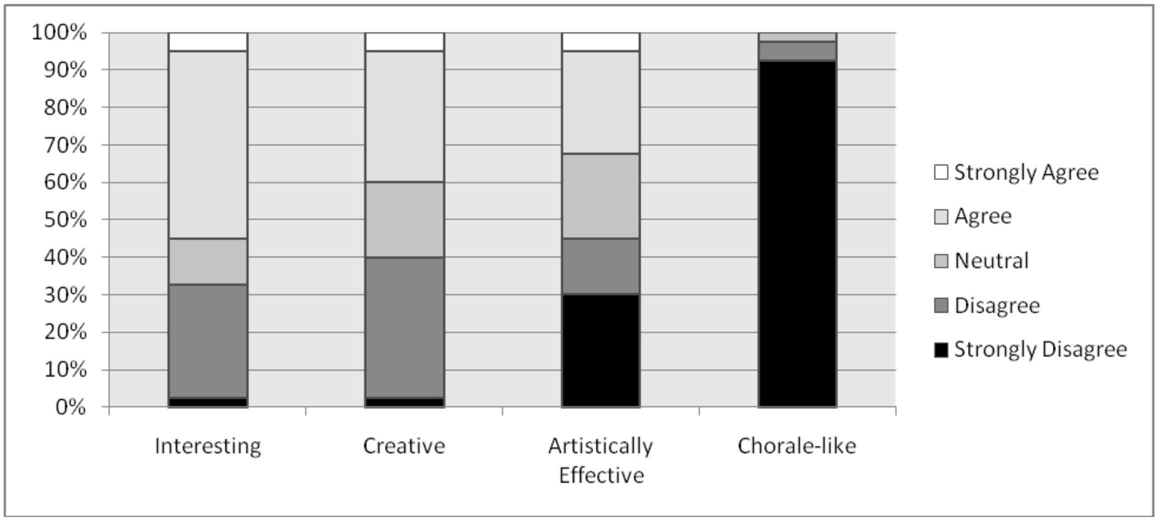


Figure 4. Combined composer ratings of test condition music.

The comments from the composers regarding the five songs they heard from the test group repeatedly mentioned these words and ideas: Atonal, contrapuntal, too sustained, dissonance, randomness, and lack of shape. It was very similar to the composers' descriptions of the small group music, though a few songs in each group had unique comments. The composers' per-song ratings and comments regarding the test crowdsource-trained interactive GA may be found in Appendix P.

Small Control vs. Crowdsourced Test

In this section, the differences between answers to individual questions and groups of questions are compared between the control and test group songs. First is the difference in effectiveness. At this point and later in the chapter, effectiveness is used to describe the grouping of Like, Artistically Effective, Creative, and Interesting, where applicable (e.g. composers did not rate Like).

The null hypothesis that "small-group-trained and crowdsourcing-trained compositional interactive GAs produce music that is equally effective" is not rejected at the 95% confidence level because the p-value calculated by a *t* test (shown in Table 1) is 0.463.

Table 1

t Test of Test Minus Control Differences in Effectiveness from All, Reviewers, and Composers

	N	Mean	St. Dev.	Min	Q1	Median	Q3	Max
Combined	16	.015	6.19	-8.00	-3.75	-0.67	0.75	19.00
Reviewers	8	.013	3.77	-8.00	-4.00	-2.50	1.75	19.00
Composers	8	.017	8.24	-4.67	-1.17	-0.33	0.00	8.67

The difference in effectiveness, shown below, was calculated by subtracting each reviewer and composers control group ratings for quality from their test group ratings for effectiveness. It shows a few outliers, but for most of the subjects, the test music and control music were very similar in effectiveness.There was more dispersion and variance in opinion regarding the difference in effectiveness between test and control music between reviewers as shown by the Figure 5 and the difference in standard deviation.

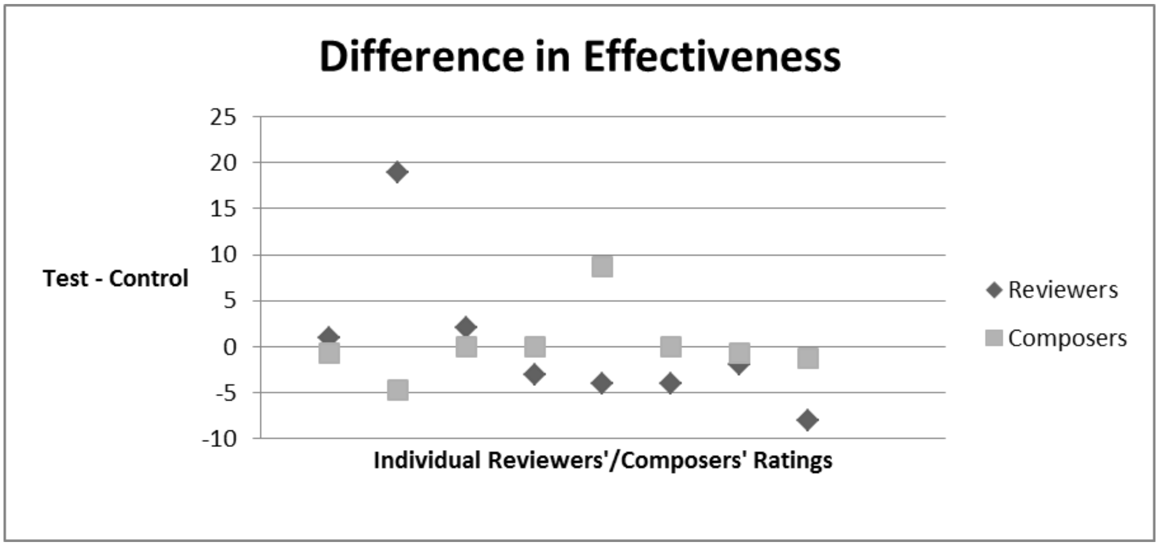


Figure 5. Difference between reviewers'/composers' ratings of test vs. control effectiveness.

In a randomization test run to compare the differences between Reviewers' and Composers' combined ratings of effectiveness in the test and control groups, the observed difference in means is very similar (near 0) between the groups, as seen in Figure 6. This indicates that the differences between effectiveness ratings of the test group vs. control group may have happened by chance and are not significant.

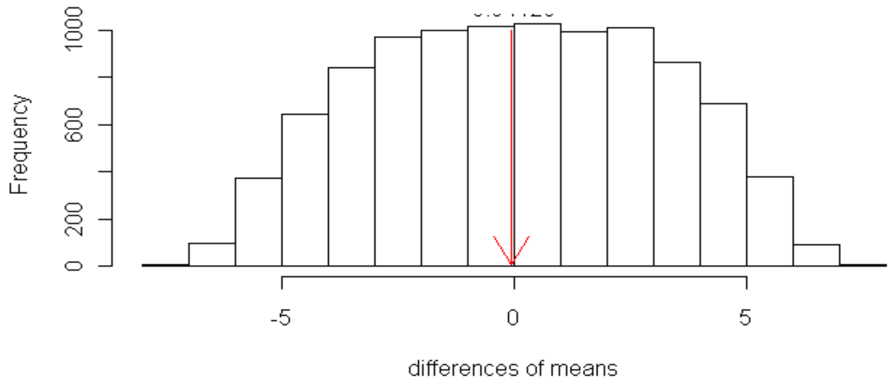


Figure 6. Randomization test for ratings of test vs. control effectiveness.

For originality, the ratings "This music sounds similar to things I've heard before" (reviewers only) were reversed. In Figure 7, test music originality ratings of individual reviewers were subtracted from their control music originality ratings. There was one outlying reviewer who rated the test music considerably higher than the control music, and for the rest, the two groups were quite close in originality.

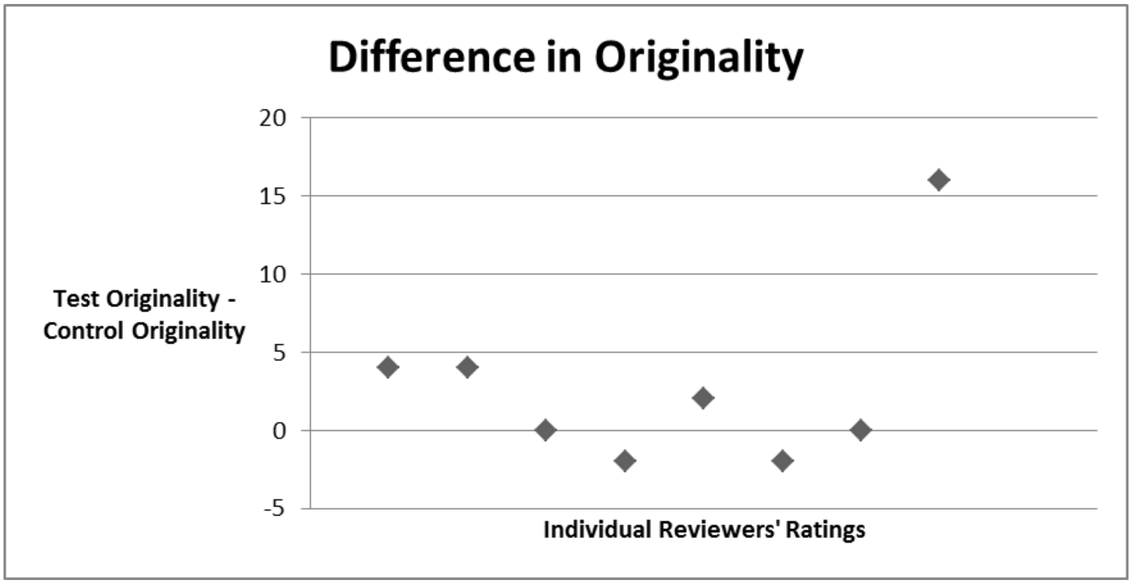


Figure 7. Difference between reviewers ratings of test vs. control originality.

Composers agreed or disagreed with the statement “This music is chorale-like”, and Figure 8 shows their chorale-like ratings for the control group subtracted from their chorale-like ratings for the test group. One outlying composer felt strongly that the control music was more chorale-like than the test music, and for the rest, all the music was similar in its chorale-likeness.

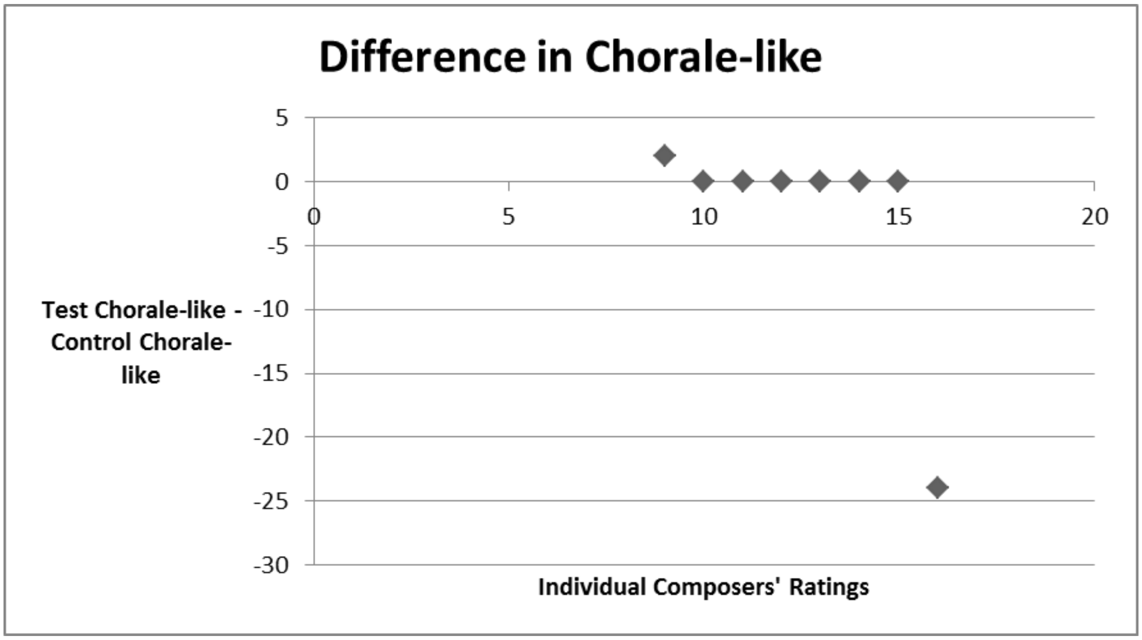


Figure 8. Difference between composers’ ratings of test vs. control chorale-likeness.

Reviewers’ Question by Question Comparision

In the following section, ratings from individual statements posed to reviewers are evaluated – first “I like this music”, then “This music is artistically effective,” and then “This music sounds similar to things I’ve heard before”. Figure 9 shows the combined Like ratings from all reviewers for each song in both conditions. With a maximum score of 50, it indicates that the songs were fairly well-liked, with the test songs being slightly better liked than the control songs.

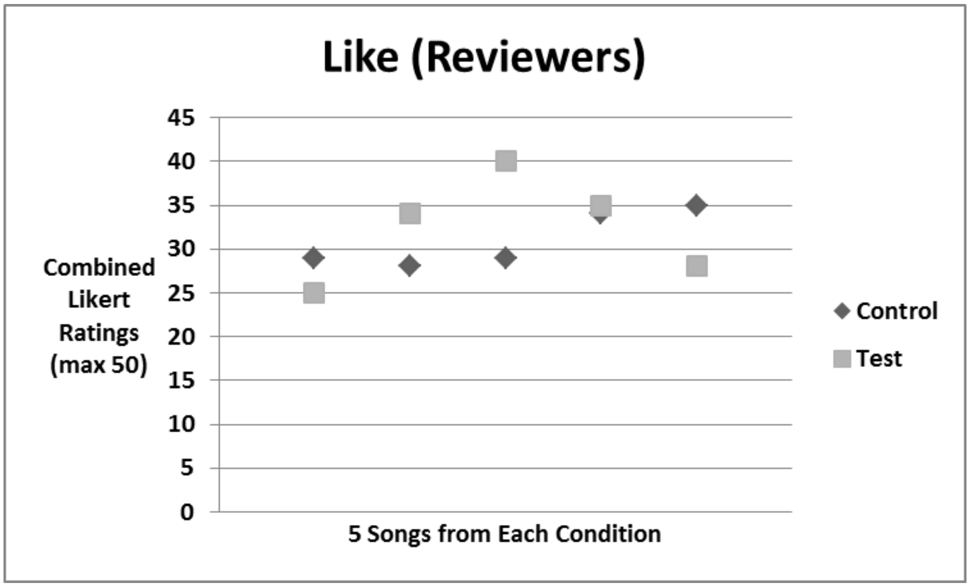


Figure 9. Combined reviewer ratings for “I like this music”.

Figure 10 shows the combined artistic effectiveness ratings from all reviewers for each song in both conditions. With a maximum score of 50, it indicates that the songs were seen as fairly artistically effective, with the test songs rated slightly more artistically effective than the control songs.

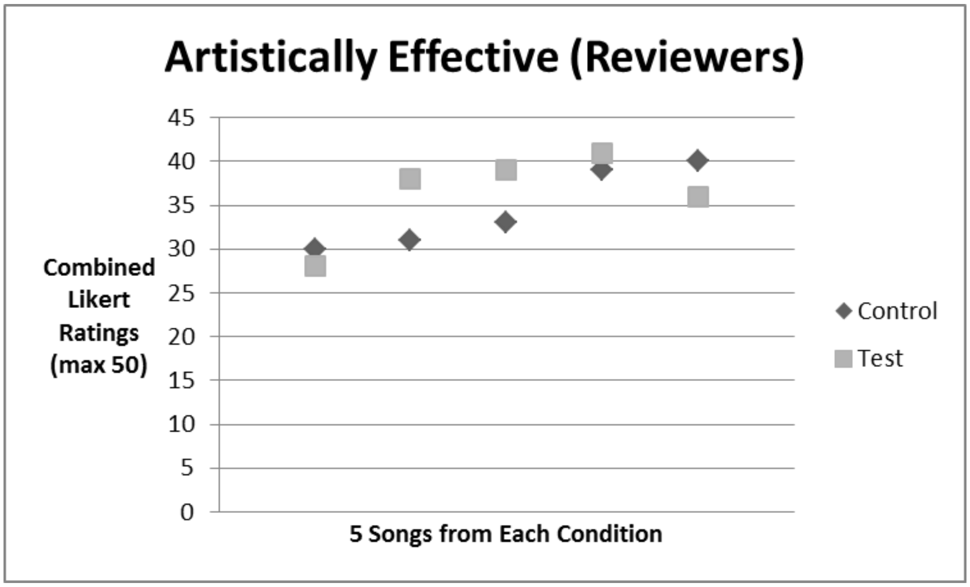


Figure 10. Combined reviewer ratings for “This music is artistically effective”.

Figure 11 shows the combined Originality (reverse of Similar) ratings from all reviewers for each song in both conditions. It indicates that the control and test songs had about the same average originality, though there was more spread in the control songs. With the responses centered around 25 – 30 out of 50, they believed the songs to be neither very original or un-original.

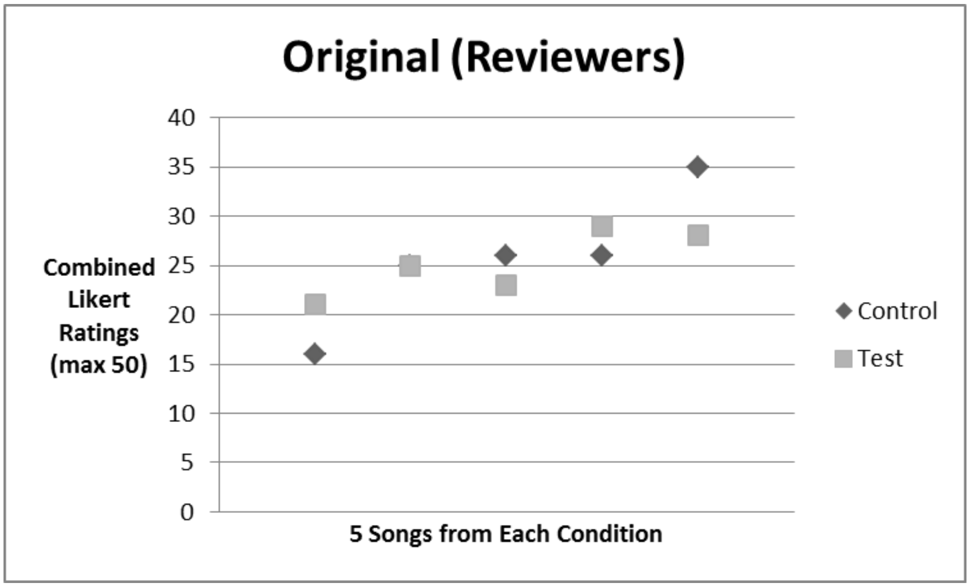


Figure 11. Combined reviewer ratings for “This music sounds similar to things I’ve heard before”.

Composers’ Question by Question Comparision

In the following section, ratings from individual statements posed to composers are evaluated – “This music is interesting”, “This music is creative”, “This music is artistically effective,” and “This music is chorale-like”. Figure 12 shows the combined Interesting ratings from all composers for each song in both conditions. The most interesting song was from the control group and the least interesting from the test group, but otherwise the test songs were rated as more interesting than the control songs. With ratings between 25 – 40 out of a possible 50, the composers believed the songs to be more interesting than not.

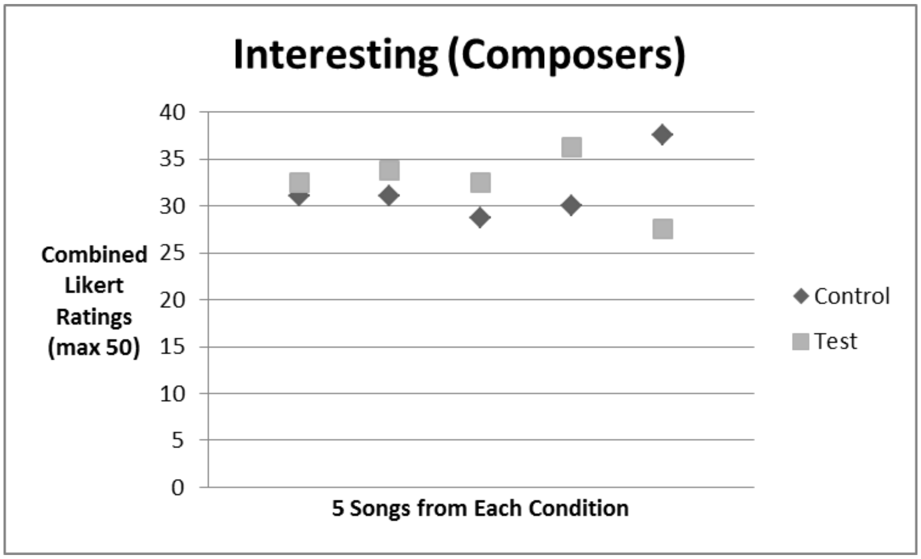


Figure 12. Combined composer ratings for “This music is interesting”.

Figure 13 shows the combined Creative ratings from all composers for each song in both conditions. It indicates that the songs were believed to be moderately creative (between 25 – 40 where the maximum is 50) and that the control and test condition creativity levels were almost identical.

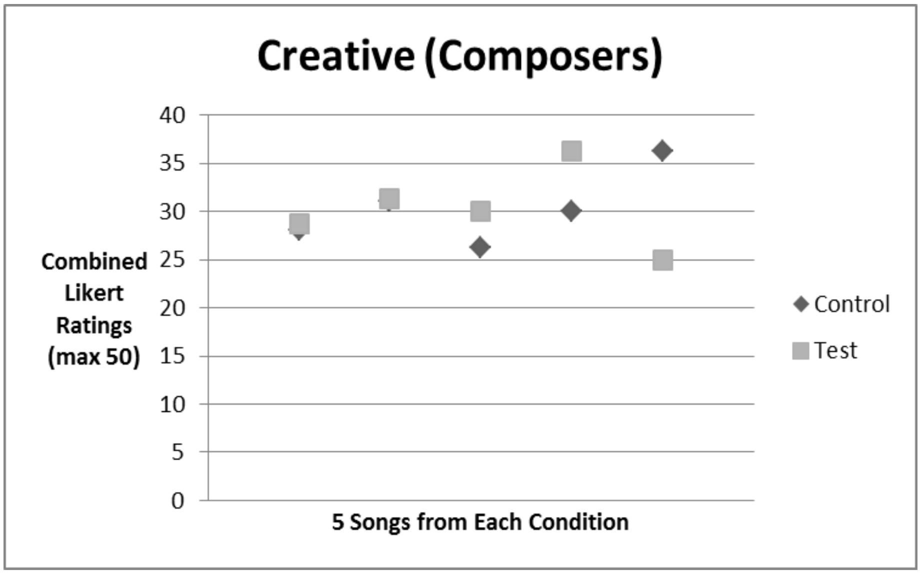


Figure 13. Combined composers ratings for “This music is creative”.

Figure 14 shows the combined Artistically Effective ratings from all composers for each song in both conditions.The scatterplot suggests that the songs were less artistically effective than they were interesting and creative, from the composers’ perspectives, and that both conditions’ songs had about the same level of artistic effectiveness.

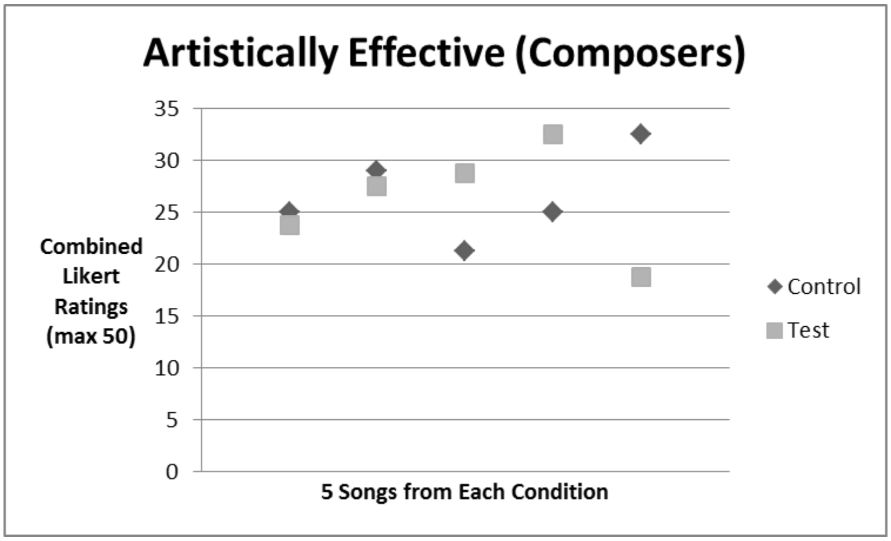


Figure 14. Combined composer ratings for “This music is artistically effective”.

Figure 15 shows the combined Chorale-like ratings from all composers for each song in both conditions. This question was clearly the most disagreed with. In fact, with the minimum score being 8 (1 point from each composer), the ratings could not have been much lower. The control songs were all rated more chorale-like than the test songs.

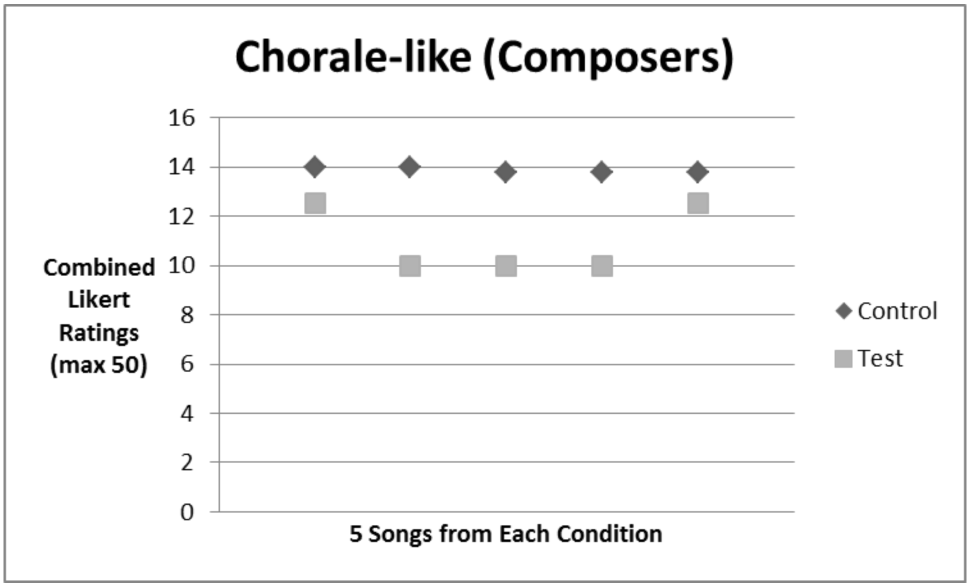


Figure 15. Combined composer ratings for “This music is chorale-like”.

Reviewers vs. Composers

In this section, trends in ratings by reviewers versus trends in ratings by composers are evaluated. When the control group ratings were combined and the test group ratings were combined, as shown in Figure 16, the reviewers rated both groups highest on the "This music is artistically effective" statement and highest on "This music sounds similar to things I've heard before". This means that the musical originality was *low* compared to the other factors.

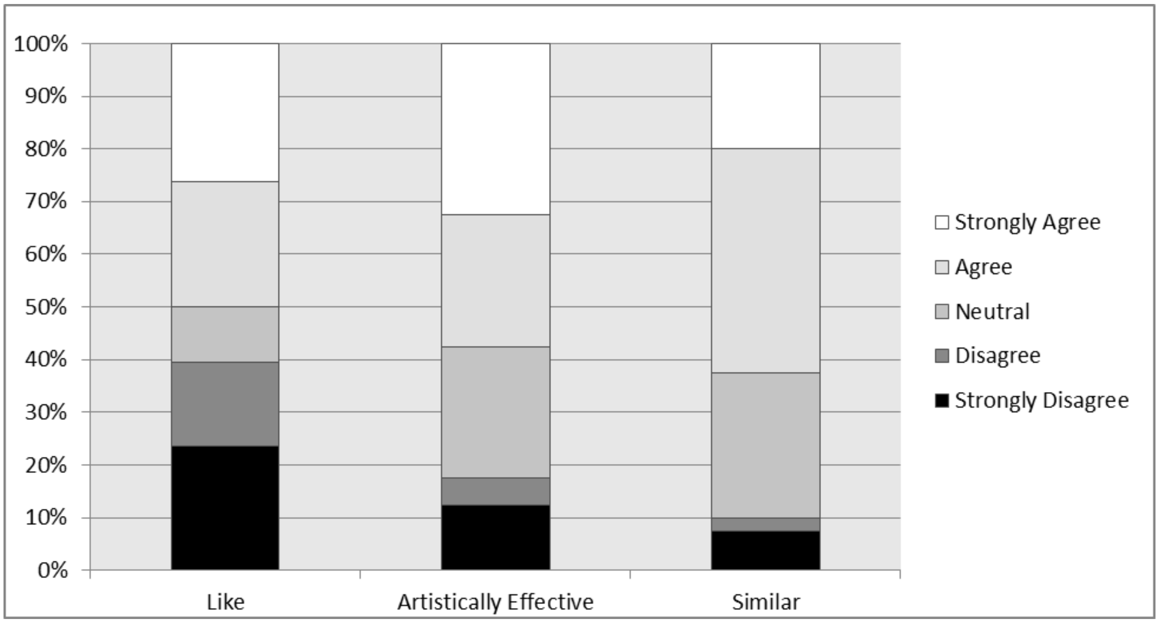


Figure 16. Combined reviewer ratings of all music.

When the control group ratings were combined and the test group ratings were combined, as shown in Figure 17, the composers rated both groups highest on the “This music is interesting" statement and lowest on the “This music is chorale-like” statement.

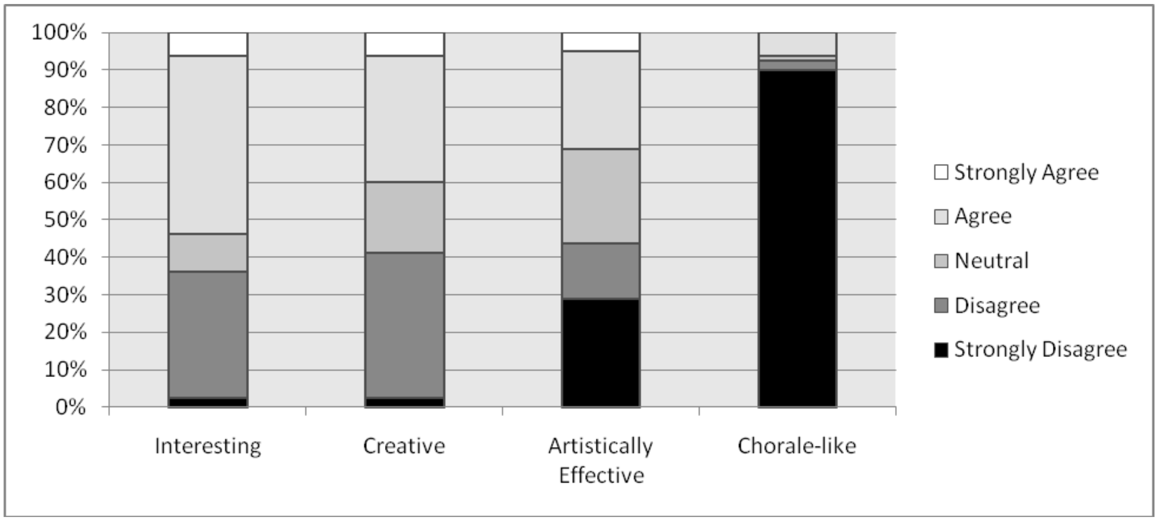


Figure 17. Combined composer ratings of all music.

Both reviewers and composers rated their agreement with the statement “This music is artistically effective”. Figure 18 shows each song’s ratings by reviewers and by composers. There was a wide variance in answers and there were higher ratings from reviewers, as a group, than composers. Perhaps this indicates that artistic effectiveness is a very subjective measurement and that composers have higher expectations of musical artistry.

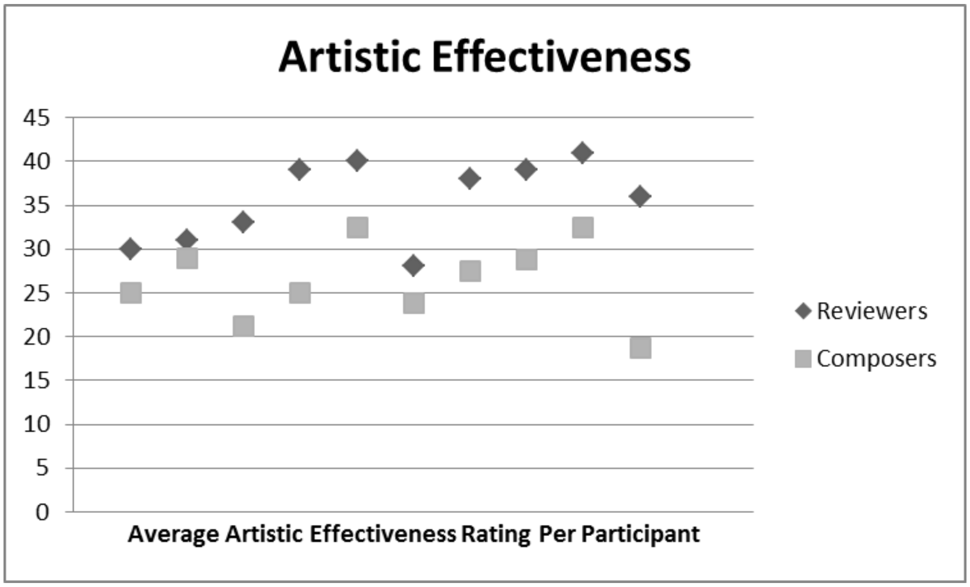


Figure 18. Reviewers’/Composers’ Artistic Effectiveness Ratings

To verify the appearance of higher ratings from reviewers, a paired *t* test (Table 2) was run. It showed a p-value of 0.000.

Table 2

Paired t Test of Per-Song Reviewer vs. Composer Ratings

	N	Mean	St.Dev.	SE Mean
Reviewers	10	35.50	4.65	1.47
Composers	10	26.60	4.67	1.48
Difference	10	8.90	4.65	1.47

To further verify that reviewers gave higher ratings than composers, a randomization test for paired data was also run. It showed an approximated p-value of 0.0011. This, along with the paired *t* test p-value of 0.000, proves that there *is* a statistically significant pattern of reviewers being more generous with their scoring than composers.

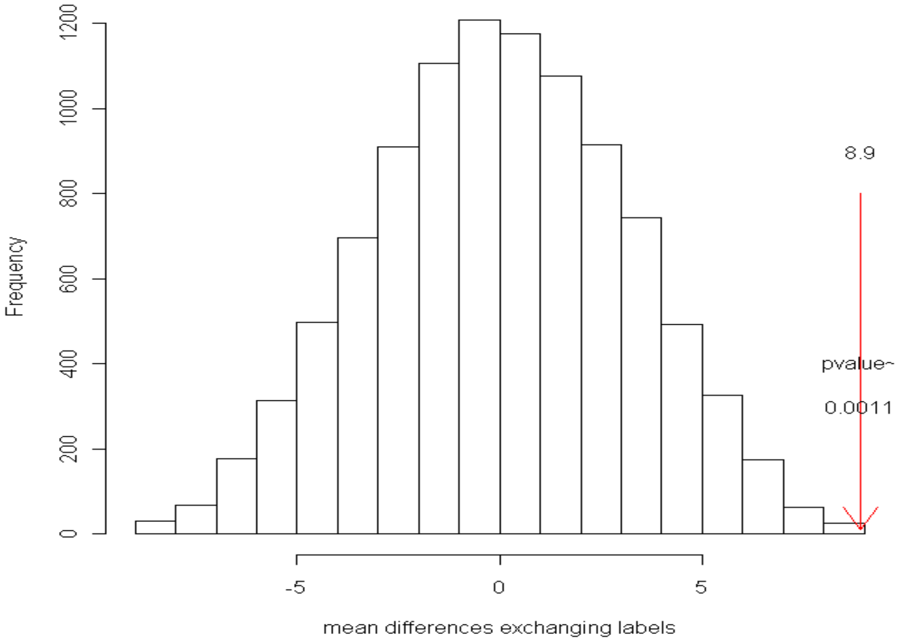


Figure 19. Randomization Test for All Reviewer vs. Composer Ratings

Reviewers and composers did not rate their agreement with the same statements (with the exception of "This music is artistically effective"); thus, their responses to the other individual statements could not be compared. The overall comments from reviewers and composers regarding the study and all the music they heard may be found in Appendices Q and R, respectively.

Song by Song Comparison

According to the effectiveness sub-group of questions (Like, Artistically Effective, Creative, and Interesting), the best song was a tie between song 5 (http://www.jessicakeup.com/research/5.mp3) from the control group and song 4 from the test group (http://www.jessicakeup.com/research/4.mp3).

According to the combined reviewer ratings of all statements, the best song is song 5 (http://www.jessicakeup.com/research/5.mp3) from the control group and the worst song is song 1(http://www.jessicakeup.com/research/1.mp3) from the test group. According to the combined composer ratings of all statements, the best song is song 5 (http://www.jessicakeup.com/research/5.mp3) from the control group and the worst song is song 4 (http://www.jessicakeup.com/research/4.mp3) from the test group.

Summary of Results

The reviewers' opinions of the test and control groups and composers' opinions of the test and control groups were given first. Most of their responses were moderate, except that the composers did not agree that any of the music was very chorale-like. In comparing the ratings for the test group and control group, there were not statistically significant differences in which was more effective. There was a significant difference in ratings given by reviewers and composers, in that reviewers tended to give higher ratings (to both test and control songs) than the composers did.

Chapter 5

Conclusions

Implications

First, it is important to note that due to small number of participants (*N = 16*) reviewing the music created by the interactive GAs, the differences in ratings were not statistically significant, with one exception. There was a definite difference between scores given (to both groups) by reviewers and scores given (to both groups) by composers. Further research with a larger number of reviewers and/or composers could answer the research question more definitively, and with statistical significance.

The null hypothesis - small-group-trained and crowdsourcing-trained compositional interactive GAs produce music that is equally effective – cannot be rejected at this time. The rest of the results suggest that the small control group-trained music and the large crowdsource-trained music had small differences that may be attributed to chance. In some cases like Like and Interesting the test music performed slightly better, and in others like Chorale-like and Original, the control music performed slightly better.

Recommendations

Genetic Algorithm Setup

If this particular modification of MC is to be used again, it should be further tested and adjusted, particularly regarding the mutation rate. The songs seem to have converged too quickly for the purposes of this experiment and reached local optimas. In both the small group with 11 generations and the large group with 200 generations, the five best songs in the last generation of each sounded very similar to each other, even starting with the same motives and only differing in the middles and endings. In this research, genes had a 1 in 12 chance of being selected for mutation, and if selected, the note, octave, and duration each had a 1 in 2 chance of being randomly changed. Darwin Tunes' mutation occurred with a chance of 1/1500 per node within a chromosome, but they did not report problems with convergence (DarwinTunes, 2010c).

One could argue that the reason the music created by the test condition was more effective than the control is that there were 189 more generations, and therefore, the music was more evolved and better, regardless of the human training input. It would be interesting to compare a 200-generation GA run with the same programmatic fitness functions and *no* human input. For future research, a crowdsourced interactive GA with 200 (or some other large number) generations could be compared to a non-interactive GA with 200 generations that was trained only by programmatic fitness functions, to compare the effectiveness of the music each creates.

While the composers' ratings varied with regard to interest, creativity, and artistic effectiveness, their responses showed an overwhelming belief that the music (created by both conditions) was not chorale-like. There are some changes to be considered in future research that may create a more chorale-like style. First, the programmatic fitness function rules need to be further refined or supplemented by additional constraints. Some of the rules included in this work were specific to chorales, but many others were merely general guidelines that would apply to most tonal music. The other way to make the music more chorale-like would be to increase the relative weights of the rules. The human input from mTurk was more heavily weighted than all the other rules combined. For example, a rating of “middle” from a set of three added 100,000 to the score and a rating of “worse” added 1,000,000 to the score. In contrast, a song with a too-high percentage of rests to notes was penalized at double the difference between the actual and max percentages (e.g. a song with 30% rests instead of 10% or less would have had 40 points added to its score). Every eighth note that occurred on the first beat of a measure added 20 points to the song’s score. A lower score was more desirable. The disparity in scoring between human ratings and everything else was because the GAs were supposed to be primarily interactive. The programmatic rules were only in place to be used as a tie breaker, but the emphasis on the human ratings may have been too much.

Regarding the choice of a chorale-like style, it was a limitation that the music contained only soprano and bass parts, and it therefore lacked the typical four-part texture that includes an alto and tenor voice. Another rule, namely that the chorales only used notes from the C major scale, is also problematic, since chorales typically tonicize chords and modulate to closely related keys. To create such harmonic shifts would call for the inclusion of chromatic pitches and many additional constraints.

Had the author been aware of the Turkit toolkit, it would have been used in development (Little et al., 2009). It manages automatic, iterative postings of HITs to mTurk for HITs where the contents of one batch are dependent on the previous. It was created by Little et al., (2009) for tests with iterative tasks related to “image description, copy editing, handwriting recognition, and sorting” and would likely have been able to handle many of the error checking and reposting functions that were written for this research.

Training Process

Before the experiment was conducted, it was expected that the HITs would take a little more than 60 seconds to complete, for an effective rate of more than $6.00/hour. That was considerably higher than the HITs other researchers have found or posted themselves. However, they were judging hour rates based on the time it actually took to complete the HIT. Since that proved to be almost 4 ½ minutes, the workers were only paid a rate equivalent to $1.35 an hour. It is suspected that they were not listening to the songs over-and-over, but rather waiting on the page to load, multitasking, etc. More detailed recordkeeping of their click activity might be able to answer that question.

The rating method of choosing the best, middle, and worst song out of three was chosen for several reasons outlined in Chapter 3. Using a scale, such as 1 – 5 or 1 – 10, can be inconsistent because participants have different interpretations of what each score means and an individuals' strictness or generosity with scores may vary over time. During the experiment, though, one trainer offered his opinion about why he wished there were other rating options available. He expressed concern about getting sets of three unusually good or unusually bad songs. He believed that the best in some groups were worse than the worst in others, and had no way to accurately rate them as such. Preference judgments could be a way around that problem. With preference judgments two options are offered at a time and the participant chooses the better one. As various pairs are evaluated, the entire list of songs could be sorted by quality (Urbano et al., 2010).

The one-HIT-per-worker limit that was removed near the beginning of the large test group training may have had an effect on the results. While it still maintained adherence to the premise of work undertaken by the collective online intelligence, it would have been more in the spirit of crowdsourcing to have input from more unique participants (Surowiecki, 2005). The workers who completed a large number of HITs had a bigger influence on the music that was eventually created than the workers who completed a few HITs. Since the answers to their questions were purely subjective, most of the measures recommended by Amazon and other researchers to prevent cheating - such as known correct answers and qualification tests - were not applicable. Therefore, there was a risk that one of the Turkers who completed a large number of HITs could have done so carelessly or maliciously and negatively affected the outcome.

Averaged ratings from multiple Turkers might have been helpful in ensuring quality ratings. As Xu and Bailey (2011, p. 1185) reasoned in their crowdsourcing of visual design critiques, "[s]ince workers may have different styles and diverse aesthetic experiences, we adopt a voting scheme to prevent the introduction of errors." DarwinTunes, too, used a voting scheme (DarwinTunes, 2010d). Even voting is not always a good solution; in situations where cheating workers are abundant, the honest and correct answers may be outvoted (Lease, et al., 2011).

If a voting scheme is not implemented, it could be tried again with the one-per-person restriction in place, perhaps with higher pay to encourage participation. However, that is probably a less effective solution because it conflicts with the typical Turker workflow. Due to the learning curve of a HITs and other factors like qualification tests, many Turkers do streaks of work where they find worthwhile types of HITs and complete them as long as possible (Heer & Bostock, 2010).

Review Process

As a side effect of the premature convergence of the interactive GAs to similar-sounding music, one of the survey statements became less relevant. Some participants expressed concern about “This music sounds similar to things I’ve heard before.” They were unsure if they were supposed to compare it with other music from this study, where the answer was likely “agree” or “strongly agree” or with all the music they had heard prior to the study, where the answer was more likely negative. Thus, the reviewers’ answers to that question were not very meaningful. The composers were not asked that question, though it is important to note that three of them thought it impossible to differentiate between all of the control and test songs, and gave identical ratings to all ten.

Though the ratings for the crowdsourced music were slightly higher than that of the small-group-trained music, neither set of songs had particularly high ratings. Reviewers and composers expressed some valid concerns about the amount of dissonance and the feeling of atonality that made it seem random. Perhaps, if a similar experiment is repeated, the melody should be left to evolve on its own programmatically for a number of generations before participants are brought in to train it, as Unehara and Onisawa (2003) suggested. Alternatively, using a minimum programmatic fitness function (in which clearly subpar songs are discarded without sending them to trainers) could allow the songs to become more musically refined and effective with the same amount of reviewer effort (de Freitas & Guimarães, 2011).

The reviewers seemed less polarized by the test condition music, as there were more Agree and Neutral responses (and fewer Strongly Agree and Strong Disagree responses) for Like and Artistically Effective, as compared to the control condition music. This supports the idea discussed in Chapter 1 that music created by a crowdsourced interactive GA might have more popular appeal overall, but might not be loved or hated since it was created by consensus of so many people.

The qualitative comments from both composers and reviewers indicated that the music would have been more effective with more variety. They specifically mentioned variations in tempo and instrumentation, but other factors such as variations in key signatures, time signatures, modalities, and number of parts could be implemented as well.

Summary

Computers have been used for tools in sound synthesis, sound processing, music theory analysis, composition, and performance. AI has been applied to creation and composition. When GAs used to create music, the fitness function may be programmatic based on music theory regarding voice leading and harmonic progressions. Computer-created music created by GAs with programmatic fitness functions tend to be homogeneous and non-interesting (Biles, 2007; Roads, 1985).

In an Interactive GA (Unehara & Onisawa, 2003), human listeners gauge the quality of a composition; Interactive GAs are subject to a fitness bottleneck because it takes a lot of time and energy for humans to rate and review generations of chromosomes, thus limiting the effectiveness of the music created by the Interactive GA.

Existing methods of music creation with Interactive GA can produce higher quality output than compositions without human input; however, they are by a fitness bottleneck. (Biles 2007; Chen, 2007; Fu et al., 2009; Gartland-Jones & Copley, 2003;

Khalifa et al., 2007; Oliwa, 2008; Unehara & Onisawa, 2003). Humans must attentively listen and precisely rate a substantial amount of audio information to train a musical GA; they may take too long or be overwhelmed in doing so and this fitness bottleneck is repeatedly referred to as a limiting factor in musical GA research (Biles, 2007; Tokui & Iba, 2000). By training interactive GAs for music composition with online collective intelligence, or “crowdsourcing”, it was possible to supply the algorithm with adequate training data without requiring much input from any one evaluator.

This research was intended to show whether applying crowdsourcing to the human review fitness function of musical GAs yields more effective music, as compared to the small groups typically chosen to provide feedback to a compositional interactive GA. Those small groups are limited by a fitness bottleneck, because it takes a great deal of time and effort to fully train the GA.

Crowdsourcing is the outsourcing of work to the collective online intelligence. A wide variety of work types have been crowdsourced, and it is most effective for use in tasks that are easy for humans but difficult for computers, such as image tagging and relevance rating. It is subject to misuse by careless or malicious users, but there are techniques to check responses for validity and mitigate that risk. Amazon mTurk is a crowdsourcing community where requestors can post tasks and Turkers will do the work in exchange for small payments.

Crowdsourcing has been applied to the GA fitness bottleneck in other domains and it has been applied to music recommendation systems. In DarwinTunes, it was tried with a compositional interactive GA to see what type of music could be created. An experimental study was conducted to test the following hypothesis: When the training of a compositional interactive GA is crowdsourced, as opposed to being delegated to an individual or a small group, the fitness bottleneck is overcome and the resulting music is more effective.

This was accomplished by establishing an interactive GA that created music in a two-part chorale-like style. A small group of group of eleven participants recruited through traditional means trained one instance of the musical interactive GA, as a control group. Those results were compared to another instance of the same musical interactive GA trained by a large crowdsourced group recruited from mTurk.

After the music was created, another small group of eight reviewers was recruited, along with eight composers, to subjectively rate and give feedback on the music. This was done to determine which training method for interactive compositional GAs was more effective. The data was analyzed to gather the following information.

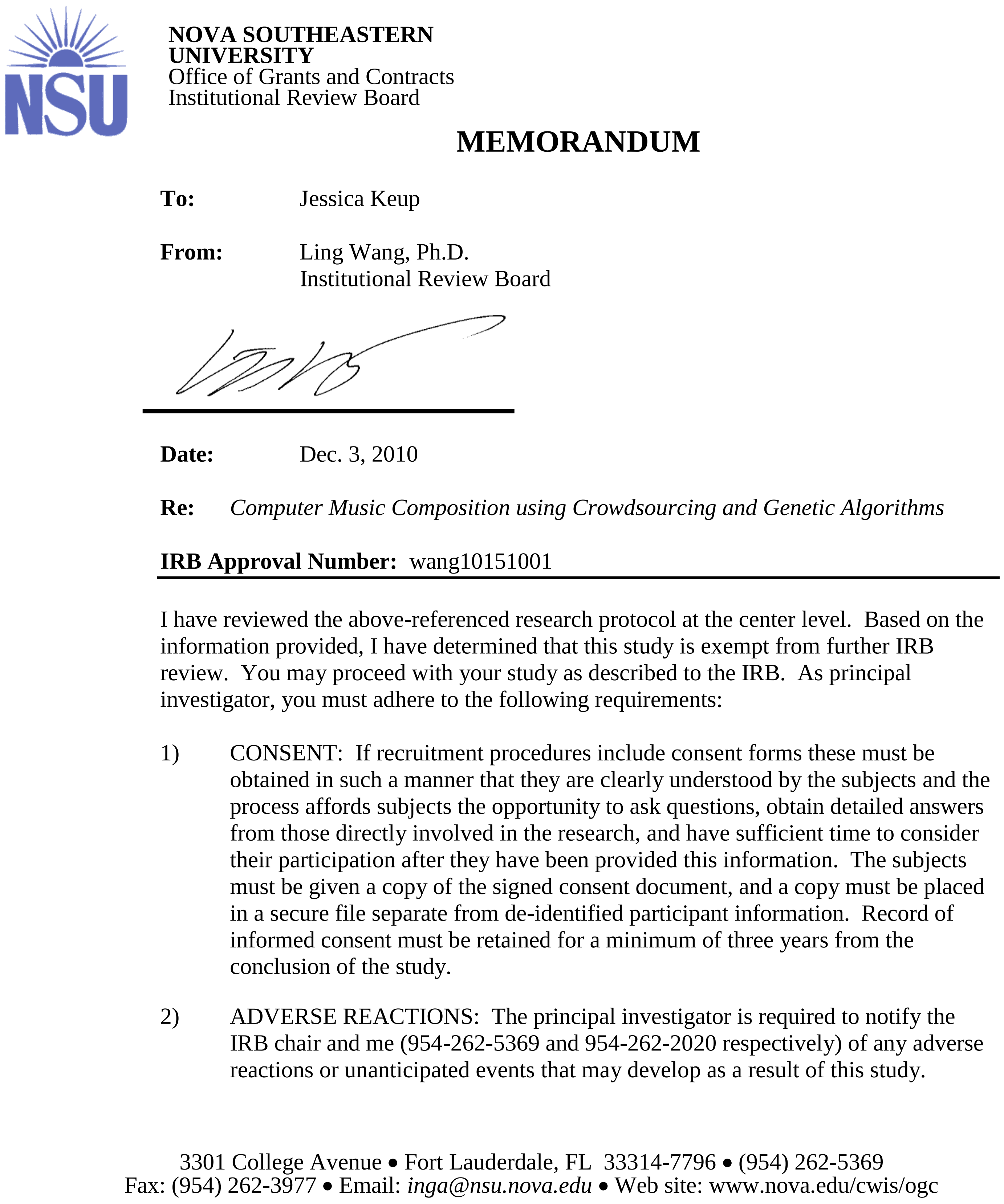
The songs from the large, crowdsourced test group scored slightly higher (with all questions combined) than the songs from the small-group-trained control group both in the combination of all reviewer ratings and all composer ratings. Specifically, the reviewers found the test music to be more artistically effective and more likeable, but less original. The composers found the test music to be more interesting, less creative, less artistically effective, and less chorale-like. This suggests that crowdsourcing might be a more effective training method, but that the difference in sums of Likert-scale answers was small enough that it could be attributed to random chance.

Several modifications or additions to the methodology were suggested, should the experiment be repeated or a similar experiment run. For the genetic algorithm, the mutation rate may need to be changed, more rules should perhaps be added and the existing rules might need different weights. It would be interesting to create music with a 200-generation GA instance using *only* the programmatic fitness functions and compare that to the crowdsourced music. The Turkit toolkit could have been used to simplify HIT management on mTurk.

Rather than asking for best, middle, and worst ratings, a preference judgment might be a better choice. HITs should probably either be limited to one-per-Turker (which slows and limits the work) or the music ratings should be voting-based (which requires multiple workers to listen to the same music). The GA could use the programmatic fitness functions to eliminate obviously bad choices or to evolve on its own a while before human reviewers are brought in. Finally, the similarity/originality question should have been clarified for reviewers and composers. While the above suggestions could have made the results clearer or research process smoother, the research showed that crowdsourcing on mTurk has potential as a solution to the fitness bottleneck in compositional GAs.

Appendix A

Nova Southeastern University IRB Approval



NOVA SOUTHEASTERN UNIVERSITY
Office of Grants and Contracts
Institutional Review Board

MEMORANDUM

To: Jessica Keup

From: Ling Wang, Ph.D.
Institutional Review Board

Date: Dec. 3, 2010

Re: *Computer Music Composition using Crowdsourcing and Genetic Algorithms*

IRB Approval Number: wang10151001

I have reviewed the above-referenced research protocol at the center level. Based on the information provided, I have determined that this study is exempt from further IRB review. You may proceed with your study as described to the IRB. As principal investigator, you must adhere to the following requirements:

1) CONSENT: If recruitment procedures include consent forms these must be obtained in such a manner that they are clearly understood by the subjects and the process affords subjects the opportunity to ask questions, obtain detailed answers from those directly involved in the research, and have sufficient time to consider their participation after they have been provided this information. The subjects must be given a copy of the signed consent document, and a copy must be placed in a secure file separate from de-identified participant information. Record of informed consent must be retained for a minimum of three years from the conclusion of the study.

2) ADVERSE REACTIONS: The principal investigator is required to notify the IRB chair and me (954-262-5369 and 954-262-2020 respectively) of any adverse reactions or unanticipated events that may develop as a result of this study.

Reactions or events may include, but are not limited to, injury, depression as a result of participation in the study, life-threatening situation, death, or loss of confidentiality/anonymity of subject. Approval may be withdrawn if the problem is serious.

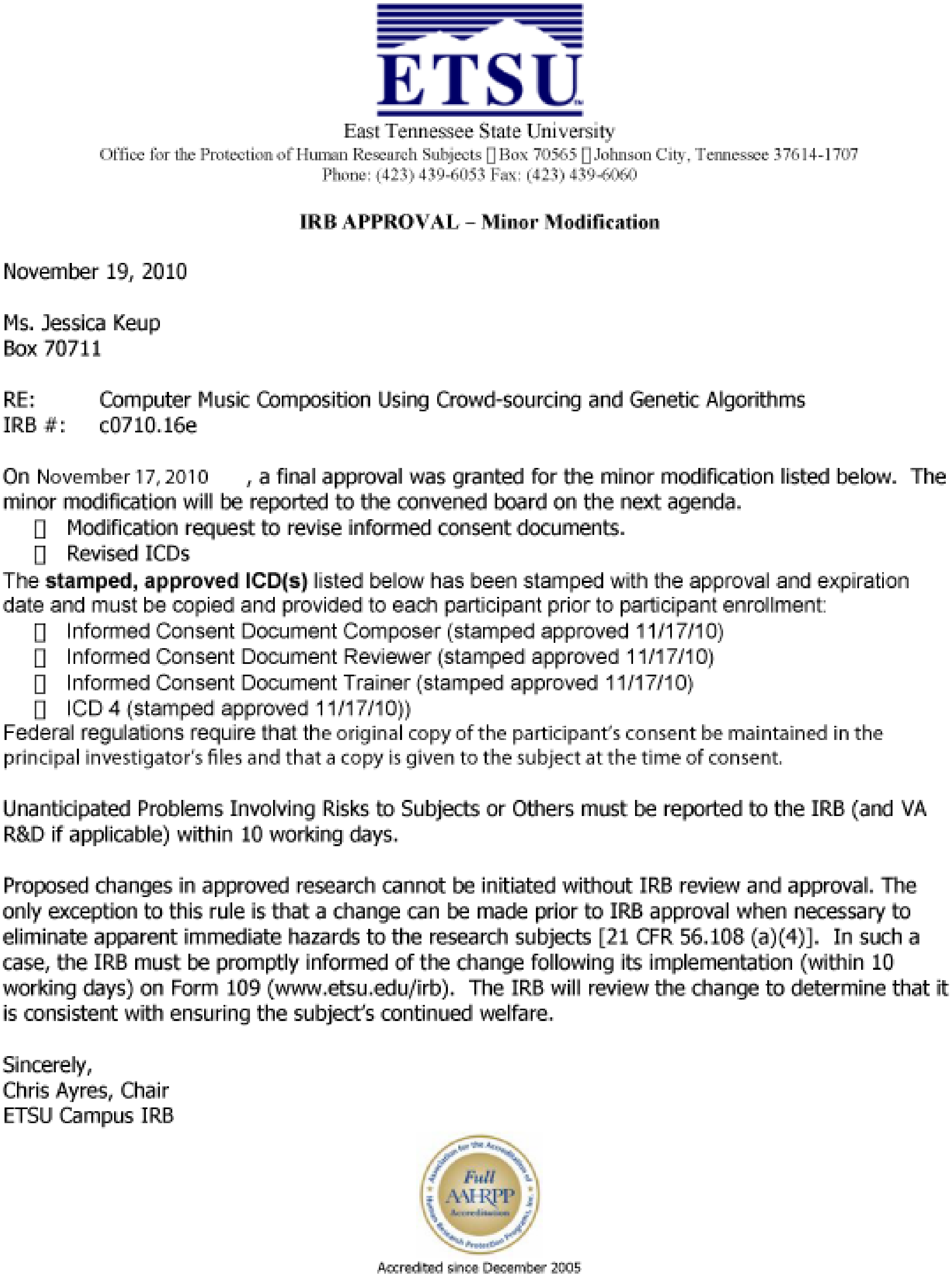
3) AMENDMENTS: Any changes in the study (e.g., procedures, number or types of subjects, consent forms, investigators, etc.) must be approved by the IRB prior to implementation. Please be advised that changes in a study may require further review depending on the nature of the change. Please contact me with any questions regarding amendments or changes to your study.

The NSU IRB is in compliance with the requirements for the protection of human subjects prescribed in Part 46 of Title 45 of the Code of Federal Regulations (45 CFR 46) revised June 18, 1991.

Cc: Protocol File
Office of Grants and Contracts

Appendix B

East Tennessee State University IRB Approval



ETSU

East Tennessee State University
Office for the Protection of Human Research Subjects □ Box 70565 □ Johnson City, Tennessee 37614-1707
Phone: (423) 439-6053 Fax: (423) 439-6060

IRB APPROVAL – Minor Modification

November 19, 2010

Ms. Jessica Keup
Box 70711

RE: Computer Music Composition Using Crowd-sourcing and Genetic Algorithms
IRB #: c0710.16e

On November 17, 2010 , a final approval was granted for the minor modification listed below. The minor modification will be reported to the convened board on the next agenda.

□ Modification request to revise informed consent documents.
□ Revised ICDs

The **stamped, approved ICD(s)** listed below has been stamped with the approval and expiration date and must be copied and provided to each participant prior to participant enrollment:

□ Informed Consent Document Composer (stamped approved 11/17/10)
□ Informed Consent Document Reviewer (stamped approved 11/17/10)
□ Informed Consent Document Trainer (stamped approved 11/17/10)
□ ICD 4 (stamped approved 11/17/10))

Federal regulations require that the original copy of the participant's consent be maintained in the principal investigator's files and that a copy is given to the subject at the time of consent.

Unanticipated Problems Involving Risks to Subjects or Others must be reported to the IRB (and VA R&D if applicable) within 10 working days.

Proposed changes in approved research cannot be initiated without IRB review and approval. The only exception to this rule is that a change can be made prior to IRB approval when necessary to eliminate apparent immediate hazards to the research subjects [21 CFR 56.108 (a)(4)]. In such a case, the IRB must be promptly informed of the change following its implementation (within 10 working days) on Form 109 (www.etsu.edu/irb). The IRB will review the change to determine that it is consistent with ensuring the subject's continued welfare.

Sincerely,
Chris Ayres, Chair
ETSU Campus IRB

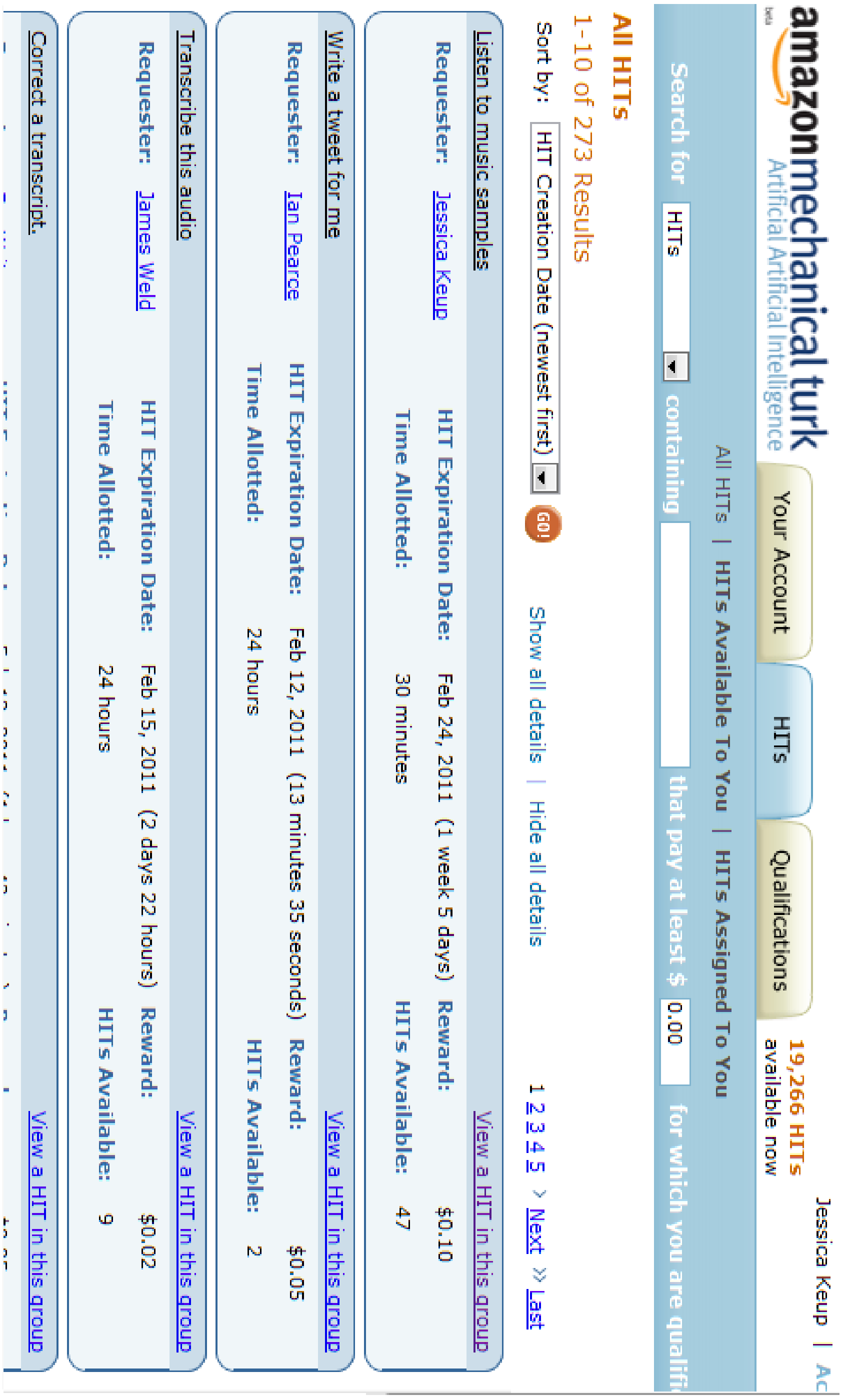
Full AAHRPP Accreditation
Association for the Accreditation of Human Research Protection Programs, Inc.

Accredited since December 2005

Appendix C

Screenshots of User Interface

Listing of HITs Screen



19,266 HITs available now

Jessica Keup | Ac

All HITs | HITs Available To You | HITs Assigned To You

Search for HITs containing that pay at least $ 0.00 for which you are qualifi

All HITs

1-10 of 273 Results

Sort by: HIT Creation Date (newest first) GO! Show all details | Hide all details 1 2 3 4 5 > Next >> Last

Listen to music samples View a HIT in this group

Requester: Jessica Keup HIT Expiration Date: Feb 24, 2011 (1 week 5 days) Reward: $0.10

Time Allotted: 30 minutes HITs Available: 47

Write a tweet for me View a HIT in this group

Requester: Ian Pearce HIT Expiration Date: Feb 12, 2011 (13 minutes 35 seconds) Reward: $0.05

Time Allotted: 24 hours HITs Available: 2

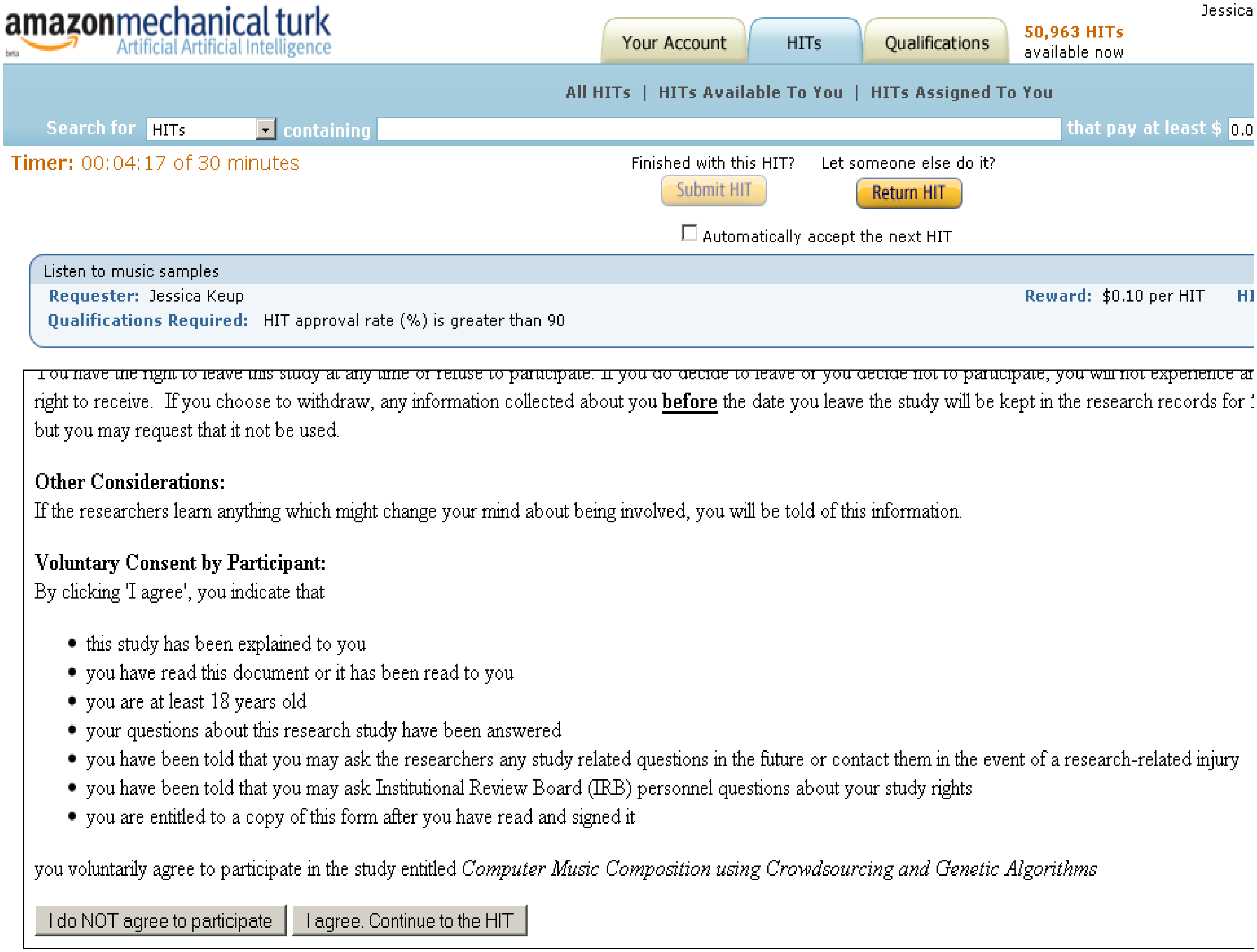
Transcribe this audio View a HIT in this group

Requester: James Weld HIT Expiration Date: Feb 15, 2011 (2 days 22 hours) Reward: $0.02

Time Allotted: 24 hours HITs Available: 9

Correct a transcript. View a HIT in this group

Consent Screen



Timer: 00:04:17 of 30 minutes

Listen to music samples

Requester: Jessica Keup

Reward: $0.10 per HIT

Qualifications Required: HIT approval rate (%) is greater than 90

You have the right to leave this study at any time or refuse to participate. If you do decide to leave or you decide not to participate, you will not experience ar right to receive. If you choose to withdraw, any information collected about you **before** the date you leave the study will be kept in the research records for : but you may request that it not be used.

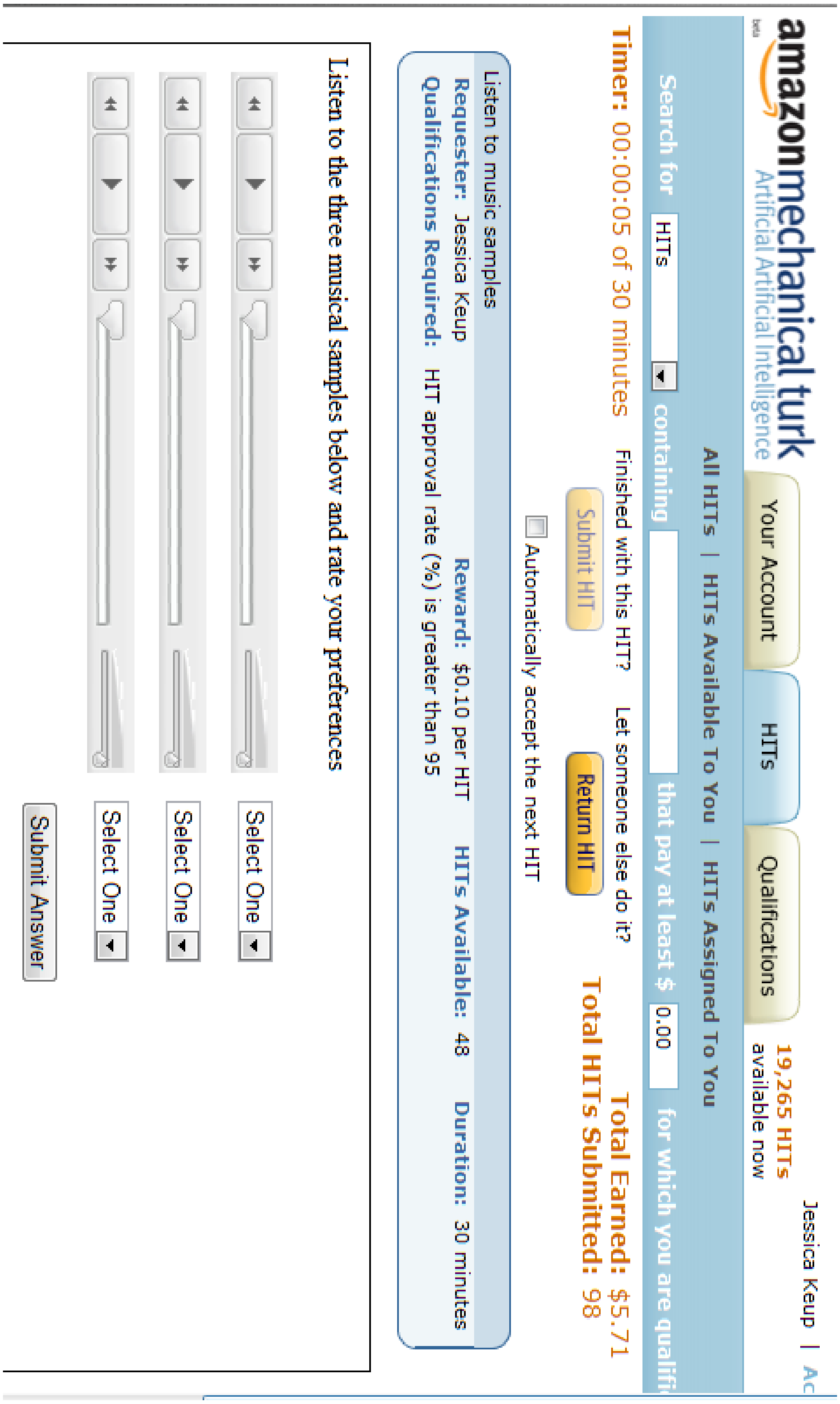
Other Considerations:
If the researchers learn anything which might change your mind about being involved, you will be told of this information.

Voluntary Consent by Participant:
By clicking 'I agree', you indicate that

- this study has been explained to you
- you have read this document or it has been read to you
- you are at least 18 years old
- your questions about this research study have been answered
- you have been told that you may ask the researchers any study related questions in the future or contact them in the event of a research-related injury
- you have been told that you may ask Institutional Review Board (IRB) personnel questions about your study rights
- you are entitled to a copy of this form after you have read and signed it

you voluntarily agree to participate in the study entitled *Computer Music Composition using Crowdsourcing and Genetic Algorithms*

Listening Screen



amazonmechanical turk beta
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
19,265 HITs available now
Jessica Keup | Ac
All HITs | HITs Available To You | HITs Assigned To You
Search for HITs containing that pay at least $ 0.00 for which you are qualifi

Timer: 00:00:05 of 30 minutes

Finished with this HIT? Submit HIT

Let someone else do it? Return HIT

Total Earned: $5.71

Total HITs Submitted: 98

☐ Automatically accept the next HIT

Listen to music samples

Requester: Jessica Keup

Reward: $0.10 per HIT

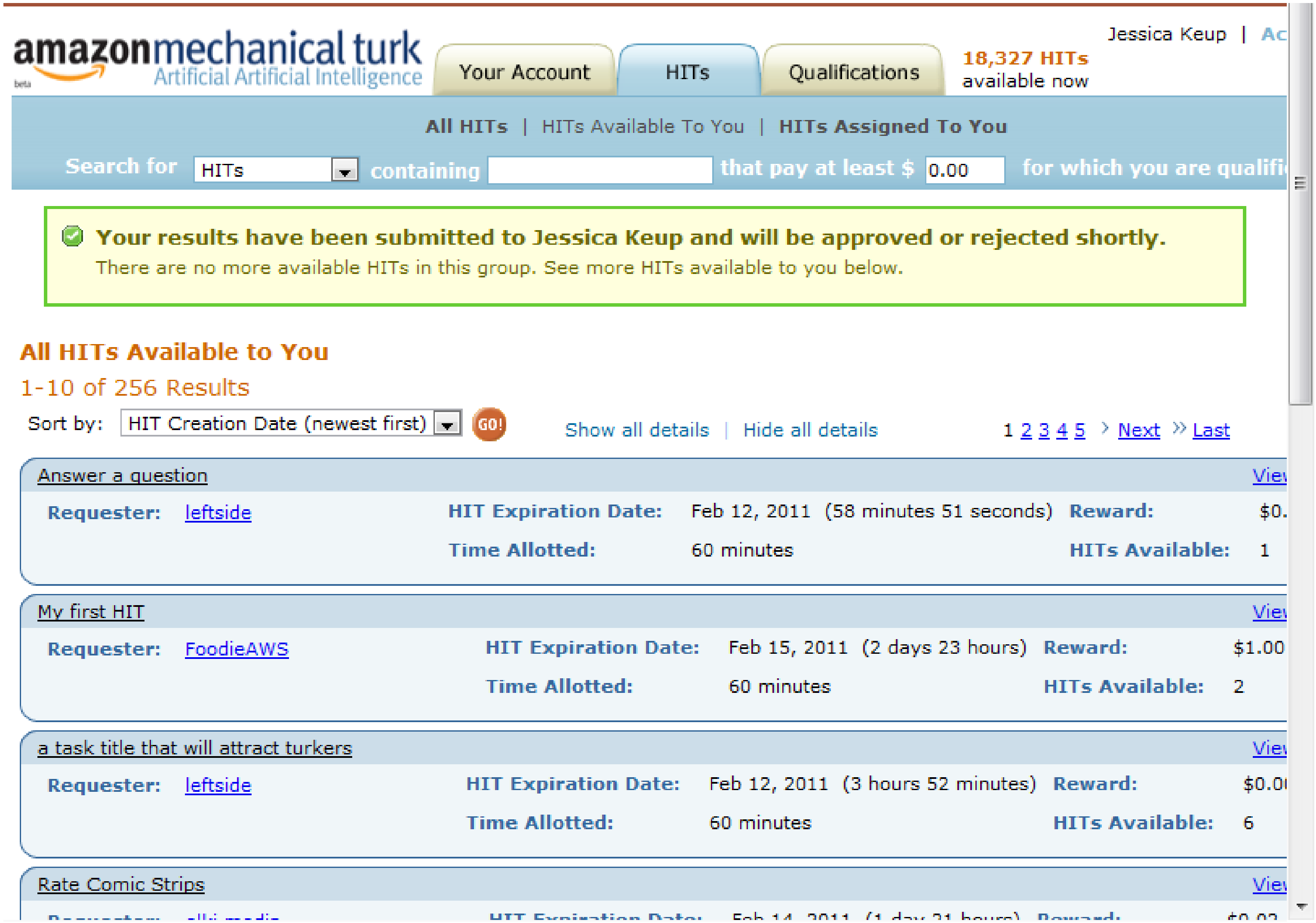
HITs Available: 48

Duration: 30 minutes

Qualifications Required: HIT approval rate (%) is greater than 95

Listen to the three musical samples below and rate your preferences

Confirmation/Thank You Screen



Your results have been submitted to Jessica Keup and will be approved or rejected shortly.

There are no more available HITs in this group. See more HITs available to you below.

All HITs Available to You

1-10 of 256 Results

Answer a question

Requester: leftside HIT Expiration Date: Feb 12, 2011 (58 minutes 51 seconds) Reward: $0.

Time Allotted: 60 minutes HITs Available: 1

My first HIT

Requester: FoodieAWS HIT Expiration Date: Feb 15, 2011 (2 days 23 hours) Reward: $1.00

Time Allotted: 60 minutes HITs Available: 2

a task title that will attract turkers

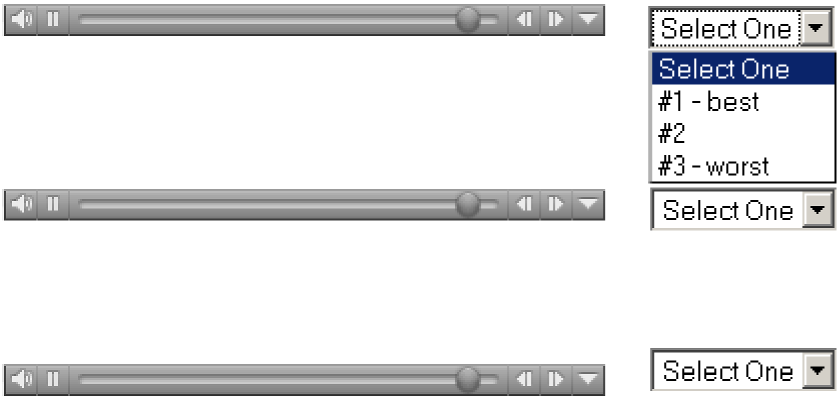
Requester: leftside HIT Expiration Date: Feb 12, 2011 (3 hours 52 minutes) Reward: $0.0

Time Allotted: 60 minutes HITs Available: 6

Rate Comic Strips

Appendix D
Trainer Instructions

Listen to the three musical samples below and rate your preferences.



Appendix E

Trainer Recruitment

Participants needed for research in musical artificial intelligence

We are looking for volunteers to take part in a study to evaluate different methods of training artificial intelligence programs to compose music.

As a participant in this study, you would be asked to listen to samples of music created by an artificial intelligence composition program and evaluate them. You must be at least 18 years old to participate.

Your participation would involve one session lasting approximately one hour.
In appreciation for your time, you will receive a $10 Amazon.com giftcard.

For more information about this study, or to volunteer for this study, please contact:

Jessica Keup
ETSU Department of Computer and Information Sciences
Nova Southeastern University School of Computer and Information Sciences
(423) 439-6963 or
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Musical AI Training Research
(423) 439-6963
keup@etsu.edu

Appendix F

Trainer Consent Form

Consent Form for Participation in the Research Study Entitled
Computer Music Composition using Crowdsourcing and Genetic Algorithms

Funding Source: None.
IRB protocol #: wang10151001

Principal investigator
Jessica Keup, MHCI
ETSU Box 70711
Johnson City, TN 37614
(423) 439-6963

Co-investigator
Maxine Cohen, Ph.D.
3301 College Avenue
Fort Lauderdale, FL 33314
(954) 262-2072

For questions/concerns about your research rights, contact:
Human Research Oversight Board (Institutional Review Board or IRB)
Nova Southeastern University
(954) 262-5369/Toll Free: 866-499-0790
IRB@nsu.nova.edu

Site Information
East Tennessee State University
Department of Computer and Information Sciences
807 University Pkwy
Johnson City, TN 37614

What is the study about?
You are invited to participate in a research study. The goal of this study is to compare training methods of genetic algorithms that create music.

Why are you asking me?
We are inviting you to participate because we need a large number of people to listen to and review music. There will be approximately 5,043 participants in this research study.

What will I be doing if I agree to be in the study?
You will listen to 25 sets of 3 very short songs and rank them to indicate which you think is best. These tasks should take no more than 1 hour to complete.

Is there any audio or video recording?
There is no audio or video recording. Only your responses to the questions will be kept.

What are the dangers to me?
Risks to you are minimal, meaning they are not thought to be greater than other risks you experience everyday. If you have questions about the research, your research rights, or if you experience an injury because of the research please contact Ms. Keup at (423) 439-6963. You may also contact the IRB at the numbers indicated above with questions about your research rights.

Initials: _________ Date: _________ **Page 1 of 2**

Are there any benefits to me for taking part in this research study?
There are no benefits to you for participating.

Will I get paid for being in the study? Will it cost me anything?
It will not cost you anything to participate in the study. You will receive a $10 Amazon gift card to compensate you for your time.

How will you keep my information private?
Your responses will only include your ratings of the music, and no personally identifying information will be gathered. All information obtained in this study is strictly confidential unless disclosure is required by law. The IRB, regulatory agencies, or Dr. Cohen may review research records.

What if I do not want to participate or I want to leave the study?
You have the right to leave this study at any time or refuse to participate. If you do decide to leave or you decide not to participate, you will not experience any penalty or loss of services you have a right to receive. If you choose to withdraw, any information collected about you **before** the date you leave the study will be kept in the research records for five years from the conclusion of the study but you may request that it not be used.

Other Considerations:
If the researchers learn anything which might change your mind about being involved, you will be told of this information.

Voluntary Consent by Participant:
By signing below, you indicate that

- this study has been explained to you
- you have read this document or it has been read to you
- you are at least 18 years old
- your questions about this research study have been answered
- you have been told that you may ask the researchers any study related questions in the future or contact them in the event of a research-related injury
- you have been told that you may ask Institutional Review Board (IRB) personnel questions about your study rights
- you are entitled to a copy of this form after you have read and signed it
- you voluntarily agree to participate in the study entitled *Computer Music Composition using Crowdsourcing and Genetic Algorithms*

Participant's Signature: ___________________________ Date: ________________

Participant’s Name: ______________________________ Date: ________________

Signature of Person Obtaining Consent: _____________________________

Date: ___________________________

Appendix G

Reviewer Instructions

For this study, you will be asked to listen to ten musical excerpts. You may listen to them as many times as you like. You will be asked to rate each excerpt on a number of aspects and describe your opinion of its merits or shortcomings.

Excerpt 1

Question	**Strongly Disagree**	**Somewhat Disagree**	**Neutral**	**Somewhat Agree**	**Strongly Agree**
a. I like this music					
b.This music is artistically effective					
c.This music sounds similar to things I’ve heard before					

What, if any, emotion(s) does it evoke?

What, if anything, was memorable about it?

What, if anything, were its shortcomings?

[repeated for Excerpts 2 – 10]

Do you have any **overall** comments about the selections?

Appendix H

Composer Instructions

For this study, you will be asked to listen to ten musical excerpts. You may listen to them as many times as you like. You will be asked to rate each excerpt on a number of aspects and describe your opinion of its merits or shortcomings.

Excerpt 1

Question	**Strongly Disagree**	**Somewhat Disagree**	**Neutral**	**Somewhat Agree**	**Strongly Agree**
a. This music is interesting					
b.This music is creative					
c.This music is artistically effective					
d.This music is chorale-like					

What, if anything, was memorable about it?

____

____

What, if anything, were its shortcomings?

____

____

[repeated for Excerpts 2 – 10]

Do you have any **overall** comments about the selections?

____

____

____

____

Appendix I

Reviewer Recruitment

Participants needed for research in musical artificial intelligence

We are looking for volunteers to take part in a study to evaluate different methods of training artificial intelligence programs to compose music.

As a participant in this study, you would be asked to listen to samples of music created by an artificial intelligence composition program and evaluate them. You must be at least 18 years old to participate.

Your participation would involve one session lasting approximately 30 minutes.
In appreciation for your time, you will receive a $5 Amazon.com giftcard.

For more information about this study, or to volunteer for this study, please contact:

Jessica Keup
ETSU Department of Computer and Information Sciences
Nova Southeastern University School of Computer and Information Sciences
(423) 439-6963 or
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Musical AI Training
Research
(423) 439-6963
keup@etsu.edu

Appendix J

Reviewer Consent Form

Consent Form for Participation in the Research Study Entitled
Computer Music Composition using Crowdsourcing and Genetic Algorithms

Funding Source: None.
IRB protocol #: wang10151001

Principal investigator	Co-investigator
Jessica Keup, MHCI	Maxine Cohen, Ph.D.
ETSU Box 70711	3301 College Avenue
Johnson City, TN 37614	Fort Lauderdale, FL 33314
(423) 439-6963	(954) 262-2072

For questions/concerns about your research rights, contact:
Human Research Oversight Board (Institutional Review Board or IRB)
Nova Southeastern University
(954) 262-5369/Toll Free: 866-499-0790
IRB@nsu.nova.edu

Site Information
East Tennessee State University
Department of Computer and Information Sciences
807 University Pkwy
Johnson City, TN 37614

What is the study about?
You are invited to participate in a research study. The goal of this study is to compare training methods of genetic algorithms that create music.

Why are you asking me?
We are inviting you to participate because we need a large number of people to listen to and review music. There will be approximately 5,043 participants in this research study.

What will I be doing if I agree to be in the study?
You will listen to 10 very short songs, rate them on a number of scales, and describe your opinions of them. These tasks should take no more than 30 minutes to complete.

Is there any audio or video recording?
There is no audio or video recording. Only your responses to the questions will be kept.

What are the dangers to me?
Risks to you are minimal, meaning they are not thought to be greater than other risks you experience everyday. If you have questions about the research, your research rights, or if you experience an injury because of the research please contact Ms. Keup at (423) 439-6963. You may also contact the IRB at the numbers indicated above with questions about your research rights

Initials: _________ Date: _________

Are there any benefits to me for taking part in this research study?
There are no benefits to you for participating.

Will I get paid for being in the study? Will it cost me anything?
It will not cost you anything to participate in the study. You will receive a $5 Amazon gift card to compensate you for your time.

How will you keep my information private?
Your responses will only include your ratings and descriptions of the music, and no personally identifying information will be gathered. All information obtained in this study is strictly confidential unless disclosure is required by law. The IRB, regulatory agencies, or Dr. Cohen may review research records.

What if I do not want to participate or I want to leave the study?
You have the right to leave this study at any time or refuse to participate. If you do decide to leave or you decide not to participate, you will not experience any penalty or loss of services you have a right to receive. If you choose to withdraw, any information collected about you **before** the date you leave the study will be kept in the research records for five years from the conclusion of the study but you may request that it not be used.

Other Considerations:
If the researchers learn anything which might change your mind about being involved, you will be told of this information.

Voluntary Consent by Participant:
By signing below, you indicate that

- this study has been explained to you
- you have read this document or it has been read to you
- you are at least 18 years old
- your questions about this research study have been answered
- you have been told that you may ask the researchers any study related questions in the future or contact them in the event of a research-related injury
- you have been told that you may ask Institutional Review Board (IRB) personnel questions about your study rights
- you are entitled to a copy of this form after you have read and signed it
  you voluntarily agree to participate in the study entitled *Computer Music Composition using Crowdsourcing and Genetic Algorithms*

Participant's Signature: ____________________________ Date: ________________

Participant’s Name: _______________________________ Date: ________________

Signature of Person Obtaining Consent: _____________________________

Date: ____________________________

Appendix K

Composer Recruitment Form

Participants needed for research in musical artificial intelligence

We are looking for volunteers to take part in a study to evaluate different methods of training artificial intelligence programs to compose music.

As a participant in this study, you would be asked to listen to samples of music created by an artificial intelligence composition program and evaluate them. You must be at least 18 years old to participate.

Your participation would involve one session lasting approximately thirty minutes
In appreciation for your time, you will receive a $25 Amazon.com giftcard.

For more information about this study, or to volunteer for this study, please contact:

Jessica Keup
ETSU Department of Computer and Information Sciences
Nova Southeastern University School of Computer and Information Sciences
(423) 439-6963 or
keup@etsu.edu

Appendix L

Composer Consent Form

Consent Form for Participation in the Research Study Entitled
Computer Music Composition using Crowdsourcing and Genetic Algorithms

Funding Source: None.
IRB protocol #: wang10151001

Principal investigator
Jessica Keup, MHCI
ETSU Box 70711
Johnson City, TN 37614
(423) 439-6963

Co-investigator
Maxine Cohen, Ph.D.
3301 College Avenue
Fort Lauderdale, FL 33314
(954) 262-2072

For questions/concerns about your research rights, contact:
Human Research Oversight Board (Institutional Review Board or IRB)
Nova Southeastern University
(954) 262-5369/Toll Free: 866-499-0790
IRB@nsu.nova.edu

Site Information
East Tennessee State University
Department of Computer and Information Sciences
807 University Pkwy
Johnson City, TN 37614

What is the study about?
You are invited to participate in a research study. The goal of this study is to compare training methods of genetic algorithms that create music.

Why are you asking me?
We are inviting you to participate because we need a large number of people to listen to and review music. There will be approximately 5,043 participants in this research study.

What will I be doing if I agree to be in the study?
You will listen to 10 very short songs, rate them on a number of scales, and describe your opinions of them. These tasks should take no more than 30 minutes to complete.

Is there any audio or video recording?
There is no audio or video recording. Only your responses to the questions will be kept.

What are the dangers to me?
Risks to you are minimal, meaning they are not thought to be greater than other risks you experience everyday. If you have questions about the research, your research rights, or if you experience an injury because of the research please contact Ms. Keup at (423) 439-6963. You may also contact the IRB at the numbers indicated above with questions about your research rights.

Initials: _________ Date: _________

Are there any benefits to me for taking part in this research study?
There are no benefits to you for participating.

Will I get paid for being in the study? Will it cost me anything?
It will not cost you anything to participate in the study. You will receive a $25 Amazon gift card to compensate you for your time.

How will you keep my information private?
Your responses will only include your ratings and descriptions of the music, and no personally identifying information will be gathered. All information obtained in this study is strictly confidential unless disclosure is required by law. The IRB, regulatory agencies, or Dr. Cohen may review research records.

What if I do not want to participate or I want to leave the study?
You have the right to leave this study at any time or refuse to participate. If you do decide to leave or you decide not to participate, you will not experience any penalty or loss of services you have a right to receive. If you choose to withdraw, any information collected about you **before** the date you leave the study will be kept in the research records for five years from the conclusion of the study but you may request that it not be used.

Other Considerations:
If the researchers learn anything which might change your mind about being involved, you will be told of this information.

Voluntary Consent by Participant:
By signing below, you indicate that

- this study has been explained to you
- you have read this document or it has been read to you
- you are at least 18 years old
- your questions about this research study have been answered
- you have been told that you may ask the researchers any study related questions in the future or contact them in the event of a research-related injury
- you have been told that you may ask Institutional Review Board (IRB) personnel questions about your study rights
- you are entitled to a copy of this form after you have read and signed it
- you voluntarily agree to participate in the study entitled *Computer Music Composition using Crowdsourcing and Genetic Algorithms*

Participant's Signature: ___________________________ Date: ________________

Participant’s Name: ______________________________ Date: ________________

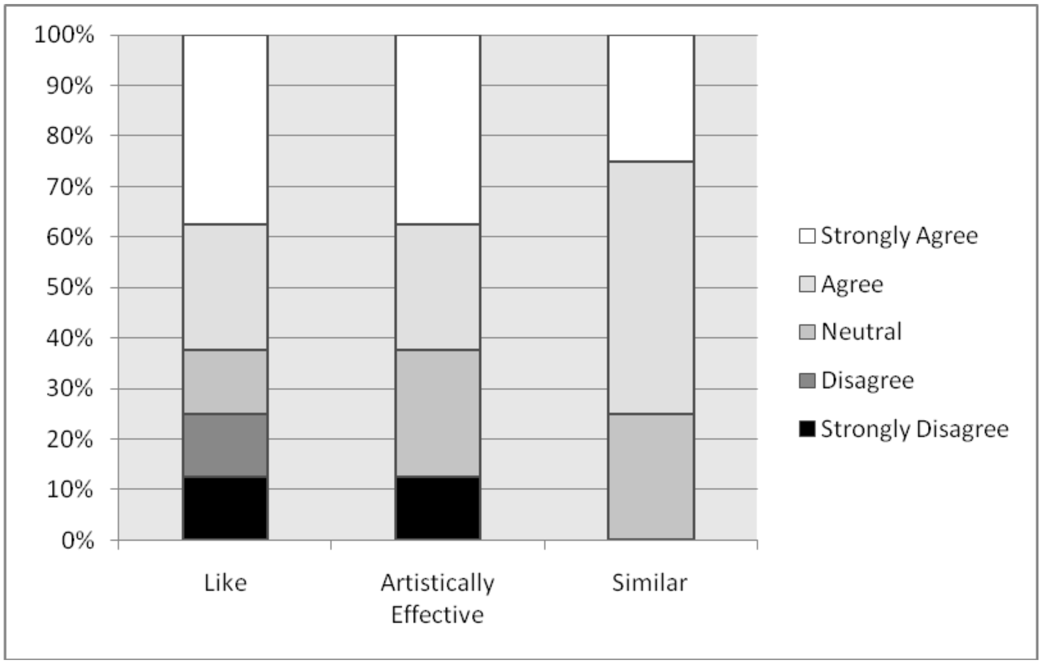
Signature of Person Obtaining Consent: _____________________________

Date: ___________________________

Appendix M

Reviews of Control GA Music by the Non-Musically Trained

Song 1:



What, if any, emotion(s) does it evoke?

- still a dark feeling
- same as previous
- this captures the same emotions as *[above]* - curiosity mostly
- suspense, again
- made me feel medieval
- Reminiscent of the previous example, but with much deeper, darker inclinations. If the previous example was a broken ballet dancer music box, this one is the small boy that’s breaking it while it plays.

What, if anything, was memorable about it?

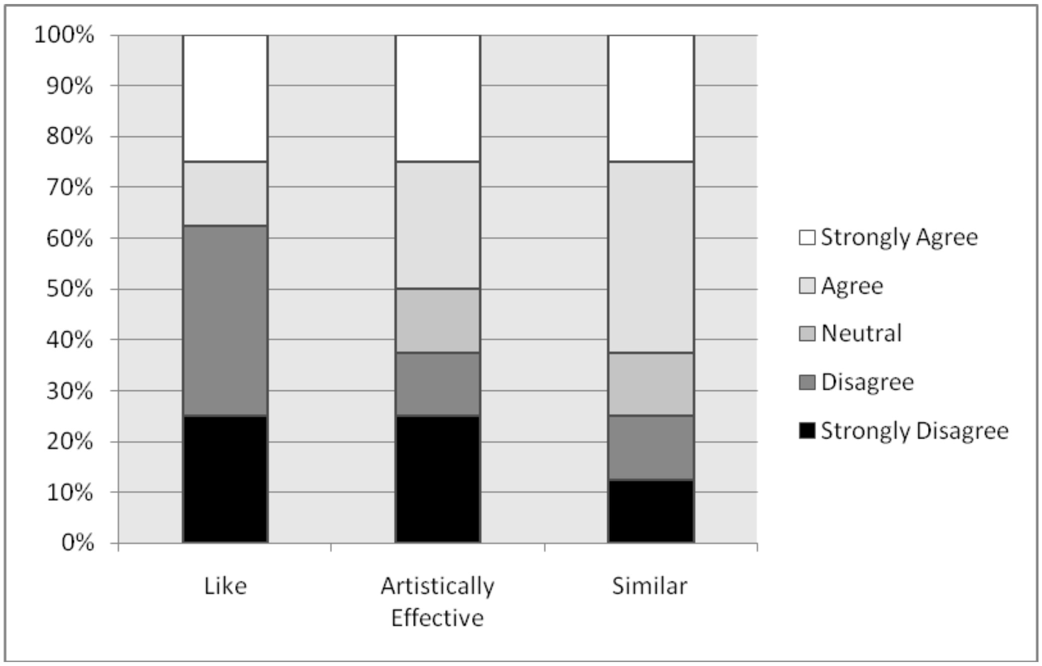
- felt like a snip of a story being told

- ballet
- again, I heard a melody and counter-melody in their infancy
- again, it was Debussy-like. It would not be stuck in my head as melody, but the chords and whole tones are memorable.
- Immediate emotional response. I didn't have to wait for it.
- I liked the 'big' feel of the score
- Nice melodies here, and the countermelody/harmonies were nice

What, if anything, were its shortcomings?

- I don't like the low part
- Same as *[previous]*
- None
- I don't think the melody & countermelody belonged together in this one, the combination of the two was very jarring.

Song 2:



What, if any, emotion(s) does it evoke?

- dark/maybe explaining a story
- curiosity
- confusion
- not sure
- the opening made me anxious
- Another bright piece, this one brings images of children playing while outside while a storm approaches

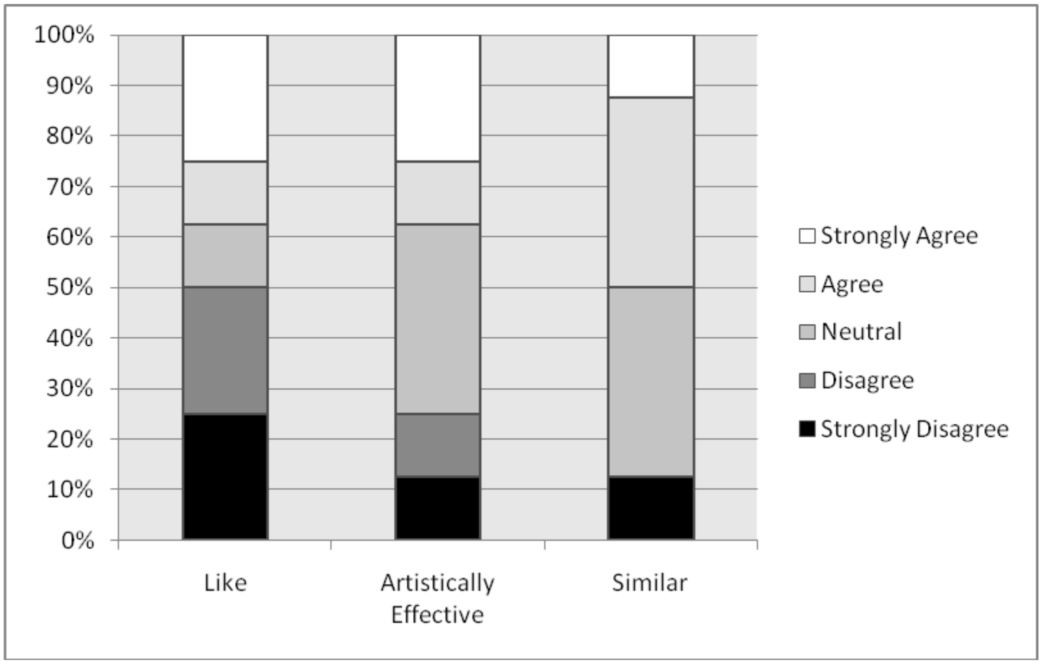
What, if anything, was memorable about it?

- reminds me about an intense video game scene music
- parallel octaves in the middle and weird syncopation
- nothing
- sounds “clanky”
- I liked the melodies and countermelodies, the dissonances were effective

What, if anything, were its shortcomings?

- Tunes are not in accordance
- Same as *[previous 9]*
- It was not as successful as the others like it because it was not harmonious
- Disorganized
- Too staccato, and sounded ‘tinny’
- None noticeable

Song 3



What, if any, emotion(s) does it evoke?

- again still a mix of dark and happy feeling
- suspense again, as one traditionally associates with dodecaphonic music in a movie or an opera. Tension.
- I wanted to slow it down or speed it up. It should not have all been the same tempo. To be affective it would need to be played with emotion
- A little more upbeat
- Again, a bit too much chaos in this one, like a youngster with a bit of piano lessons under their belt, just banging away and trying new things without really knowing how things “should” sound

What, if anything, was memorable about it?

- this reminded me of a mystery theme music
- not much
- The anxiety of it. The excerpt itself is not memorable.
- Has a “melody”

- Some of the faster melodic movement was intriguing, but seemed out of place with the rest of the piece

What, if anything, were its shortcomings?

- Again, like example *[1]*, it was too rhythmic and awkwardly syncopated

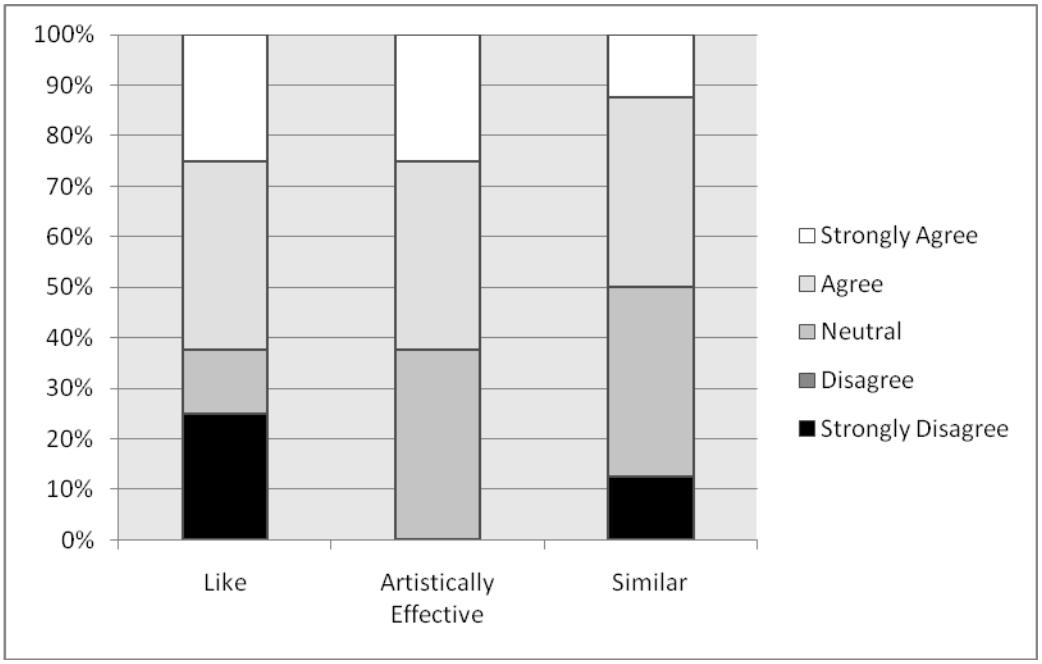
- Mentioned in *[1st question]*. It feels too mechanical, therefore it doesn't make sense.

- Not bad – for a computer

- I didn't like the pace . . . seemed to start and stop abruptly.

- The great distance between the highs and lows as well as the seeming incongruity between melodic and harmonic progression makes this one a bit hard to listen to, unless you're a fan of experimental music ☺

Song 4



What, if any, emotion(s) does it evoke?

- this had a darker feeling than most of the previous music

- suspense

- wonder. It’s slightly scary, but at the same time evokes curiosity.
- Sounds like one of the earlier excerpts. Sadness – somber
- Had a cheerful cadence.
- Longing, nostalgia, a bit morose

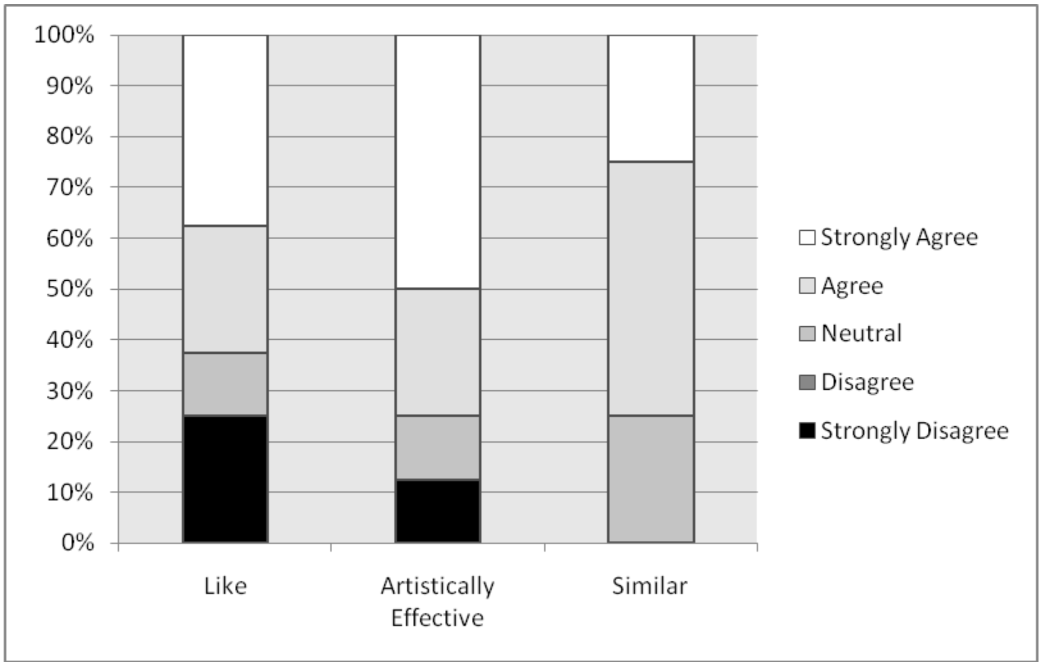
What, if anything, was memorable about it?

- felt like the music was trying to explain a story
- it was too rhythmic for dodecaphonic music – its awkward syncopation was memorable.
- It reminds me of Debussy. That piece of music could be made into a symphony piece.
- Very low notes, very high notes. Low bass – high treble
- Good harmonics and countermelodies

What, if anything, were its shortcomings?

- The tunes are not in accordance
- It sounded like Schoenberg. I hate Schoenberg ☺
- At the end, the rhythm seems a bit jazzy and doesn’t fit the mood.
- Some of the countermelodies were extremely off-putting from the rest of the overall tone

Song 5



What, if any, emotion(s) does it evoke?

- dark/sad or explanation feeling
- amusement
- None, it makes me curious to the thinking of the composer when composing. More analytical.
- Could be used in a drama for transition between scenes
- Has a bright feel to it.
- Black Swan ☺ Images of a broken music box ballet dancer rotating in a caterwauling motion, instead of the gracefulness one would expect

What, if anything, was memorable about it?

- Piano
- Not much, other than no major third at the end ☺
- It would be hard to repeat, since it was so disjunct. Not memorable.
- This one is very “dancelike”, dreamlike, the melody in particular was nice.

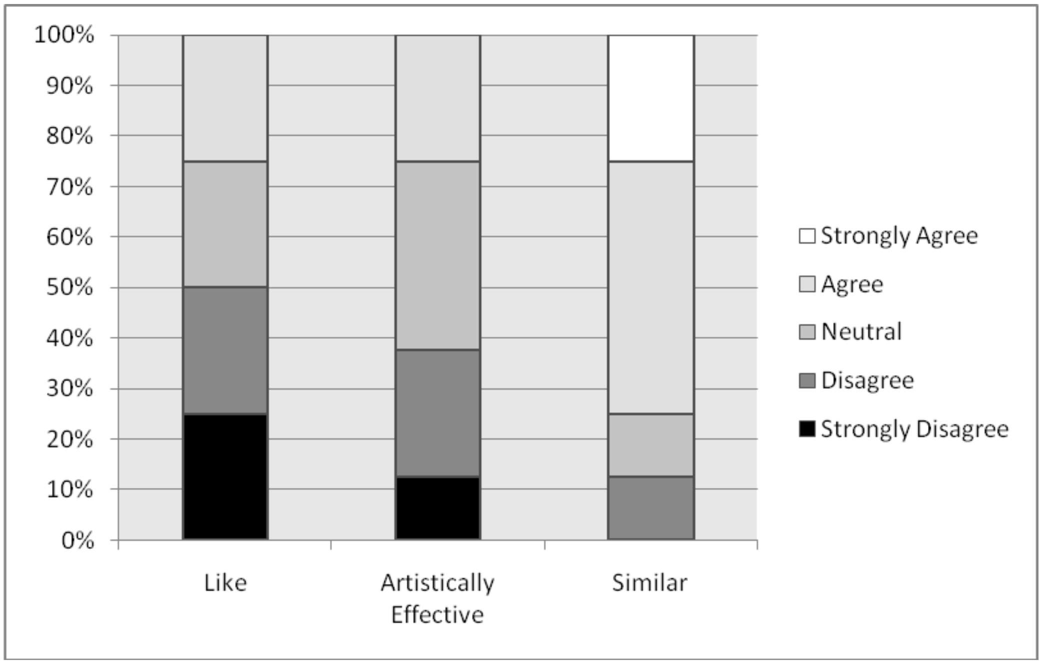
What, if anything, were its shortcomings?

- Same
- It sounds like modern compositions in some ways, but to me it sounds like a child randomly playing notes.
- None to note

Appendix N

Reviews of Test GA Music by the Non-Musically Trained

Song 6



What, if any, emotion(s) does it evoke?

- dark emotion
- same as examples *[1 & 2]*
- anxiousness. Frustration.
- Not sure
- Dark and moody, sinister, chaotic

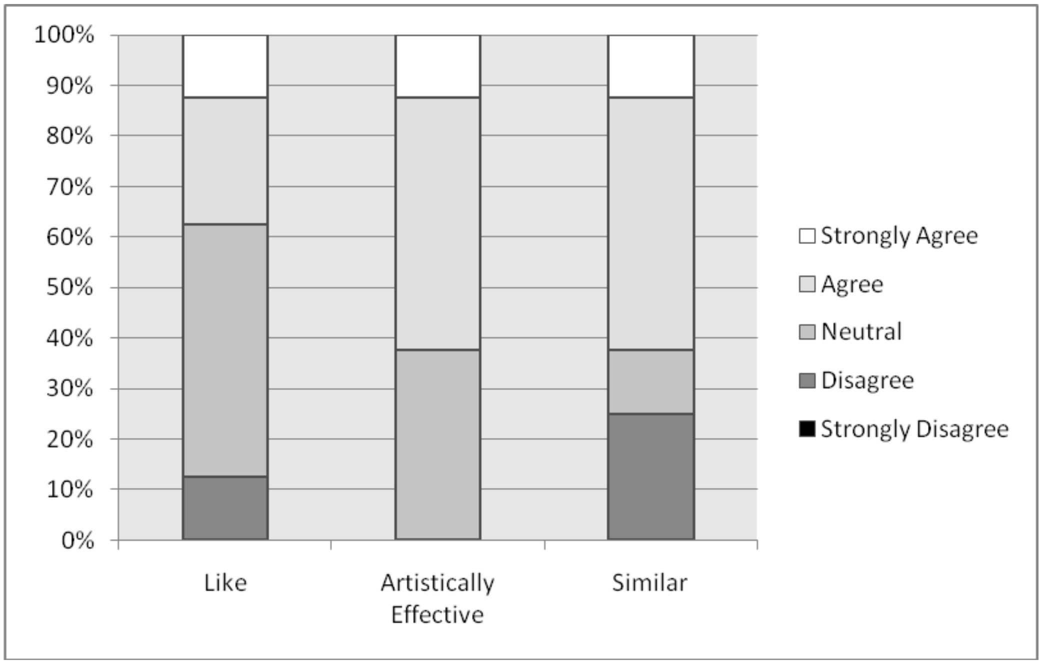
What, if anything, was memorable about it?

- I remembered a major third at the end
- Nothing. Only the bit of melody lead in the beginning.
- Lots of dissonance and large intervals that stand out

What, if anything, were its shortcomings?

- didn't flow very well
- The tunes are not compatible
- Same as *[previous]*
- It did not flow together. The notes seemed random; out of place as well as the rhythm.
- Too many pauses.
- Rhythm could be enhanced
- Some of the intervals are a bit too striking

Song 7



What, if any, emotion(s) does it evoke?

- this made me "perk" up. Still dark but very interesting
- again, curiosity

- Again, a sense of mystery, but towards the end a bit of frustration.
- Suspense
- Playful, adventurous, trepidation

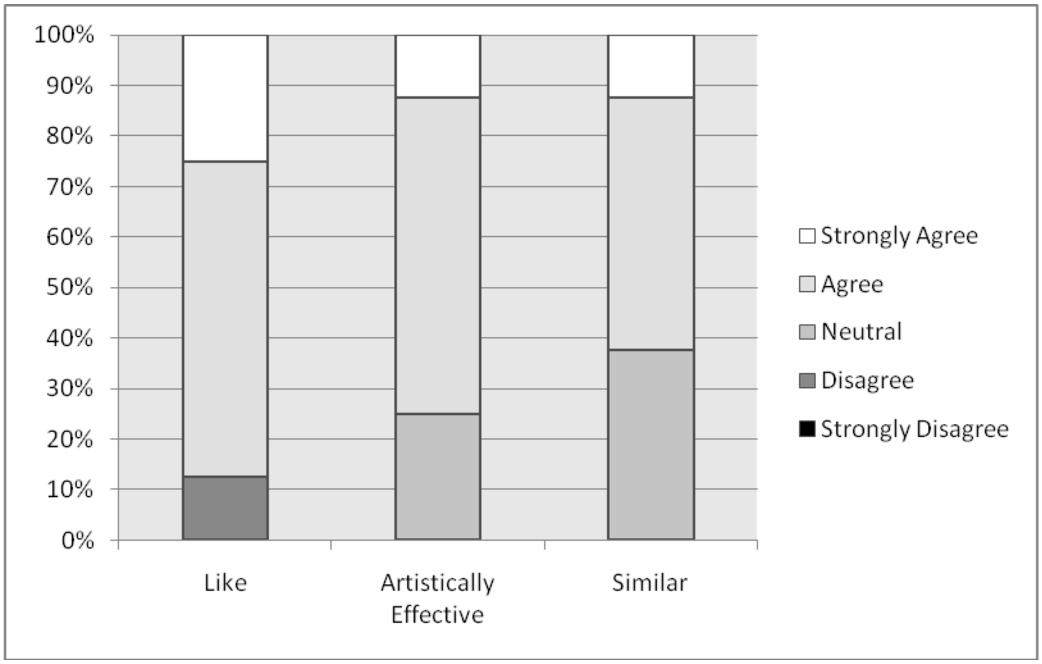
What, if anything, was memorable about it?

- this song felt more like a classical artist
- a few seconds into the piece I thought I heard a hint of a Chopin etude
- The beginning riff. I don’t think the whole thing was memorable because there is no resolutions.
- I liked the strong opening, somewhat like a movie soundtrack.
- This felt more cohesive and “bright”

What, if anything, were its shortcomings?

- The ending part sounds not like the end
- This might have sounded better with variable temp, maybe a couple of tenutos
- Not following through with the set-up expectations.
- Still a bit irregular
- Some of the harmonies tended to stretch into the newer motive, leading to an inconsistent tone at times

Song 8



What, if any, emotion(s) does it evoke?

- this was still dark feeling but had a little happier feeling
- curiosity
- mystery and curiosity
- dramatic
- I imagine this is something that Bette Midler would write when she were drunk ☺. Again, very much like a modern “ballad”. Picture a lounge singer, that’s had the same gig for 50 years, and it’s the end of the night and there’s only a handful or regulars at the bar, and this is what comes out.

What, if anything, was memorable about it?

- still sounded like video game music
- toward the end the harmonies chanced on something resembling an altered dominant
- The melody. This could be a successful version for the murder-mystery theme.
- Reminds me of a soap opera

- I really liked the lower countermelody on this one.

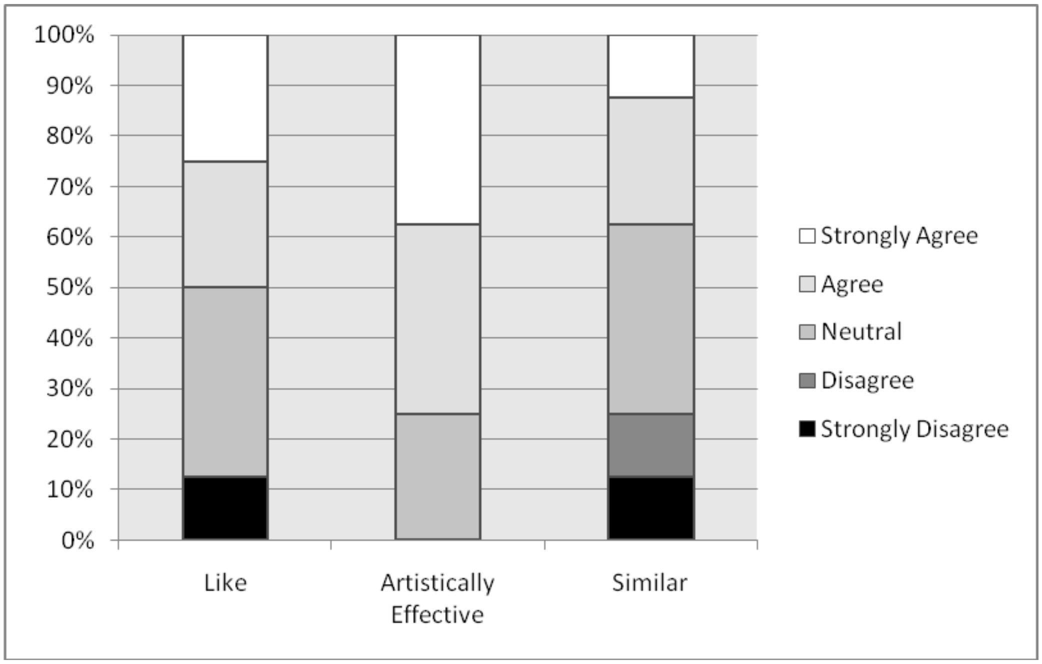
What, if anything, were its shortcomings?

- not compatible
- same as *[previous]*
- Too steady of a tempo. It felt like it needed to rest on some notes to communicate the feeling of suspense

.

- Started out strong but seemed to lose something at the end.

Song 9



What, if any, emotion(s) does it evoke?

- dark w/some happy moments
- give me the feeling like water under a layer of ice
- same as *[previous]*
- mystery
- uplifting

- There’s a hint of mainstream “torchsong” or “ballad” style here

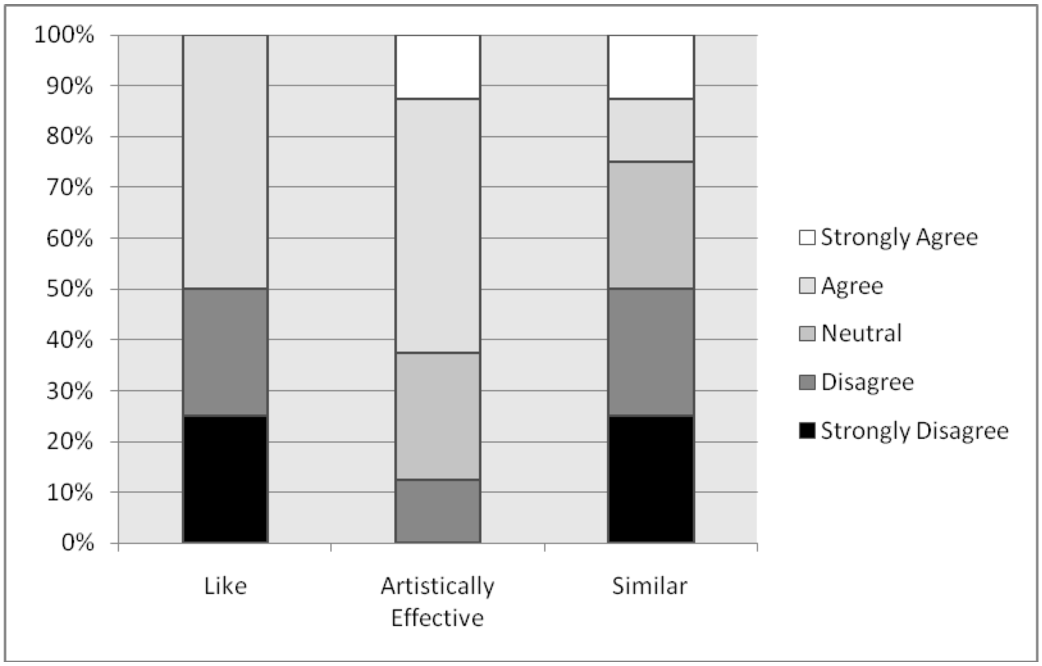
What, if anything, was memorable about it?

- story being told
- at one point in the middle I heard a curious counter-melody. This excerpt is more interesting than *[previous]*.
- The beginning is something catchy – it could even be the beginning of a murder-mystery show theme song.
- It seemed well composed
- Nice harmonics and chord progressions here, some of the melodies were very nice

What, if anything, were its shortcomings?

- Same as *[previous]*
- There was not enough repetition of the initial theme that was set up.
- In both the melody & the harmony (or countermelody) there were too many extreme highs and lows that just distracted the listener from the rest of the piece

Song 10



What, if any, emotion(s) does it evoke?

- dark/sad feeling
- give me a feeling of a dream about ocean
- frustration
- confusion but a want for the melody to come through = frustration
- somber
- This is another one that brings to mind unrequited or forgone love . . . I could see this in some sort of broadway musical

What, if anything, was memorable about it?

- story trying to be told
- I heard the major third at the end again
- The melody is coming through more. I can remember the absence of much harmony.
- Irregular
- The melody & countermelody meshed much better here

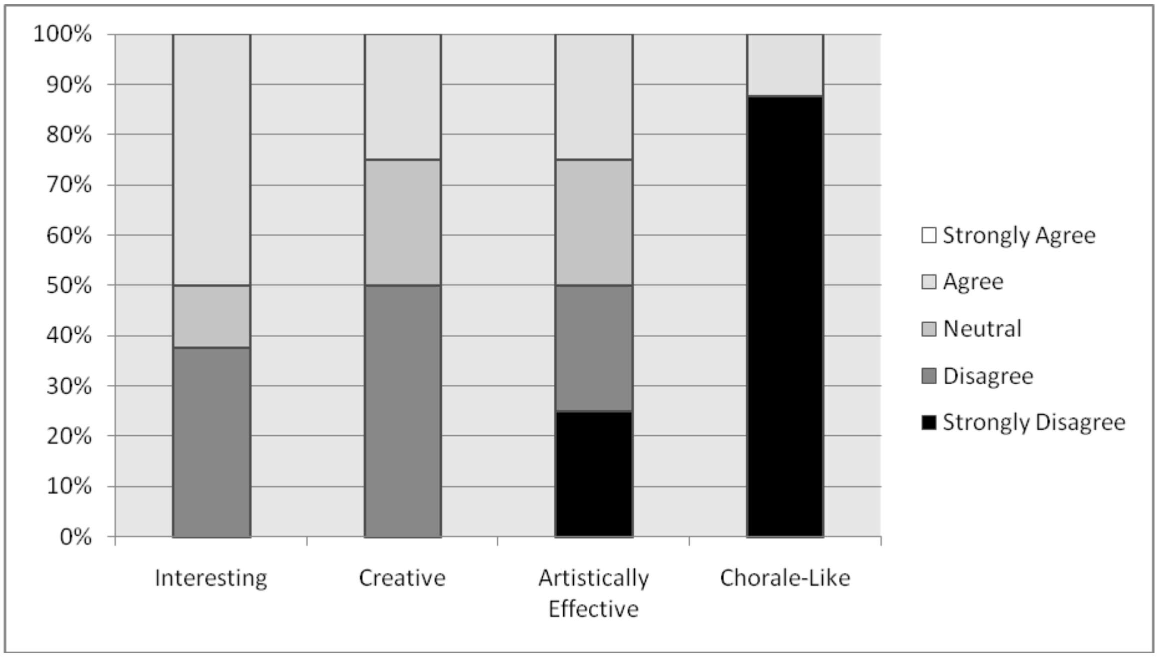
What, if anything, were its shortcomings?

- I don’t like the low part
- I think the awkward syncopation interferes with what could be interesting melody lines.
- The bass seemed to be competing with the treble melody.
- This excerpt did not sound like anything that I have heard in the past.
- Too simple, overly repetitive.
- While the melody & countermelody went well together, it felt at times as if they were competing with one another, and there was quite a bit of “dead space” on either end, where the harmonics weren’t carried over to maintain the theme/tone

Appendix O

Reviews of Control GA Music by Composers

Song 1:



What, if anything, was memorable about it?

- Dissonant “chord” progressions
- The initial phrase imitation signaled ‘structure in time.’
- Nothing
- Same comments as *[Song No. 1]*
- Opening motive is answered in bass.
- Abrupt ending

What, if anything, were its shortcomings?

- No phrase shape
- However, the ‘structure in time’ made no effort to organize via meter
- Randomness

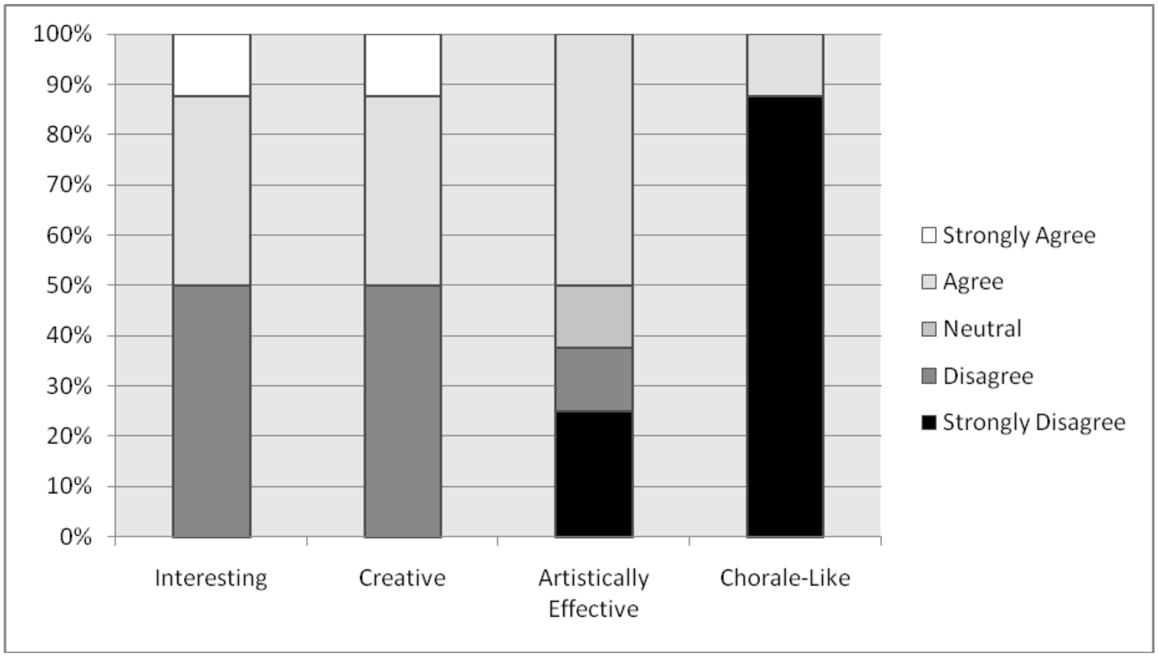
- Same as *[song 1]*. It’s just notes, raw material, that could generate something interesting with lots of sculpting.

- Same comments as *[Song No. 1]*

- Lacks coherence throughout.

- Lacked sense of form

Song 2



What, if anything, was memorable about it?

- Layered

- Is this *[Song 7]* revisited? How different are the two?

- With rhythmic interplay that develops and a harmonic vocabulary that hints at tonality while still being quite dissonant and modern, this is the most interesting of the bunch.

- Opening gesture reminded me of *[song 2]*

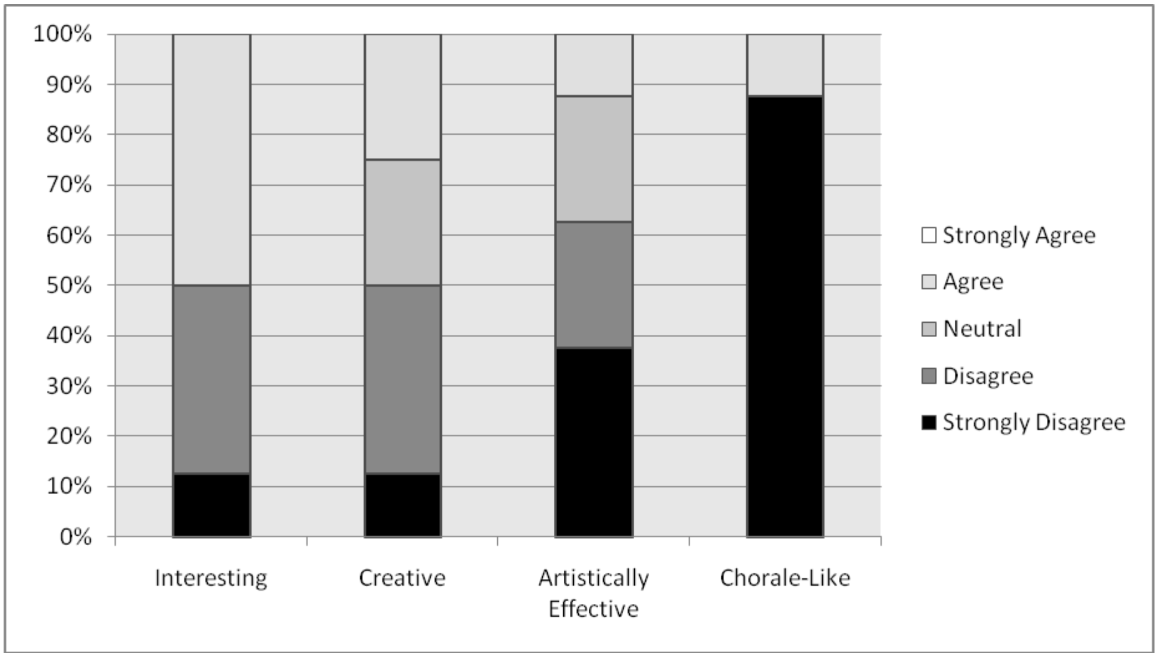
- Same comments as *[Song No. 1]*

- Opening motive is answered in bass.
- Bass line at end!

What, if anything, were its shortcomings?

- No shape or direction
- Same as *[song 1]*
- Same comments as *[Song No. 1]*
- More of the same comments from *[number 4]*
- Lacks coherence throughout.

Song 3



What, if anything, was memorable about it?

- Layered linear "harmony"
- The final cadence almost seemed tonal; thanks!
- Reminded me of *[song 3]*

- Same comments as *[Song No. 1]*

- Opening motive is answered in bass. Similar to *[#3, 5, 7, 8]*, but ending suggests tonality.

What, if anything, were its shortcomings?

- Lacks forward motion and obvious progression

- Random, and the ending especially seems really out of character. The tonal resolution just doesn't fit – it sounds cheesy after what came before.

- Same as *[song 1]*

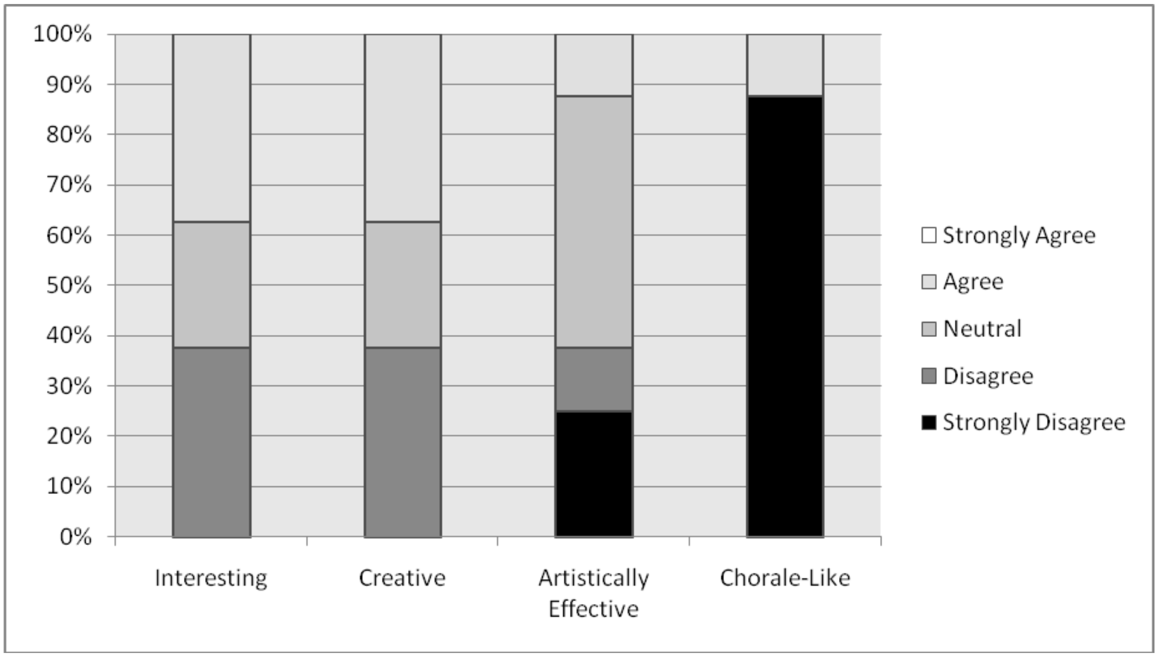
- Same comments as *[Song No. 1]*

- Seemed to start out somewhere but never really moved in a particular direction

- Lacks coherence throughout.

- Lacks energy 1/3 way through, then at end

Song 4



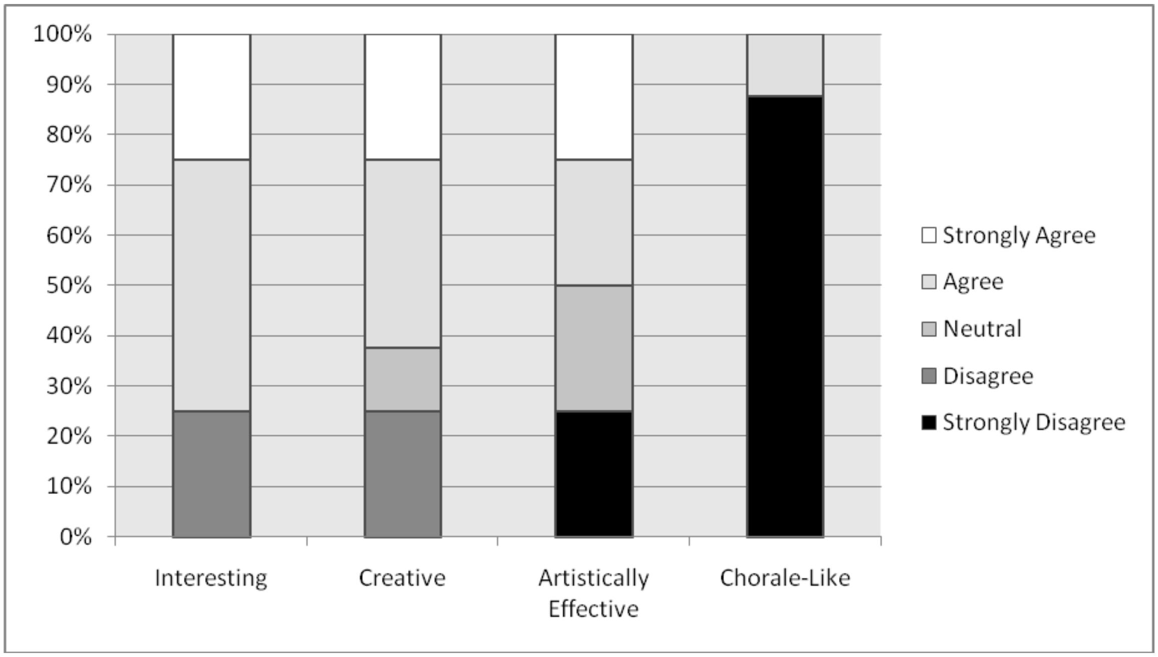
What, if anything, was memorable about it?

- A little more consonant than *[1 – 8]*
- Mostly, that it could become a continuation of *[Song 1]*
- The rhythms suggest a sense of organization, but it isn’t quite there.
- Nothing
- Same comments as *[Song No. 1]*
- Opening motive is answered in bass.

What, if anything, were its shortcomings?

- No direction or forward phrase motion
- See comments for *[Song 1]*; the single ultimate pitch made the cadence even more stable
- Same as *[song 1]*
- Same comments as *[Song No. 1]*
- seemed more random
- Lacks coherence throughout

Song 5



What, if anything, was memorable about it?

- A little more consonant than *[1 – 8]*
- Both in structure and cadence it is atonal. Its counterpoint makes the many 2nds and 7ths listenable
- It has a bit of a sense of evolution and growth, but not much.
- Nothing
- Same comments as *[Song No. 1]*
- Rhythmic continuity. Harmonic continuity – following a perceivable logic and course.
- Opening motive is answered in bass.
- Good placement of low C pedal

What, if anything, were its shortcomings?

- No direction or forward phrase motion

- Most objectives of classical (tonal) composition (form through phrase construction, cadences, etc) were not attempted here.

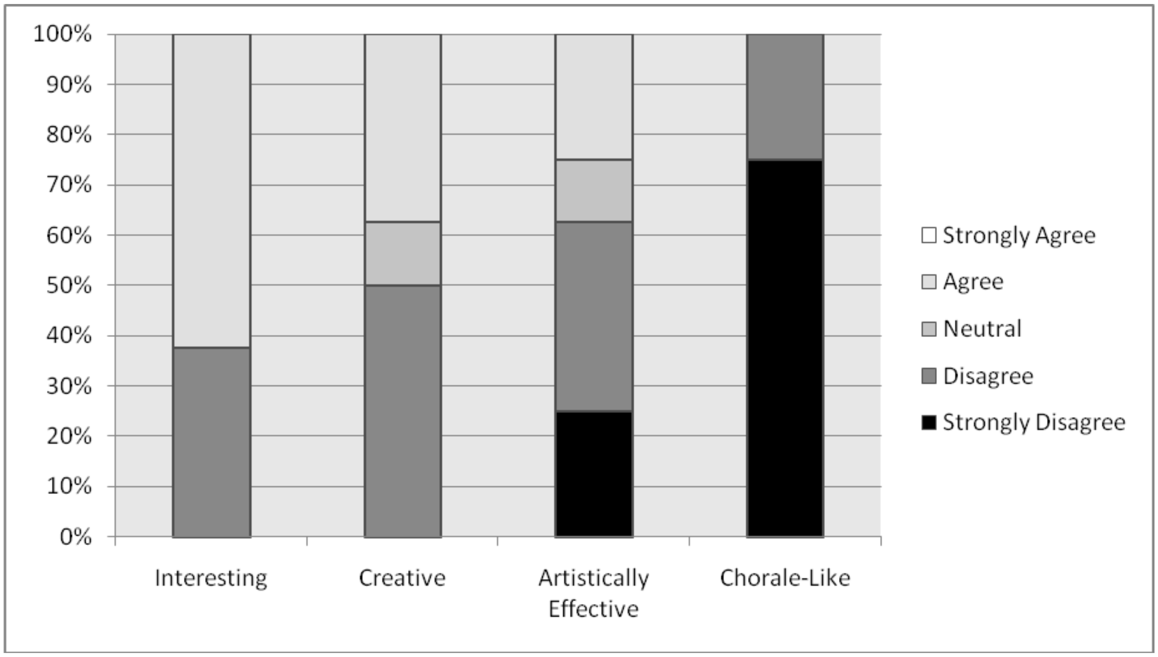
- Same comments as *[Song No. 1]*

- Lacks coherence throughout.

Appendix P

Reviews of Test GA Music by Composers

Song 6



What, if anything, was memorable about it?

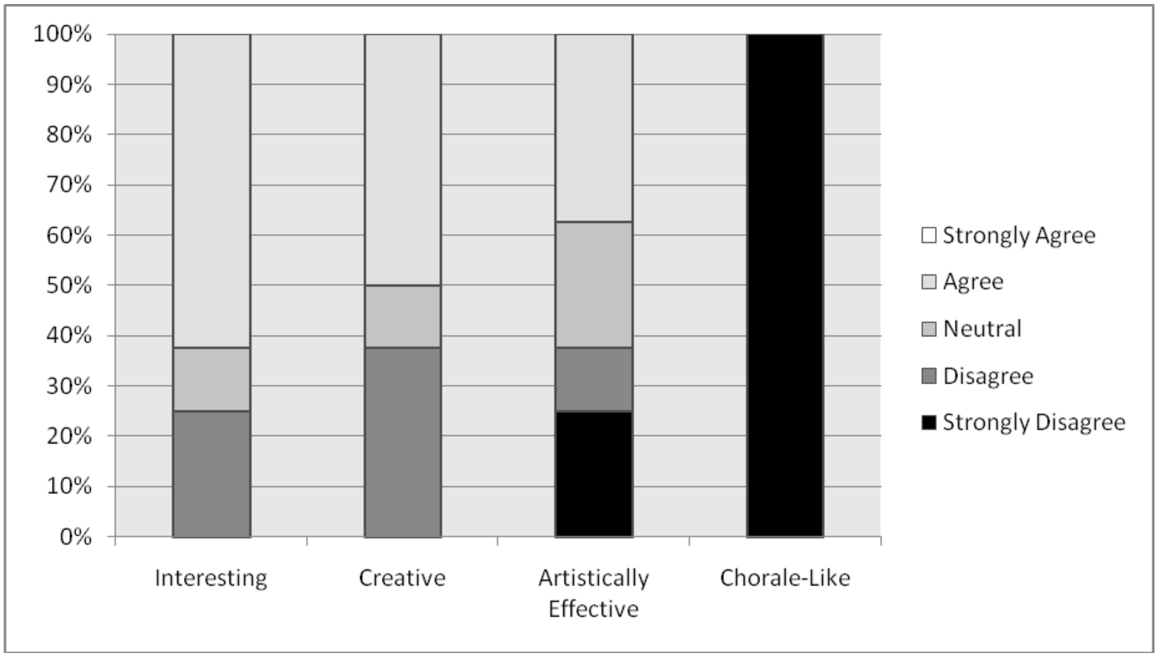
- Layered dissonance
- The single note (somewhat melodic) reiteration is more conjunct, thus is slightly more chorale-like
- Opening gesture reminded me of *[song 1]*.
- Same comments as *[Song No. 1]*
- harmonic continuity
- Like *[#1]* and *[#2]*, it is atonal, contrapuntal. Starts with conversation between treble and bass. Ending 3d is nice.

What, if anything, were its shortcomings?

- No phrase shape

- No sense of unity or direction
- Same as *[song 1]*.
- Same comments as *[Song No. 1]*
- Like *[#1]* and *[#2]*, it lacks unity that could have been created by better development of melodic or rhythmic motives.

Song 7



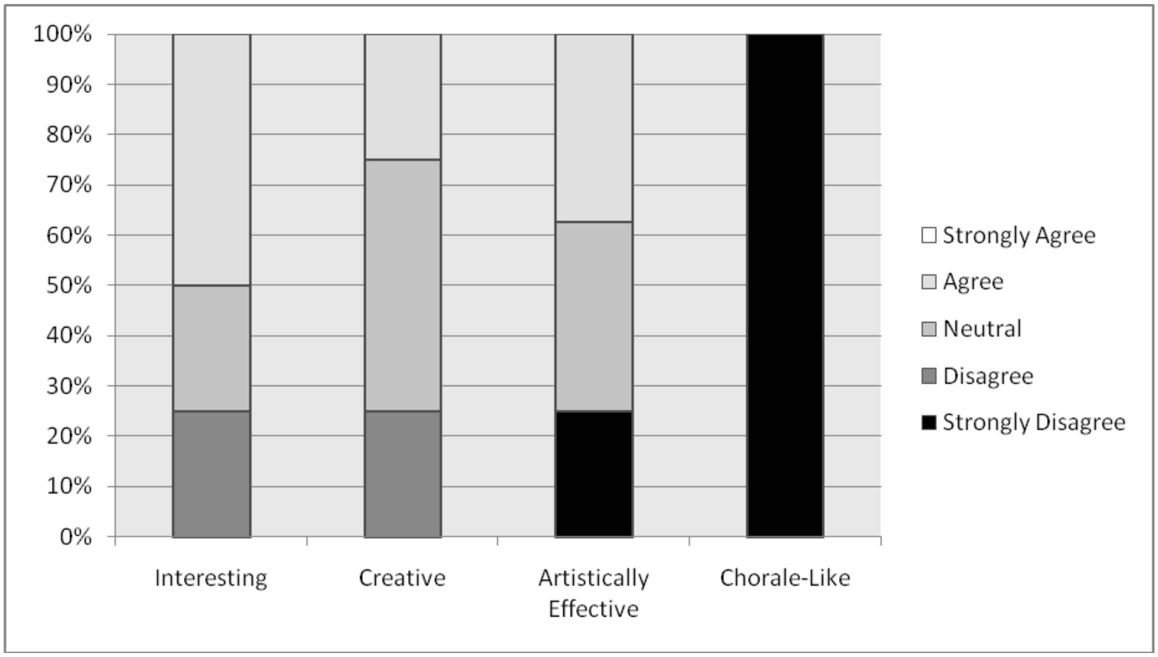
What, if anything, was memorable about it?

- Feels rhythmic with multi-voiced lines
- By this time I am visualizing the movie scenes the Songs portray.
- Nothing
- Same comments as *[Song No. 1]*
- The harmonic sense seemed to dissipate about halfway through
- Atonal, contrapuntal. Starts with conversation between treble and bass.

What, if anything, were its shortcomings?

- Emphasized dissonance. No obvious phrasing or shape.
- It seems quite random.
- This seems more like raw material that can be sculpted into something "creative" or "artistically effective." The lack of nuanced articulation, dynamics, and the perpetually depressed sustain pedal its most noticeable shortcomings.
- Same comments as *[Song No. 1]*
- Opening motive could have been better manipulated for more unity.

Song 8



What, if anything, was memorable about it?

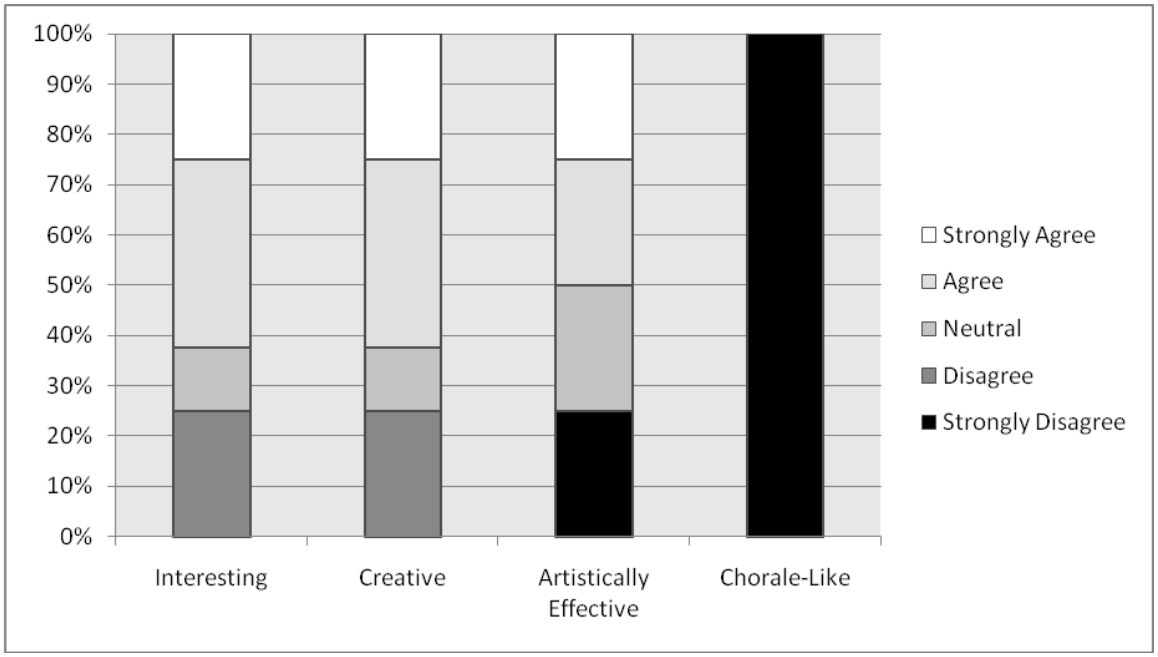
- Layered rhythmic motifs.
- Its seemingly thicker texture had me listening more closely for inferred chord progression
- The implication of tonality in the low voice helps this one. The back-and-forth of high to low voicing gives it a sense of organization.

- The material reminded me of *[song 1]*.
- Same comments as *[Song No. 1]*
- Atonal, contrapuntal. Starts with conversation between treble and bass.

What, if anything, were its shortcomings?

- The sustained sounds made it too dissonant to be at all pleasant. Very little discernable phrasing.
- I couldn’t count; were all chromatic pitches used equally?
- Same as *[song 1]*
- Same comments as *[Song No. 1]*
- Opening motive could have been better manipulated for more unity.

Song 9



What, if anything, was memorable about it?

- Feels rhythmic with multi-voiced lines

- Only that other Songs seem to use the same creative formulae

- This one has the tension-and-release that we look for in most types of music. That makes it feel like it's going somewhere.

- Reminded me of *[song 1]*

- The melodic material seems interesting and with the "right" harmonies, whether dissonant or consonant, has some interesting potential.

- Harmonic and rhythmic continuity. Good emotional quality, a sense of angst.

- Like *[#1], [#2]*, and *[#4]*, atonal and contrapuntal. Starts with conversation between treble and bass.

- The low D comes in at a nice place

What, if anything, were its shortcomings?

- The sustained sounds made it too dissonant to be at all pleasant. Very little discernable phrasing.

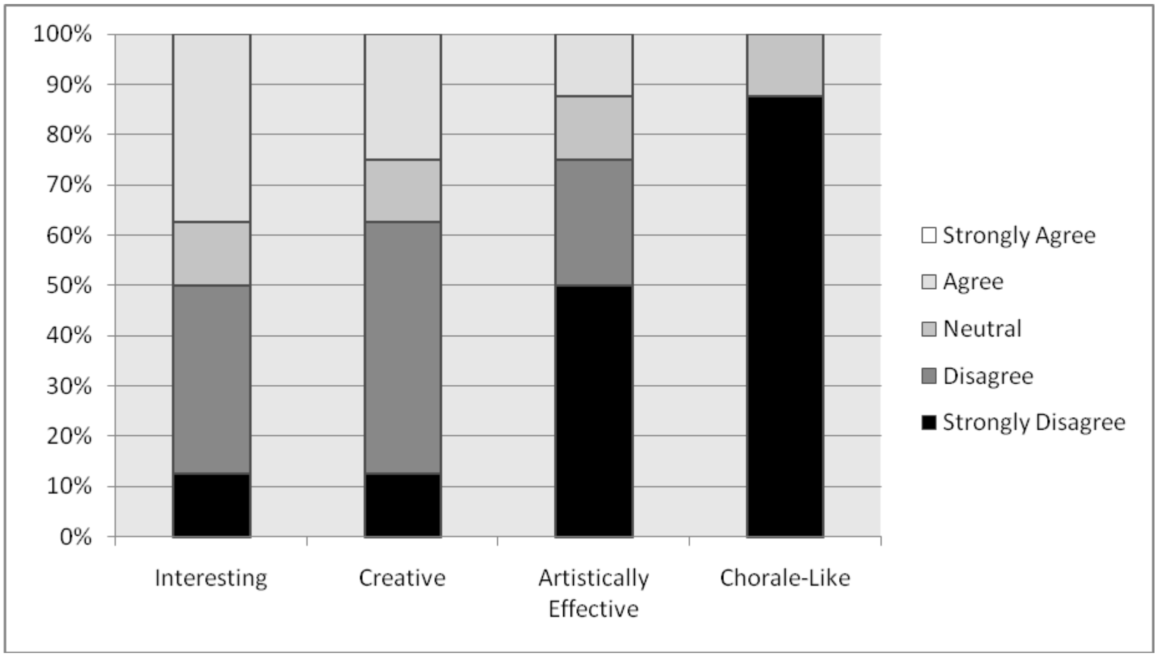
- By this time, the similarity of the Songs is making them less artistically effective

- Same as *[song 1]*

- The midi-like playback aside, it reminds me of a composition student who wants to stretch his or her use of harmony but doesn't really understand how. It uses dissonant harmonies just for the sake of dissonance. There's no apparent logic to it, it seems random in its "progression". Less pedal would possibly help.

- Opening motive could have been better manipulated for more unity.

Song 10



What, if anything, was memorable about it?

- Feels layered
- The reiteration (conjunct) was lost again; the cadential chord was interesting
- The pauses give it a better sense of phrasing than most of the pieces.
- Opening reminded me of *[song 1]*
- Same comments as *[Song No. 1]*
- Like *[#1, 2, 4 and 6]*, atonal, contrapuntal. Starts with conversation between treble and bass. Ending 3d is nice.

What, if anything, were its shortcomings?

- No shape or direction
- feels quite random
- Same as *[song 1]*
- Same comments as *[Song No. 1]*

- The randomness of the harmonic center leads me down no particular path. I feel that I’m left unattended.

- Opening motive could have been better manipulated for more unity.

Appendix Q

Overall Comments from Reviewers

- The music tied together. If each clip was wound together I believe it would tell a story. A story about a mystery or tragic time in someones life. Also could see some of this music used in a RPG or a mystery movie.

- The music might be more effective if all lines moved, like a Count Basie sax section; or if the tempo varied, or with a database of chords (while leaving the inversion/spelling of those chords up to the program) giving added weight to the probability of the root going up a 4th, down a 5th, or down a minor 2nd.

- There were obviously 2 themes and each had a successful version to my ears. However, the absence of variation in tempo bothered me. They lacked emotion and feeling.

- I am impressed by the fact that these selections were AI composed

- Some seemed well written, while others felt as if they were just notes strung out on a page. I would like to hear a longer selection, as a couple of them had a good hook but ended before I got fully engaged.

- The sounds are great but too short to grasp the rhythm

- Overall, I feel like these are a decent sampling of pieces. They are all fairly dark, but most have interesting bits and pieces. Several combine those well, others don’t. I’d say that something similar could be written by a student taking their first composition course and perhaps only having made it through the first half of the course, learning the basics, and some of the rules of “how things should sound”, but missing out on all the really important nuances of what is pleasing to the ear. Granted, this is all very subjective, as I tend to find darker, experimental music extremely satisfying, while others might not. Most people also tend to not prefer songs set in a minor key, but those are among my favorite. Overall, I liked most of these, and would welcome the opportunity to hear longer pieces

Appendix R

Overall Comments from Composers

Do you have any overall comments about the selections?

- These selections have characteristics similar to MIDI realizations of music written by humans. The primary problem with these melody fragments AND MIDI realizations of all types is the lack of forward motion and phrase progression. GOOD human performances will always "lead" the listener toward the end of the phrase and will leave the melody with a sense of completion. This sense was missing. Whether a human performer could have provided this type of forward motion and phrasing in THIS music is still an unanswered question. I don't feel that these renditions had much, if any, motion. And the shape of the phrase was not readily evident.

- I am obsessively partial to the harmonic overtone series and the construction of music suggested by its structure. Unfortunately that makes me partial to tonal music and therefore, all of its impact on pitch, duration, timbre and intensity common to music composition. And since 'this music is chorale-like' was an evaluation criterion, all the rules of voice leading would also come into play. Little in the Songs seemed to embrace these elements of music.
  I do enjoy atonal music for the intellectual lengths to which it goes to disallow a single pitch being a 'tonic.' AI music, as represented here, seems to embrace this approach.
  If I had another lifetime unencumbered by the vicissitudes of daily life, I would think it possible to personally invite AI into the 'HOS' and allow AI to mathematically proceed through an evolution of relationships to the musical elements (and the forms, melodies and harmonies that have come to us). It would be amusing to see if in several weeks AI would have traveled a course similarly to what has occurred in millennia of western music history.
  Thanks, it's been fun!

- With such short excerpts, I found myself drawn to the pieces that had more suggestion of organization, and this included the presence of tonal harmony or harmonic implications. Any of these pieces – including the ones that seem completely pointless in these truncated forms - have the potential to be interesting and artistically satisfying if developed into longer works.
  The lack of expression in the playback is a hurdle to accepting this music. If interpreted with nuance and expression all of these pieces would fare much better.
  I'm puzzled by the question of whether the pieces are "chorale-like". There's nothing remotely suggestive of a chorale in any of them. Perhaps if they were slowed down and played on a sound with more sustain, but even then I don't think "chorale" would really be the effect.

- Even though I was unable to articulate precisely what at the time of listening, clearly something was memorable about the first few selections due to the fact that I recognized them later. My response on the shortcomings on the first song is a fair representation of my overall thoughts on the selections.

- I'm sorry that my answers turned out to be the same for each selection. What I said in response to the first selection is true for all of the selections. There's potential of course, but there doesn't seem to be any logic behind the use of harmonies. I wasn't looking for beauty per se, but something that was musical. The midi-like stiffness will always take away some of the musicality, but as I stated before the lack of harmonic logic seems to me to be the real downfall. This is not to say that it can't work and I have no clue how you teach a technology the concepts of consonance and dissonance or how you give it suggestions on the use of extended harmonies, but it does create some interesting and exciting possibilities. I think the pieces are strongest melodically even with similarities between each tune. To me very typical of a composition student (AI or not) trying to find his/her/its way in an extended harmonic landscape.

- I think there is an influencing effect of familiarity due to listening to these as a group, because they start to seem like variations on a theme, and I think there may be an inherent preference for things that sound familiar since there is a unifying logic among the samples. I think this is influencing the grading of them.

- Each example begins with a promise of logic and coherence but does not follow through. I miss the human input.

- I was surprised how similar they, on the surface, were. Hard to contrast. Small, detailed differences.
  I broke them into two sets of 5 based on the opening motive. Within each set I then rated them and used those five positions as my answers for Questions A. B. C.
  None of the examples were what I would consider chorale-like.
  For me the placement of the initial low note was critical in setting up some sense of a note.
  Also a continuous sounding of notes (approx 16th notes) seemed more effective. This holds true in Bach.

References

Alonso, O., Rose, D.E., and Stewart, B. (2008). Crowdsourcing for Relevance Evaluation. *SIGIR Forum 42*(2) 9-15.

Amazon.com, Inc. (2010). *Amazon Mechanical Turk*. Retrieved September 12, 2011 from https://www.mturk.com/mturk/welcome.

Amazon.com, Inc. (2010). *Best Practices Guide* Retrieved February 12, 2010, from http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf

Amazon.com, Inc. (2010). *Requester User Interface Guide*. Retrieved September 12, 2011 from http://s3.amazonaws.com/awsdocs/MechTurk/latest/MTURK-UI.pdf

Amazon.com, Inc. (2010). *Technical Documentation: Amazon Mechanical Turk (API Version: 2008-08-02) Beta*. Retrieved February 12, 2010, from http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1852&categoryID=28

Bhagwan, V., Grandison, T., & Gruhl, D. (2009). Sound index: Charts for the people, by the people. *Communications of the ACM*, *52*(9), 64-70.

Biles, J. (1994). GenJam: A genetic algorithm for generating jazz solos. Paper presented at the Proceedings of the *International Computer Music Conference (ICMC'94),* Aarhus, Denmark. 131-137.

Biles, J. (2007). Evolutionary computation for musical tasks. In Miranda, E. R. and Biles, J. A., editors, *Evolutionary Computer Music*, chapter 2, pages 28–51. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Birchfield, D. (2003). Generative model for the creation of musical emotion, meaning, and form. Paper presented at the *ETP '03: Proceedings of the 2003 ACM SIGMM Workshop on Experiential Telepresence*, Berkeley, California. 99-104.

BootBe, Inc. (2011). *World's Best Creative Department*. Retrieved August 5, 2011 from http://www.bootb.com/en/

Carnegie Mellon University (2010). *The Official CAPTCHA site*. Retrieved July 26, 2011 from http://www.captcha.net/

Carvalho, V. R., Lease, M., & Yilmaz, E. (2011). Crowdsourcing for search evaluation. *SIGIR Forum*, *44*(2), 17-22.

CastingWords (2010). *Audio transcription services: plus video, CDs, podcasts and more.* Retrieved February 24, 2010 from http://castingwords.com/

Celma, Ò., & Lamere, P. (2008). If you like the Beatles you might like...: A tutorial on music recommendation. Paper presented at the *MM '08: Proceeding of the 16th ACM International Conference on Multimedia,* Vancouver, British Columbia, Canada. 1157-1158.

Chen, Y. (2007). Interactive music composition with the CFE framework. *SIGEVOlution, 2*(1), 9-16.

Chung, K., Chiu, C., Xiao, X., & Chi, P. (2009). Stress outsourced: A haptic social network via crowdsourcing. Paper presented at the *CHI EA '09: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, Boston, MA, USA. 2439-2448.

CloudCrowd (2009). *CloudCrowd: We're working on it. Lots of us.* Retrieved August 5, 2011 from http://www.cloudcrowd.com/home.

Composers Forum (2011). *American Composers Forum Members.* Retrieved July 24, 2011 from http://www.composersforum.org/artists_membersearch_browse.cfm?state=Tennessee&x=14&y=11.

Craane, J. (2009). *Jamie Craane's Blog: Introduction to Genetic Algorithms with JGAP*. Retrieved September 15, 2009 from http://jcraane.blogspot.com/2009/02/introduction-to-genetic-algorithms-with.html.

Craane, J. (2009). *Melodycomposition: application for melody composition using genetic algorithms*. Retrieved September 15, 2009 from http://code.google.com/p/melodycomposition/.

CrowdFlower (2011). *CrowdFlower: Enterprise Crowdsourcing Solutions.* Retrieved August 5, 2011 from http://crowdflower.com/.

CrowdSPRING (2011). *CrowdSPRING: The world's #1 marketplace for logos and graphic design.* Retrieved August 5, 2011 from http://www.crowdspring.com/.

DarwinTunes (2010). *Let's evolve music*. Retrieved May 10, 2010 from http://darwintunes.org/evolve-music.

DarwinTunes (2010). *The population memetics of DarwinTunes.* Retrieved April 16, 2010 from http://darwintunes.org/sites/default/files/The_population_memetics_of_DarwinTunes.pdf.

DarwinTunes (2010). *The technology behind DarwinTunes.* Retrieved April 16, 2010 from http://darwintunes.org/sites/default/files/The_technology_behind_DarwinTunes.pdf

DarwinTunes (2010). *Welcome to DarwinTunes.* Retrieved April 16, 2010 from http://darwintunes.org/.

DarwinTunes (2010). Audio Snapshots. Retrieved August 16, 2011 from http://darwintunes.org/audio-snapshots.

de Freitas, A. R. R., & Guimarães, F. G. (2011). Originality and diversity in the artificial evolution of melodies. Paper presented at the *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation,* Dublin, Ireland. 419-426.

De Prisco, R., Zaccagnino, G., & Zaccagnino, R. (2010). EvoBassComposer: A multi-objective genetic algorithm for 4-voice compositions. Paper presented at the *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation,* Portland, Oregon, USA. 817-818.

Dictionary.com (2011). *Music Theory | Define Music Theory at Dictionary.com.* Retrieved July 28, 2011 from http://dictionary.reference.com/browse/music+theory.

Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., & Stuckenschmidt, H. (2010). Crowdsourcing the assembly of concept hierarchies. Paper presented at the *Proceedings of the 10th Annual Joint Conference on Digital Libraries,* Gold Coast, Queensland, Australia. 139-148.

Encyclopedia Britannica (2011). *Genetic drift.* Retrieved July 28, 2011 from http://www.britannica.com/EBchecked/topic/228886/genetic-drift.

Evolectronica (2011). *Channel 1*. Retrieved February 12, 2011 from http://evolectronica.com/channel1.

Evolectronica (2011). *Evolectronica news*. Retrieved February 12, 2011 from http://evolectronica.com/news.

Evolectronica (2011). *Press: information, background and contacts.* Retrieved February 12, 2011 from http://evolectronica.com/press-info.

fellowforce (2007). *World's Portal to Open Innovation*. Retrieved August 6, 2011 from http://www.fellowforce.com/.

Fu, T., Wu, T., Chen, C., Wu, K., & Chen, Y. (2006). Evolutionary interactive music composition. Paper presented at the *GECCO '06: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation,* Seattle, Washington, USA. 1863-1864.

Gabrani, G., Bhargava, P., Bhawana, B., & Gill, G. S. Use of genetic algorithms for indian music mixing. *Ubiquity, 2008* (March), 1-10.

Ganjisaffar, Y., Javanmardi, S., & Lopes, C. (2009). Leveraging crowdsourcing heuristics to improve search in wikipedia. Paper presented at the *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration,* Orlando, Florida. 1-2.

Gartland-Jones, A., & Copley, P. (2003). The suitability of genetic algorithms for musical composition. *Contemporary Music Review, 22*(3), 43-55.

GAVAB Research Group (2007). *Spieldose: A genetic music composition tool*. Retrieved April 30, 2009, from http://www.gavab.es/recursos_en.html.

Geisler, G., Willard, G., & Whitworth, E. (2010). Crowdsourcing the indexing of film and television media. Paper presented at the *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47,* Pittsburgh, Pennsylvania. 82:1-82:10.

Gill, S. (1963).A Technique for the Composition of Music in a Computer. *The Computer Journal, 6*(2), 129-133.

Harper, F. M., Raban, D., Rafaeli, S. & Konstan, J. A. (2008). Predictors of answer quality in online Q & A sites. Paper presented at the *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems,* Florence, Italy. 865-874.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. Paper presented at the *Proceedings of the 28th International Conference on Human Factors in Computing Systems,* Atlanta, Georgia, USA. 203-212.

Hintikka, K. A. (2008). Web 2.0 and the collective intelligence. Paper presented at the *MindTrek '08: Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, Tampere, Finland. 163-166.

Horton, J. J., & Chilton, L. B. (2010). The labor economics of paid crowdsourcing. Paper presented at the Proceedings of the *11th ACM Conference on Electronic Commerce,* Cambridge, Massachusetts, USA. 209-218.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired, 14*(6). Retrieved June 14, 2009 from http://www.wired.com/wired/archive/14.06/crowds.html.

Hsueh, P., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. Paper presented at the *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing,* Boulder, Colorado. 27-35.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception,* Vol. 19, No. 1, pp. 1-64.

InnoCentive (2011). *Welcome to InnoCentive: Where the World Innovates.* Retrieved August 5, 2011 from http://www.innocentive.com/.

Jacob, B. (2009). *Algorithmic Composition*. Retrieved April 22, 2009, from http://www.ece.umd.edu/~blj/algorithmic_composition/.

Jagadeesan, A. P., Lynn, A., Corney, J. R., Yan, X. T., Wenzel, J., Sherlock, A., et al. (2009). Geometric reasoning via internet CrowdSourcing. Paper presented at the *SPM '09: 2009 SIAM/ACM Joint Conference on Geometric and Physical Modeling,* San Francisco, California. 313-318.

Jensen, J. H., & Haddow, P. C. (2011). Evolutionary music composition based on zipf's law. Paper presented at the *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation,* Dublin, Ireland.

Khalifa, Y., & Al-Mourad, M. B. (2006). Autonomous evolutionary music composer. Paper presented at the *GECCO '06: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation,* Seattle, Washington, USA. 1873-1874.

Khalifa, Y. M. A., Khan, B. K., Begovic, J., Wisdom, A., & Wheeler, A. M. (2007). Evolutionary music composer integrating formal grammar. Paper presented at the *GECCO '07: Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation*, London, United Kingdom. 2519-2526.

Kittur, A., Chi, ,E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. Paper presented at the *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems,* Florence, Italy. 453-456.

Kirke, A., & Miranda, E. R. (2009). A survey of computer systems for expressive music performance. *ACM Computing Surveys, 42*(1), 1-41.

Koblin, A. M. (2009). The sheep market. Paper presented at the *C&C '09: Proceeding of the Seventh ACM Conference on Creativity and Cognition,* Berkeley, California, USA. 451-452.

Krieger, M., Stark, E. M., & Klemmer, S. R. (2009). Coordinating tasks on the commons: Designing for personal goals, expertise and serendipity. Paper presented at the *CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems,* Boston, MA, USA. 1485-1494.

Last.fm (2009) *Listen to free music with internet radio and the largest music catalogue online*. Retrieved May 6, 2009, from http://www.last.fm/.

Lease, M., Carvalho, V. R., & Yilmaz, E. (2011). Crowdsourcing for search and data mining. *SIGIR Forum, 45*(1), 18-24.

Ledlie, J., Odero, B., Minkov, E., Kiss, I., & Polifroni, J. (2009). Crowd translator: On building localized speech recognizers through micropayments. *SIGOPS Operating Systems Review, 43*(4), 84-89.

Legaspi, R., Hashimoto, Y., & Numao, M. (2006). An emotion-driven musical piece generator for a constructive adaptive user interface, Paper presented at the *9th Pacific Rim International Conference on Artificial Intelligence*, 890-894.

Legaspi, R., Hashimoto, Y., Moriyama, K., Kurihara, S., & Numao, M. (2007). Music compositional intelligence with an affective flavor. Paper presented at the *IUI '07: Proceedings of the 12th International Conference on Intelligent User Interfaces,* Honolulu, Hawaii, USA. 216-224.

Levisohn, A., & Pasquier, P. (2008). BeatBender: Subsumption architecture for autonomous rhythm generation. Paper presented at the *ACE '08: Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology,* Yokohama, Japan. 51-58.

Li, C., Buyuktur, A. G., Hutchful, D. K., Sant, N. B., & Nainwal, S. K. (2008). Portalis: Using competitive online interactions to support aid initiatives for the homeless. Paper presented at the *CHI '08: CHI '08 Extended Abstracts on Human Factors in Computing Systems,* Florence, Italy. 3873-3878.

Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2009). TurKit: Tools for iterative tasks on mechanical turk. Paper presented at the *Proceedings of the ACM SIGKDD Workshop on Human Computation*, Paris, France. 29-30.

Liu, Y., Lehdonvirta, V., Kleppe, M., Alexandrova, T., Kimura, H., & Nakajima, T. (2010). A crowdsourcing based mobile image translation and knowledge sharing service. Paper presented at the *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, Limassol, Cyprus. 6:1-6:9.

Mannes, A. E. (2009). Are we wise about the wisdom of crowds? the use of group judgments in belief revision. *Management Science, 55*(8), 1267-1279.

Mason, W., & Watts, D. J. (2009). Financial incentives and the performance of crowds. Paper presented at the HCOMP '09: Proceedings of the *ACM SIGKDD Workshop on Human Computation,* Paris, France. 77-85.

Mathews, M.V. (1963). The Digital Computer as a Musical Instrument. *Science, 142*(3592), 553-557.

McDermott, J., O'Neill, M., & Griffith, N. J. L. (2010). Interactive EC control of synthesized timbre. *Evolutionary Computing 18*(2), 277-303.

Meehan, J. R. (1979). An artificial intelligence approach to tonal music theory. Paper presented at the *ACM 79: Proceedings of the 1979 Annual Conference*, 116-120.

Meffert, K. & Rotstan, N. (2009). *JGAP: Java Genetic Algorithms Package*. Retrieved September 16, 2009 from http://jgap.sourceforge.net/.

Merriam-Webster (2001). *Genre – Definition and More from the Free Merriam-Webster Dictionary.* Retrieved July 28, 2011 from http://www.merriam-webster.com/dictionary/genre.

Miletto, E. M., Flores, L. V., Pimenta, M. S., Rutily, J., & Santagada, L. (2007). Interfaces for musical activities and interfaces for musicians are not the same: The case for codes, a web-based environment for cooperative music prototyping. Paper presented at the *ICMI '07: Proceedings of the 9th International Conference on Multimodal Interfaces,* Nagoya, Aichi, Japan. 201-207.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review, 63*, 81-97.

Milne, D., & Witten, I. H. (2008). Learning to link with wikipedia. Paper presented at the *CIKM '08: Proceedings of the 17th ACM Conference on Information and Knowledge Management,* Napa Valley, California, USA. 509-518.

Miranda, E. R. (2004). At the crossroads of evolutionary computation and music: Self-programming synthesizers, swarm orchestras and the origins of melody. *Evolutionary Computation 12*(2), 137-158.

Miranda, E., & Todd, P. M. (2007). Computational Evolutionary Musicology. Evolutionary Computer Music, 218-249.

Nelson, G.L. (1993). Sonomorphs: An application of genetic algorithms to growth and development of musical organisms. Paper presented at the *4th Biennial Art and Technology Symposium,* New London, Connecticut. 155-169

Nelson, G.L. (1995). Further adventures of the Sonomorphs. Paper presented at the *Fifth Biennial Art & Technology Symposium,* New London, Connecticut. 51-64

Nichols, E., Morris, D., & Basu, S. (2009). Data-driven exploration of musical chord sequences. Paper presented at the *IUI '09: Proceedings of the 13th International Conference on Intelligent User Interfaces,* Sanibel Island, Florida, USA. 227-236.

Ning, C., & Zhou, S. (2010). The music pattern: A creative tabletop music creation platform. *Computers in Entertainment, 8*(2), 13:1-13:15.

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. Paper presented at the *Proceedings of the International Conference on Multimedia Information Retrieval*, Philadelphia, Pennsylvania, USA. 557-566.

Numao, M., Takagi, S., & Nakamura, K. (2002). Constructive adaptive user interfaces: Composing music based on human feelings. Paper presented at the *Eighteenth National Conference on Artificial Intelligence, Edmonton, Alberta, Canada*. 193-198.

Oliwa, T. M. (2008). Genetic algorithms and the abc music notation language for rock music composition. Paper presented at the *GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation,* Atlanta, GA, USA. 1603-1610.

Pandora Radio (2009). *Listen to Free Internet Radio, Find New Music*. Retrieved May 6, 2009, from http://www.pandora.com/.

Reis, G., & Vega, F. F. (2007). Electronic synthesis using genetic algorithms for automatic music transcription. Paper presented at the *GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation,* London, England. 1959-1966.

Roads, C. (1985). Research in music and artificial intelligence. *ACM Computing Surveys, 17*(2), 163-190.

Roman, D. (2009). Crowdsourcing and the question of expertise. *Communications of the ACM, 52*(12), 12-12.

Sadie, S. (Ed.). (1988). *The Norton/Grove Concise Encyclopedia of Music*. W. W. Norton & Company.

Sánchez, A., Pantrigo, J., Virseda, J., & Pérez, G. (2007). Spieldose: An interactive genetic software for assisting to music composition tasks. Paper presented at the *IWINAC '07: Proceedings of the 2nd International Work-Conference on the Interplay between Natural and Artificial Computation, Part I,* La Manga del Mar Menor, Spain. 617-626.

Schoenberger, J. (2009). *Musical Composition with Genetic Algorithms with Coherency Through Genotype*. Retrieved September 15, 2009, from http://www.mamageekminis.com/joy/career/research.html.

Seay, A. (1964).The Composer of Music and the Computer, *Computers and Automation*, *13*(8), 16-18.

Sharp, H., Rogers, Y. and Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction.* John Wiley and Sons.

Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2009). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th edition).* Addison-Wesley.

Spotify (2010). *Spotify – A world of music*. Retrieved May 10, 2010 from http://www.spotify.com/int/

Stewart, O., Huerta, J. M., & Sader, M. (2009). Designing crowdsourcing community for the enterprise. Paper presented at the *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation,* Paris, France. 50-53.

Stewart, O., Lubensky, D., & Huerta, J. M. (2010). Crowdsourcing participation inequality: A SCOUT model for the enterprise domain. Paper presented at the *Proceedings of the ACM SIGKDD Workshop on Human Computation*, Washington DC. 30-33.

Stolee, K. T., & Elbaum, S. (2010). Exploring the use of crowdsourcing to support empirical studies in software engineering. Paper presented at the *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement,* Bolzano-Bozen, Italy. 35:1-35:4.

Su, Q., Pavlov, D., Chow, J., & Baker, W. C. (2007). Internet-scale collection of human-reviewed data. Paper presented at the *WWW '07: Proceedings of the 16th International Conference on World Wide Web,* Banff, Alberta, Canada. 231-240.

Sugimoto, T., Legaspi, R., Ota, A., Moriyama, K., Kurihara, S., & Numao, M. (2008). Modelling affective-based music compositional intelligence with the aid of ANS analyses. *Knowledge-Based Systems, 21*(3), 200-208.

Surowiecki, J. (2005). *The wisdom of crowds.* Anchor.

Tanaka, A., Tokui, N., & Momeni, A. (2005). Facilitating collective musical creativity. Paper presented at the *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*, Hilton, Singapore. 191-198.

Tokui, N., & Iba, H. (2000). Music composition with interactive evolutionary computation. Paper presented at the *GA2000: Proceedings of the third International Conference on Generative Art*. 227-251.

Twitter (2010). *Darwintunes*. Retrieved February 8, 2011 from http://twitter.com/darwintunes.

Unehara, M., & Onisawa, T. (2003). Construction of music composition system with interactive genetic algorithm. Paper presented at the *Proceedings of the 6th Asian Design International Conference,* Tsukuba Japan. 84-89.

Unemi, T. (2003). SBEAT3: A tool for multi-part music composition by simulated breeding. Paper presented at the *ICAL 2003: Proceedings of the Eighth International Conference on Artificial Life,* Sydney, Australia. 410-413.

Urbano, J., Morato, J., Marrero, M. & Martin, D. (2010). Crowdsourcing preference judgments for evaluation of music similarity tasks. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010),* Geneva, Switzerland. 9 - 16

Von Ahn L., & Dabbish, L. (2004). Labeling images with a computer game. Paper presented at the CHI '04: Proceedings of the *SIGCHI Conference on Human Factors in Computing Systems,* Vienna, Austria. 319-326.

Von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. Paper presented at the *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Montréal, Québec, Canada. 55-64.

Wikipedia (2011). *Amazon Mechanical Turk*. Retrieved July 28, 2011 from http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk.

Wilogo.com (2011). *Logo Design by Wilogo*. Retrieved August 6, 2011 from http://en.wilogo.com/.

Xia, M., Huang, Y., Duan, W., & Whinston, A. B. (2009). Ballot box communication in online communities. *Communications of the ACM*, *52*(9), 138-142.

Xu, A., & Bailey, B. P. (2011). A crowdsourcing model for receiving design critique. Paper presented at the *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems*, Vancouver, BC, Canada. 1183-1188.

Yan, T., Marzilli, M., Holmes, R., Ganesan, D., & Corner, M. (2009). mCrowd: A platform for mobile crowdsourcing. Paper presented at the *SenSys '09: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, Berkeley, California. 347-348.

Yang, J., Adamic, L. A., and Ackerman, M. S. (2008). Crowdsourcing and Knowledge Sharing: Strategic User Behavior on Taskcn. Paper presented at the *9th ACM conference on Electronic Commerce*, New York, NY, USA. 246-255.

Yee-King, M. (2000). Audio Serve - an online system to evolve modular audio synthesis circuits. Unpublished master’s thesis, University of Sussex, UK.