Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: https://www.tandfonline.com/loi/nnmr20

A framework for computer-assisted sound design systems supported by modelling affective and perceptual properties of soundscape

Miles Thorogood, Jianyu Fan & Philippe Pasquier

To cite this article: Miles Thorogood, Jianyu Fan & Philippe Pasquier (2019) A framework for computer-assisted sound design systems supported by modelling affective and perceptual properties of soundscape, Journal of New Music Research, 48:3, 264-280, DOI: 10.1080/09298215.2019.1612924

To link to this article: https://doi.org/10.1080/09298215.2019.1612924

JOURNAL OF NEW MUSIC RESEARCH
2019, VOL. 48, NO. 3, 264–280
https://doi.org/10.1080/09298215.2019.1612924

Routledge
Taylor & Francis Group

A framework for computer-assisted sound design systems supported by modelling affective and perceptual properties of soundscape

Miles Thorogood[a], Jianyu Fan[b] and Philippe Pasquier[b]

[a]University of British Columbia, Kelowna, BC, Canada; [b]Simon Fraser University, SIAT, Surrey, BC, Canada

ABSTRACT

Autonomously generating artificial soundscapes for video games, virtual reality, and sound art presents several non-trivial challenges. We outline a system called Audio Metaphor that is built upon the notion that sound design for soundscape compositions is emotionally informed. We first define the problem space of generating soundscape compositions referencing the sound design and soundscape literature. Next, we survey the state-of-the-art soundscape generation systems and establish the characteristics and challenges for evaluating these types of systems. We then describe the Audio Metaphor system that aims to model the soundscape generation problem using a method of soundscape emotion recognition and segmentation based on perceptual classes, and an autonomous mixing engine utilising optimisation and prediction algorithms to generate a soundscape composition. We evaluate the soundscape compositions generated by Audio Metaphor by comparing them with those created by a human expert and also those generated randomly. Our analysis of the evaluation study reveals that the proposed soundscape generation model is human-competitive regarding semantic and emotion-based indicators.

ARTICLE HISTORY
Received 16 November 2017
Accepted 5 April 2019

KEYWORDS
Audio analysis; machine learning; perception; emotion; composition; sound synthesis

1. Introduction

The field of sound design, which includes game sound, sound design for virtual reality (VR), and soundscape composition, develops alternative audio related solutions for sound-related problems. Soundscape composers aim at creating a type of electroacoustic music that Truax (1996) describes as 'characterised by the presence of recognisable environmental sounds and contexts, the purpose being to evoke listeners' associations, memories, and imagination related to the soundscape'. Although varying in intent, sound designers align with soundscape composition in terms of communicating an environment to the listener. For a video game, the sound designer aims to enhance the narrative experience by creating an animated set of sound effects corresponding to what is on the screen. A sound designer working on the sound design for a VR environment creates realistic soundscapes to increase a sense of place by simulating what would be heard if the user was in a real or imagined location (Serafin & Serafin, 2004).

Sound designers face numerous challenges posed by a growing number of sound files and the increased complexity of interactive environments, such as video games and virtual reality. Further, many of the tasks of the sound designer, such as producing draft sound designs in production environments, or searching and listening to sound files from a large database, can be repetitive, and providing tools for assisting in this field is valuable.

In order to unify the sound design practices described here, we will use the concept of a soundscape. The concept of a soundscape was first proposed by Schafer (1977), and framed in a communication model by Truax (2001) as a way of understanding the acoustic environment as a carrier of information. According to Truax (2012), soundscape composition has many possible applications that have distinct practices of representing the world. Looked at from a communication model, the different applications of sound design share the quality of communicating information to a listener about a real or imagined environment. Figure 1. demonstrates the relationships of different sound design contexts on the continuum moving from realistic to abstracted. The research here investigates computationally-assisted tools for sound design production, with focus on soundscapes trending toward the real end of the continuum, outlined on the figure in grey.

In creative practice, soundscape compositions can be generated by directly recording real-world environments,

CONTACT Miles Thorogood miles.thorogood@ubc.ca University of British Columbia, 1148 Research Rd., Kelowna, BC, Canada
This article was originally published with errors, which have now been corrected in the online version. Please see Correction (https://dx.doi.org/10.1080/09298215.2019.1626558)

Figure 1. Continuum of ambiance focus from real sounding environments, to more abstracted spaces. Production contexts range across the continuum.

retrieving files from a database, artificially generating them in a production environment, or by some combination of these techniques. Each of these approaches has an associated cost. For example, location recordings document a locale at a specific time. A drawback to location recording is that it is costly to visit a location and the recordings obtained can be impacted by unwanted environmental factors such as wind and traffic. Databases of soundscape recordings are collections of location recordings made by companies and enthusiasts. Databases overcome many of the costs associated with location recording, and offer a wide range of soundscapes. However, recordings are still limited to a fixed set of items. If no recording fitting a specification exists, then the sound designer must make a location recording, or create an artificial soundscape.

The most flexible method for generating soundscapes is to artificially produce them. The typical method is to combine particular environmental sounds obtained through synthesis using computational models or using parts of existing recordings. Current tools for making artificial soundscape compositions require a highly specialised skill set, and are often time-consuming. Therefore, computational tools to automate the tasks would be beneficial for sound designers. Furthermore, computational tools could make dynamic sound designs that respond to non-linear environments in video games and virtual reality possible.

Concerning generating a soundscape composition, researcher and soundwalk artist McCartney (2002) broadly outlines the basic tasks as follows:

(1) Specify an environmental context – The specification is a description of the environment to be represented in the sound design;
(2) Sound file retrieval – Sound files matching the specification, or parts thereof, will be retrieved from a database using semantic criteria, such as keyword search;
(3) Listen for salient regions in recordings – A sound file is curated for parts that hold an aesthetic interest. These regions are extracted for processing and mixing;
(4) Mix and sequence – Regions are sequenced on a timeline, and panning, attenuation, and sound effects applied to the sound design output;

Schafer (1977) found that acoustic properties alone failed to account for the human experience of sound. Based on these findings, an automated system for creating meaningful soundscapes must consider the human experience.

However, attempts at synthesising human understanding and decision processes are limited by the accuracy and extent at which the model represents human experience. The complexity of sound design is, in part, produced by the minutiae of human expectations and responses to the auditory environment. To account for this detail, sound designers select and combine multiple sounds using a variety of techniques.

We will describe Audio Metaphor, a system for generating a mix of soundscape compositions based on given descriptions and a database of existing soundscape recording files. This system utilises acoustic features, semantic information, and emotional characteristics of soundscape recordings.

We limit the scope of the research here to combine the emotion and semantic attributes of audio file selection, and apply this to the problem of mixing sounds. In doing so, we describe our contributions in advancing an autonomous soundscape generation system by developing techniques from computational creativity, machine learning, and audio signal analysis for the automation of soundscape composition tasks, including search, classification, and mixing.

This paper outlines the background and research for a soundscape generation system. We describe one such system named Audio Metaphor. After an introduction to the soundscape generation problem in Section 2, we put forward our research questions. In Section 3, we survey the field and define a soundscape generation system typology for establishing the focus of the research. We then analyse the evaluation methodology of generative systems in Section 4. Thereafter, in Section 5, we outline our current work toward defining and modelling characteristics of soundscape and sound design, including the algorithm for generating soundscapes. Sections 6 and 7 are about our evaluations and discussion, respectively. Finally, in Section 8, we present our concluding remarks.

2. Motivation and objectives

Soundscape composition is a contemporary form of electroacoustic music and sound art. A common characteristic of this type of art is the presence of recognisable environmental sounds and contexts. The purpose of these

sounds and contexts is to evoke the listeners' associations, memories, and imagination related to the soundscape.

Our research explores how a machine can autonomously generate soundscape compositions. We identify the requirements of soundscape generation by refining the tasks outlined by McCartney (2002) to four (4) modules needed by generative soundscape systems: soundscape context representation, search, classification and segmentation, and mixing.

2.1. How can a machine generate a soundscape with the features of a text-based description?

Soundscape context representation requires a computational method for describing the semantic and perceptual qualities of a soundscape. Using a survey to study how people describe soundscapes, Pedersen (2008) found that some people reference objects and events making sounds, while Davies et al. (2007) observed the population sample also used subjective feelings of the qualities of soundscapes. Although Niessen, Cance, and Dubois (2010) observed that there is a difference in the language used to describe soundscapes between cultural groups, Dubois and Guastavino (2006) found general commonalities for describing soundscapes. For example, a shared language is used as the basis of sound design for film, which as described by Sonnenschein (2001), involves the close reading of a script as a specification for sound design.

Based on the evidence in the literature, we choose to use a text-based utterance for representing a soundscape recording. We unpack an utterance representation taking into account how people communicate about soundscapes by exploring a text analysis algorithm. The algorithm maps descriptive words to items from a database with the aim of optimising search results for soundscape generation systems.

2.2. How can a machine generate soundscape compositions that have the basic structural elements?

Classification and segmentation require a set of machine learning tasks to automatically cut and label audio recordings. According to Schafer (1977), a soundscape has many types of sounds that fall into a set of perceptual categories:

- Keynote. These sounds appear to come from farther away, or occur in high frequency and belong to the aggregate of all sounds that make up the background texture of a soundscape.
- Sound signal. These sounds are typically heard in the foreground, standing out clearly against the background.
- Soundmark. A soundmark can be either background or foreground sound, and is culturally specific so that it defines a locale.

A soundscape has both sound signal and keynote sounds, which occur individually or simultaneously, and a soundscape composition may contain both these types of sounds. A sound designer needs to pay attention to these sounds when generating a soundscape composition because, similar to real soundscapes, a soundscape composition needs to represent both sound signal and keynote sounds. To obtain these different types of sounds from a recording, a sound designer will listen to soundscape recordings and extract regions to mix together.

Keynote sounds are associated with the background sound of a soundscape, while sound signal sounds are associated with the foreground. Hence, a background and foreground classifier would be a practical way to approximate these sound types.

Soundmark classification, on the other hand, requires a subjective appraisal of a listener to recognise cultural significance, and both background or foreground sounds can be soundmarks. Such sounds could be recognised by a computer that has an extensive collection of user models that map locale and semantic information to particular users. However, such a model is outside of the scope of this research.

Instead, we will concentrate on modelling background and foreground sound. A sound can be either background or foreground depending on factors including listening context and attention. Schafer (1977) outlines the following six types of sound:

(1) Natural sounds: e.g. birds, insects, rain;
(2) Human sounds: e.g. laugh, whisper;
(3) Sounds and society: e.g. party, church bells, concert;
(4) Mechanical sounds: e.g. airplane, machines, cars;
(5) Quiet and silence: e.g. dark night, wild space; and
(6) Sounds as indicators: e.g. clock, doorbell, siren.

With any of these types of sounds, the external listening context influences background and foreground classification. For example, the sound of a drop of water in the bathtub is accentuated by the bathroom environment, whereas it becomes a part of the background texture when in the ocean. The listeners' attention is the second factor in perceiving a sound as background or foreground. For example, the sound from the TV is foreground when a show is being watched, but becomes background when the viewers' attention is turned to a

conversation in the kitchen. Similarly, Truax (2001) outlines how listening is a dynamic process of different listening modes. Listening modes can treat any sound as either background or foreground depending on the level of attention being paid at any given moment.

Our research accounts for context but not attention i.e. the drop of water example will work, but the TV listening example will not. When the context is being addressed, background sounds seem to come from farther away than foreground sounds, or are continuous enough to belong to the aggregate of all sounds that make up the background texture of a soundscape. This is synonymous with a ubiquitous sound, specified by Augoyard and Torgue (2006) as 'a sound that is diffuse, omnidirectional, constant, and prone to sound absorption and reflection factors having an overall effect on the quality of the sound'. Urban drones and the hum of insects are two examples of background sound. Foreground sounds are typically heard standing out clearly against the background. At any moment there may be either background sound, foreground sound or a combination of both.

2.3. How can a machine generate soundscape compositions that are perceived to have a particular emotion?

Another characteristic of soundscapes is the perceived emotion of the soundscape (Berglund, Nilsson, & Axelsson, 2007; Botteldooren, Coensel, & Muer, 2006; Hall, Irwin, Edmondson-Jones, Phillips, & Poxon, 2013). Due to the subjective nature of emotion, it is natural that people have different emotional responses towards the same content. Therefore, it is essential to define different types of emotions. There are three types of emotion for when an individual is listening to music and soundscape or watching films:

- Intended emotion: The emotional response that the music/movie attempts to evoke in its viewers (Malandrakis, Potamianos, Evangelopoulos, & Zlatintsi, 2011).
- Perceived emotion: Emotions that are communicated and expressed by the source (Kallinen & Ravaja, 2006).
- Induced emotion: Emotional reactions that the source provokes in an audience (Kallinen & Ravaja, 2006). It is what the audience feels from the source.

The perceived emotion is the emotion the source conveys. The perceived emotion of happy songs is always 'happy'. However, the induced emotion is more subjective. The same happy music may not necessarily induce happiness in the listener. In this study, we focus on the perceived emotion of soundscape compositions. Intended emotion is the target emotion that sound designer trying to achieve. For example, the quality of a soundscape in a haunted abandoned village is different to that of one that is associated with a sunny village, on a festive day. Jianyu, Miles, and Philippe (2015), Fan, Thorogood, and Pasquier (2016), Fan, Thorogood, Tatar, and Pasquier (2018) and Fan, Tung, Li, and Pasquier (2018) conduct a series of studies to investigate soundscape emotion recognition. However, previous studies only focus on analysing soundscape emotion. To date, soundscape generation systems have not accounted for emotion.

We explore methods from affective computing for predicting the perceived emotion of a soundscape. The affective computing literature has proposed different models for representing emotion (Scherer, Bänziger, & Roesch, 2010). Two types of models that are regularly cited are categorical and dimensional models. The categorical model classifies a sample belonging to a finite number of emotions. For example, a common set of categories includes sad, happy, angry, and relaxed.

Alternatively, the dimensional model of affect views emotion as a point along dimensions - somewhere between happy and sad, for example. A widely recognised dimensional model proposed by Russell, Weiss, and Mendelsohn (1989) shows emotion as being a point in a multidimensional space of valence and arousal. The dimensional model can be used to allocate specific emotions to a particular region of the continuous space (Kim et al., 2010). For example, excited is associated with high valence and high arousal.

The dimensional model of affect is used in many survey studies as an effective tool for evaluating soundscapes (Berglund et al., 2007; Brocolini et al., 2010; Davies, Adams, Bruce, Carlyle, & Cusack, 2009), showing agreement for the perceived affect of particular soundscapes. The simplicity and success of the dimensional model is encouraging, warranting investigation as a method of computational representation of perceived emotions of soundscapes. We evaluate the feasibility of using this model for soundscape generation systems with techniques from music information retrieval, including psychoacoustic study, feature extraction, and standard statistical pattern matching.

2.4. How can a machine generate a soundscape composition the evokes the memories and associations of a real or imagined soundscape?

Broadly, mixing is the process of combining sounds and modulating volume and spectral parameters. The search space of mixing is combinatorial and inherently large. The sound design literature (Bazil, 2008; Brandon, 2005;

deBeer, 2012; Farnell, 2010) describes how an experienced sound designer obtains knowledge of techniques to apply for the highest likelihood of success in moving toward the intended audio output.

One method of specifying a soundscape is by describing an environmental context, as is common with the description of sound design on a film script. For example, *the creepy house has bats in the attic*. A sound designer uses a set of sound files obtained from a search, along with signal processing techniques to create a mix representing that specification. A sound designer must audition and subsequently segment sound files from a search using criteria such as background/foreground and emotion. Finally, sound files are sequenced on a timeline and signal processing techniques are applied to satisfy the specification and overall aesthetic continuity.

State of the art soundscape generation systems use techniques from AI, machine learning, and audio signal analysis to autonomously mix soundscape recordings and automate sound design tasks (Bruce, Davies, & Adams, 2009; Casu, Koutsomichalis, & Valle, 2014; Janer, Kersten, Schirosa, & Roma, 2011; Thorogood, Pasquier, & Eigenfeldt, 2012; Valle, Schirosa, & Lombardo, 2009). However, modelling sound design principles, such as emotion, remains an open problem. Further, evaluating such systems in comparison with human sound design still needs to be explored. In response, we outline a computational model of sound design principles, and integrate it into established search algorithms for generating soundscape compositions.

3. Related soundscape generation systems

Multiple soundscape generation systems have used different approaches to automating sound file retrieval, segmentation, and arrangement. These approaches take the direction of either operating in realtime or offline modes, displaying different levels of automation, and applying different sequencing techniques.

From the literature, two soundscape generation models emerge. The first model uses a layered approach to mixing, which reflects the process of multitrack recording software (for example, Audacity, Reaper). In this model, there is an arbitrary number of audio tracks with audio clips sequenced along a timeline. This approach calls for creativity in selecting and mixing together sounds, and is the primary means of production in sound design for linear media and soundscape composition. Another approach is to simulate sound sources in an environment based on a realistic sound spatialization. In the spatialization approach, which is common in game engines (Firelight Technologies, 2002 and Audiokinetic, 2000 for example), a virtual listener is placed in a virtual 3D environment and sounds are added to the environment with attributes such as position and motion.

Birchfield, Mattar, and Sundaram (2005) report on a semi-autonomous system that uses a selection of user models that influence sound selection and mixing parameters of the system. A database of 300 sound files was curated and annotated by the researchers according to spatial placement, location, and soundscape contexts. The system automatically selected sound files and mixed these together based on the user model, taking into account users contextual expectations of sounds. A sound designer then made decisions on applying effects.

Cano, Fabig, Gouyon, and Loscos (2004) describe a system of file selection, audio track sequencing and mixing for semi-automatic ambience generation. The system then selects sounds from text-based utterances using a keyword-spotting technique that links concepts of the keywords and returns a randomised set of sound files with the same concept from a database indexed by tags.

Cano et al. outline a soundscape model of mixing long ambient sounds with short event sounds occurring at intervals corresponding to moments of less energy in the ambience. Similarly, Salamon, MacConnell, Cartwright, Li, and Bello (2017) generates a soundscape as a series of foreground events and a single background recording. The user must supply the collection of classified sound clips.

Using a similar model of sequencing sounds, Rossignol, Lafay, Lagrange, and Misdariis (2014), describe the soundscape generation system SimScene for creating material in psychology studies on soundscape perception. Rossignol et al. manually label sets of sounds for textures and events using an urban sound taxonomy, organising sounds based on the domain, category, and sound class. They define the SimSound soundscape model as background texture with multiple types of event sounds occurring at distinct intervals. A user selects a sound event and texture class through a graphical interface. SimSound generates a texture by concatenating texture files for that class, and sequences all sounds of the selected event class according to parameters set by the user.

As another approach, Eigenfeldt and Pasquier (2011) designed a system which combined autonomous software agents and audio signal analysis to generate a continuously evolving soundscape composition. Similar to Birchfield et al. (2005), a database of sound files were curated and selected by the researchers. Software agents generated a layered soundscape composition in realtime using sound file-selection criteria negotiated on the basis of semantic tags and spectral attributes. The resulting compositions, although well-formed, were not rated as highly as similarly constrained human-made compositions.

Thorogood et al. (2012) outline a system to autonomously curate and select sound files based on text analysis of social media posts and sound file descriptions. Thereafter, sound file recommendations are pushed to a human performer for mixing. Thorogood and Pasquier (2013a) later describe a system that automates the segmentation procedure based on a machine learning approach to classify audio signal with soundscape perception model of background and foreground classes. The system then selected and applied effects for a layered soundscape based on a set of rules informed by production notes from the soundscape composition *Island* by Canadian composer Barry Truax (2009).

Another method of selecting, annotating, and mixing sound files is to capitalise on the user's propensity to generate data in online media sharing platforms (Roma et al., 2012). Roma, Herrera, and Serra (2009) use a sound file dataset and users' behaviours from the file-sharing platform Freesound (Akkermans et al., 2011) for generating continuous mixes of sound files. Through the web interface, users linked sounds together from the database into simple sound design patches: a directed graph of sounds as vertices and edges of the play order path. A linear crossfade applied between sounds completed the composition. Users then rated each other's patches that were processed using an interactive genetic algorithm to create new patches.

Finney and Janer (2010) designed a semi-autonomous system for generating soundscapes in virtual environments. Sounds used in their system were hand-picked for audio quality and semantic attributes viewed as germane to the represented environment. Sounds were mixed using an interactive map interface, and combined into background or foreground layers.

Roma et al. (2010) use a technique for autonomously labelling sound files with high-level concepts from the sound taxonomy proposed by Gaver (1993). Using these autonomously labelled files, Janer et al. (2011) and Janer, Roma, and Kersten (2011) developed a semi-autonomous system for augmented reality systems that generated soundscapes for virtual environments. Their strategy was to provide an authoring environment that let users select sounds based on the taxonomy and semantic tags, and position sounds by hand using a map interface.

Works utilising soundscape creation for virtual environments, such as from Tactical Sound Garden Toolkit by Shepard (2007) and the Urban Remix project by Freeman, DiSalvo, Nitsche, and Garret (2011), move toward the development of collaborative recording, exploration, and soundscape creation systems. These systems provided participants with a mobile interface for recording and tagging environmental sounds on a map interface. Sounds are layered based on geographic locations and spatialised as a virtual soundscape based on the user's proximity.

Yet another soundscape generation system is outlined by Casu et al. (2014), who describe a set of automated search and composition tools named SoDA for assisting sound designers with generating soundscapes. SoDA uses a database of sound files accompanied by a set of corresponding RDF documents. The user enters sound description fields by hand and the machine automation enters analysis features. SoDA's soundscape model is based on a physical model, with an omni-directional background sound, and sequences of shorter sound events occurring in a 3D space relative to where the sounds occur in physical space.

As outlined in the literature, different systems aim toward making one or more of the repetitive and time consuming tasks of working with sound more accessible to a user/author. In doing so, these systems can reveal patterns in the sound data that make manifest multiple sound design solutions that address a specific criteria encoded in the programming.

Our research here aims to address the problem of automating sound file search, retrieval, segmentation, and arrangement for sound design production using the semantic and sentiment criteria. In doing so, we model parts of the semantic and sentiment space of soundscapes that is utilised in the generation of artificial soundscapes.

4. Related evaluation methodologies

This section describes the related work in evaluating computationally assistive and creative machines in order to frame our evaluation methodology. To evaluate the Audio Metaphor system here, we design a group of experiments grounded in the computational creativity literature for testing the validity of the output of the system. As Arthur Flexer (2006) and Jordanous (2011) state, evaluating creative systems is important for elucidating progress in the research, and proposals have been put forward as the most valid means evaluating these types of systems (Pearce & Wiggins, 2001; Pease & Colton, 2011; Ventura, 2008). In general, asking if a system is capable of creativity is a dyadic question. On the one hand, the question is whether the process conducted by the machine can be considered creative. While on the other hand, the question is if the output of the system is considered to be a creative artefact.

While both of these questions are compelling in observing the creativity of a system, we are concerned with the soundscape compositions generated by Audio Metaphor. Specifically, we need to ascertain the quality

of the generated soundscape compositions in comparison to a soundscape composition made by a human expert. Notwithstanding that Audio Metaphor is modelling sound design tasks and executes these tasks simulating human processes, we focus the evaluation on whether the output of the system is considered to be a creative artefact relative to human endeavours.

Evaluating the creative output of a system such as Audio Metaphor can be achieved by either analysing the output of the system in relation to a corpus of expected outputs, such as the methodology outlines by Thomas, Pasquier, Eigenfeldt, and Maxwell (2013), or the outputs of the system can be given to human judges to listen and rate against some criteria. The former method is appropriate when a corpus of expected results exists, while the latter asks questions where the success of the system depends on a more subjective nature. As such, the soundscape compositions generated by Audio Metaphor, expressing emotion and sound design quality, are best suited to evaluation by listeners who will rate the soundscape compositions.

4.1. Evaluating for creativity support tools

Creativity support tools (CST) is a branch of human-computer interaction (HCI) studying systems for assisting in human creativity tasks. A fundamental element of HCI is evaluating systems with the appropriate metrics and evaluation methodology. For example, Candy and Edmonds (1997) identify three (3) criteria for evaluating interaction in software systems. These measures require a system to provide methods for modifying the knowledge base, examining the system as to the reason for an outcome, and comparing different alternatives. The motivating factor of these criteria is to understand the interactions of users with the system in creative applications.

Cherry and Latulipe (2014) define another set of criteria, named the Creativity Support Index (CSI), for evaluating computer-assisted creativity tools. Similar to the principles of Candy and Edmonds, CSI aims at assessing an interactive system based on the engagement of users, factoring in immersion, enjoyment, and collaboration as criteria.

Audio Metaphor aims at generation, not user interaction, and does not to aim to meet the above criteria. In particular, the knowledge base rules of Audio Metaphor are inaccessible for modification by the user, decisions of the system are opaque, and there is no functionality for comparative evaluations of alternatives. However, future work will look at more user input, looking at CST research such as these.

4.2. Evaluating computational creativity

As an alternative to evaluating with creativity support tools, we use an evaluation methodology grounded in the computational creativity literature to evaluate Audio Metaphor. Computational creativity aims at developing computer programs that output human-competitive creative behaviours and artefacts. Wiggins (2006) suggests a system is considered creative if behaviour exhibited by the system would be deemed creative were it exhibited by a human. Boden's (2004) distinction between P-creativity and H-creativity is also helpful in determining the creative behaviour of Audio Metaphor. P-creativity involves generating surprising and novel instances of existing ideas, such as a song in a music genre for example. On the other hand, H-creativity is an entirely new idea, such as a new music genre. The output of the system here is P-creative, such that the search space is bounded by a set of established mixing processes defined by sound design.

In studying soundscape generation systems, we need to ascertain if the generated soundscape compositions are considered novel and valuable in fair competition with human efforts. Pearce and Wiggins (2001) describe a simple evaluation methodology for judging if composition systems are indistinguishable from human-made pieces. In this type of discrimination test, a group of study participants presented with compositions discern if a machine generated and/or a human created the piece.

Psychometric data obtained with a Likert-scale and reported with a t-test statistic is a method widely used for evaluating creative systems. Likert-scales are a type of questionnaire used to get listener responses (Finney & Janer, 2010), with a scale for each response. For example, to the proposition 'I like Justin Beiber', a 3-point Likert-scale may have Disagree - Neutral - Agree. However, when used in evaluating creative systems, one argument suggests people are biased when it comes to machine made artefacts (Moffat & Kelly, 2006). In response, Ariza (2009) outlines the Musical Output Toy Test to account for such bias. A computer and a human composer generate a piece of music as a digital recording or a score, the two composers have access to the same set of resources, and the music must be an original composition. A participant is informed that a machine generated some of the music and is asked to distinguish the human from the machine compositions.

Although we investigate the human-versus-system dyad by comparing listener responses to soundscape compositions generated by Audio Metaphor and a human composer, we are also interested in exploring questions of sound design principles of emotion and soundscape characteristics defined in the system.

Jordanous (2011) proposed evaluating creative systems based on the creative domain specifics, the aspects of creativity that are important, and the set of standards used for evaluating the system. With this type of evaluation, deception can be used for circumventing a participant's biases of machine-generated artefacts. In which case, participants will not be informed a machine is involved.

Eigenfeldt and Pasquier (2011) adapt this approach for evaluating a soundscape generation system. In their study, listeners rank system and human generated soundscapes with questions grounded in the soundscape composition literature. We embrace the methodological approach taken by Eigenfeldt and Pasquier in defining the creative domain specifics for establishing questions for determining the system's successes. However, designing a study for obtaining psychometric data is not without problems.

When planning a study using a Likert-scale, there is no definitive consensus on the adequate scale length. Further, the validity of using a Likert-scale in a study such as ours has been called into question (Pease & Colton, 2011). In a study such as ours, a Likert-scale would have terms at either end of the scale representing ends of a continuum. A participant enters a response between these terms on a discrete scale. For example, unpleasant – pleasant is one such pair of words. If a participant enters a value of 0, they perceive the soundscape to be downright unpleasant. Responses signify less unpleasant toward the centre of the scale, and more pleasant as they move toward the pleasant term, and entirely pleasant at the end of the scale.

The length of the scale, or the number of response alternatives, is a factor in experiment design. Symonds (1924) asserts that using a 7-point scale achieves an optimal level of score reliability. Cook, Heath, Thompson, and Thompson (2001) suggest increasing the number of response alternatives increases score variance, and thus increase score reliability. As another consideration to the design of a Likert-scale, Garland (1991) identifies that the presence of a mid-point when using Likert-scales, as is the case with an odd-number of selections, results in a social desirability bias on the part of a respondent's desire to please the interviewer or appear helpful.

Matell and Jacoby (1971) suggest, when including a mid-point, the effect of social desirability bias is reduced with Likert-scales of increasing length. Although even with a 21 point scale, Pearse (2011) shows that respondents have a tendency to choose the mid-point response. Garland (1991) demonstrated that social desirability bias is minimised by eliminating the mid-point.

5. Mixing engine

We wish to model the formal definitions of soundscape properties to demonstrate the proposed framework. Our soundscape mixing engine will be based on a model of sound design originating from principles in the literature and observations of sound design practice. The information gathered in the model design will inform the exploration of a generative system. To achieve our research objectives, we adopt a multidisciplinary approach. The motivation of this approach is the complex problem space of designing systems for synthesising human creativity. In our system, we adopt and combine methods from modelling expert domains, text analysis, music information retrieval, machine learning, and artificial intelligence.

The mixing engine generates soundscape compositions that are perceived to have acoustic properties correlating with the semantic and sentiment specifications. To achieve this goal, we design a mixing engine to generate a soundscape composition from a text-based utterance and curves for the movement of valence and arousal over the duration of the soundscape composition. The engine has access to a database of audio segments indexed by semantic descriptors, valence and arousal values, and whether the sound is perceived as background or foreground. The engine creates background and foreground tracks for each of the soundscape concepts, and assigns sound recording segments to those tracks. We define the terms of our system as follows:

Input Specification. The input specification is an utterance (e.g. 'the quiet stream and the busy traffic'), together with a set of sound segments corresponding to each concept, plus the duration of the soundscape composition, ranging from 20 s to 2 min, and a vector of values between 0 and 1 representing the curves for for pleasantness and eventfulness.

Sound Source Concept. Here a sound source concept is part of an utterance that is used for searching a database for corresponding sound segments. There may be one or more concepts in an utterance, and a typical specification will have from 1 to 10 concepts. If an utterance has no concepts, then there are no segments, and a soundscape composition is not generated.

Segment. A segment is a region of a sound file related to a sound source concept. We also introduce a segment containing silence that can be assigned to a track.

Track. A track supports both a stereo and mono segments. Two tracks are created for each concept. One contains the background segments; the other contains foreground segments.

Mix. A mix has background and foreground tracks for each sound source in the set provided at the input specification.

Decision Point. A decision point is a trigger point occurring when a segment ends or at an interval of 1 s, whichever is sooner. At a decision point, a new segment, which is also possibly silent, is added to an empty track.

The engine must determine the segments to assign to tracks subject to a number of constraints. We formalise this problem by determining the variables, and the constraints needing optimisation.

5.1. Soundscape mix generation

The mixing engine generates a mix by assigning segments at decision points based on the output from the objective function. Initially, the segments are retrieved by a search algorithm from a database of field recordings indexed by user-contributed tags, which Lamere (2008) identifies as valuable to music information retrieval. We adopt the file search protocol outlined by Thorogood et al. (2012) that uses a keyword-spotting and grouping technique to optimise the diversity of sounds presented for mixing.

We then use a BF classifier, an audio file segmentation system that segments the selected recordings according to the perceptual categories, including background, foreground, and background with foreground sound (Thorogood, Fan, & Pasquier, 2015, 2016). The system was built based on a corpus of background, foreground, and background with foreground soundscape recordings. Previous analysis of annotations has demonstrated a high degree of certainty for these three categories. Thorogood et al. train a Support Vector Machine classifier and evaluated the classifier with different analysis window sizes. The results indicate the effectiveness of the classification.

Next, we use a soundscape emotion recognition system described by Thorogood and Pasquier (2013b) and Fan et al. (2016) to label the segments with pleasantness and eventfulness values. In previous studies, Fan et al. conducted a study that annotated the perceived pleasantness and eventfulness of a corpus of soundscape recordings. Then, the authors use stepwise regression to train two models to predict the perceived pleasantness and eventfulness.

Given our mix formalism, we implement a minimum-conflicts algorithm (Minton, Johnston, Philips, & Laird, 1992) that tests a series of possible solutions for finding the best segment combinations within a given time.

Variables. At a decision point, the engine executes the minimum conflicts algorithm to optimise the assignment of a segment to a track. The engine observes the target specification values for pleasantness and eventfulness. We therefore introduce variables $spec_p$, and $spec_e$ for the input specification values pleasantness and eventfulness, respectively. Similarly, the mix generates a new set of pleasantness and eventfulness values, and we introduce variables $measured_p$, and $measured_e$, to capture these. We adopt the predictive model outlined by Fan et al. (2016) for generating the mixed pleasantness and eventfulness values. Deciding on a mix now amounts to finding segments for each track at the decision point that minimises the objective function. We choose a function based on the Euclidean distance calculated by:

$$\sqrt{(measured_p - spec_p)^2 + (measured_e - spec_e)^2} \quad (1)$$

Introduction of Silence. The background tracks of a mix always have sound occurring, whereas foreground sounds tend to be intermittent depending on the soundscape. Therefore, we introduce a segment containing silence as a possible alternative selection for foreground type tracks. If a decision point chooses silence as the most viable option from the selection of segments, then the span of the silent region extends to the next decision point. Silence can be chosen at consecutive decision points.

5.2. Segment selection

Minimum-conflicts is a local-search technique effective for many optimisation problems with solutions densely distributed throughout the state space. It has the property of finding the best solution within the search space given time constraints.

As is shown in Algorithm 1, the algorithm is initialised with a random mix, and outputs a mix converging toward an optimal combination of tracks. A track is selected at random and assigned a segment that has the greatest effect on minimising the objective function. This process is repeated until a set number of iterations is reached.

5.3. Generating a mix

The assignment step is triggered at regular intervals of 0.25 s, or whenever a change in the mix occurs (i.e. the end-time of a segment), whichever is sooner. As shown in Figure 2, a benefit of our approach for building a mix at a decision point is allowing the definition of curves for modulating the soundscape pleasantness and eventfulness over time. An example of a simple dramatic arc is this: a soundscape starts as pleasant and uneventful, approaches unpleasant and highly eventful at the halfway mark, then ends toward pleasant and moderately eventful. This arc can be constructed.

6. Evaluation

The evaluation of the Audio Metaphor system here aims to ascertain if there is a perceived difference, regarding

input : maxSteps,number of steps before giving up, currentMix, assignment of segments for tracks
output: a solution mix

for *i* ← *1 to maxSteps* **do**
  if *currentMix fulfills specification* **then**
    return currentMix
  end
  track ← a randomly chosen track without a segment at the decision point
  segment ← the segment for a track retrieved from a database search that minimises the objective function (Equation 1)
  set track segment in currentMix
end
return currentMix

Algorithm 1: Minimum-conflicts algorithm for selecting the combination of segments minimising the objective function.

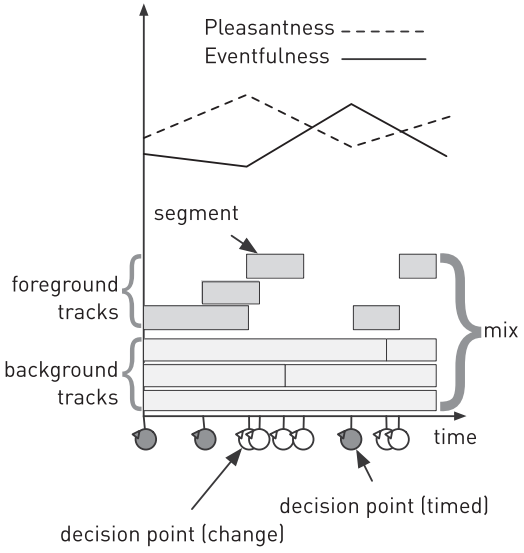

Figure 2. A graph representing the curves for pleasantness and eventfulness to modify the soundscape mix over time. The value of those curves is referenced at the corresponding selection stage to guide the decisions of the mixing engine.

semantics and emotion, between soundscape compositions generated by our system, those created by a human composer, and those randomly generated. Therefore, we conduct listening experiments for perceived pleasantness, eventfulness, and semantics, asking study participants to asses these soundscape compositions.

In this section, we outline our experiment. First, we define our research question. We define the three conditions used to generate the soundscape stimulus in our experiment. We then detail the dataset generated from these conditions. Next, we describe the instrumentation used in the experiment setup. Thereafter, we outline the questionnaire presented to participants and conclude by describing the data analysis plan.

6.1. Research question

We designed a system to generate a soundscape composition from a description with the goal that the soundscape corresponds to an acoustic environment and a particular emotion at a particular time. Therefore, we want to test the system for the perceived semantic and affective properties of the output. Further, we wish to determine if there is any difference in terms of perceived semantic and affective properties between soundscape compositions generated by our system, those created by a human composer, and a randomly generated soundscape composition.

6.2. Conditions

A soundscape specification is a sentence describing an acoustic environment, duration, and values for valence and arousal. We use a set of ten (10) soundscape specifications in this study to create an equal number of soundscape compositions with a generative mixing engine, a human soundscape composer, and a random mixing generator.

Each generator has the following set of mixing constraints:

- Only sound files from the specification can be used.
- Duration must be the length specified.
- Sound processing is constrained to segmenting a sound file, fading segments in and out to prevent clicks.

Condition 1: Machine Generated. Condition 1 includes soundscape compositions generated by Audio Metaphor. Audio Metaphor generates soundscape compositions from a database of audio files given a description that includes a text-based utterance, a desired duration, and a specific value for pleasantness and eventfulness (see Section 5).

Condition 2: Human Generated. Condition 2 includes soundscape compositions generated by a human expert. We give them the same specification of text-based utterance, a desired duration, and a specific value for pleasantness and eventfulness as in Condition 1. For each specification, the human expert has a database of audio files returned by the semantic search process carried out by Audio Metaphor – between 20 and 60 files depending on the specification. The experts used there own judgement regarding the pleasantness and eventfulness

Table 1. The list of soundscape utterances and accompanying desired pleasantness and eventfulness of the generated soundscape.

Sentence	Pleasant	Eventful
There are crows in the garbage	low	high
At the train station I hear churchbells	low	high
Quietly walking in the forest	high	low
I hear lambs in the dirt, and a horse	high	low
There is a tractor and grasshoppers	high	high
I hear birdsong and children are playing in the bush	high	high
There are bees buzzing in the city park	low	low
I hear barking in the meadow	low	low
There are bees buzzing in the city park	mid	mid
At the train station I hear churchbells	mid	mid

of particular audio files. A time limit of their discretion is allocated for the completion of the task. Experts returned their compositions no longer than 2 weeks after the task was accepted.

Condition 3: Random Generated. Condition 3 includes randomly generating soundscape compositions. The procedure is to create a background and foreground track for each concept outlined in the utterance. This condition accesses the same database of audio files segmented by the BF classifier returned from Audio Metaphor. A continuous mix is made on each track by randomly selecting from all segments associated with a concept until the given duration is reached. The random generation method disregards eventfulness and pleasantness information. Segments are sequenced end-to-end on each track, with a crossfade amount set at 0.125 s.

6.3. Dataset

We use a set of thirty (30) generated soundscape compositions for the evaluation experiment. To facilitate study participants' comprehension of soundscape semantics and emotion while not becoming fatigued by the listening process, each soundscape composition has a duration of 15 s. To create this dataset, we use a set of ten (10) soundscape descriptions that include a descriptive sentence and values for eventfulness and pleasantness (see Table 1). The set of soundscape descriptions is given to human experts, the Audio Metaphor system, and a system generates random soundscape compositions–resulting in three alternatives for each utterance.

The soundscape compositions in our study are generated from a curated selection of field-recordings. We use a dataset of 4085 audio files from the Freesound Project (Freesound, 2012) that have a Creative Commons License allowing for modification and reuse. A group of soundscape experts curated the sound files from a larger set, removing files that were of poor audio quality and those that did not match the semantics of the accompanying tags entered by Freesound users. Durations of sound files range from a few seconds up to 10 min. Following the licencing agreement as outlined by the Freesound Project, we make a list of attributions available from our project website. We also make the corpus of curated recordings availableat that site.[1]

To create the ten simple descriptive sentences, we analyse tags Freesound users labelled sound files with in the database. We examine the word frequency of tags on recordings, then group the tags regarding the following soundscape contexts: wilderness, urban, rural, and marine. We then generate random samples of tags using the probabilities associated with a tag occurrence. Finally, we connect tags with any necessary parts of speech to create a descriptive sentence. We constrain the number of tags in each sentence to being between two and three terms.

The remaining properties of the soundscape specification include a value for the desired duration, and values for the pleasantness and eventfulness. We set the time limit for each soundscape composition to sixty seconds. To explore the range of the capability of Audio Metaphor, we set the pleasantness and eventfulness values at the extremes of the affect grid, and in the middle of the grid.

A human composer receives a paper form of a specification with the pleasantness and eventfulness values denoted on a two-dimensional affect grid denoting pleasantness and eventfulness. The Audio Metaphor system receives the sentence, and pleasantness and eventfulness values. The random generation system receives only the sentence.

After receiving a completed soundscape composition from each condition, a fifteen-second interval is extracted from the sound file centred at the middle of the total duration. To prevent clicks, we apply a 0.01-second fade-in and fade-out to the extracted region. The processed region, soundscape specification and condition are logged.

6.4. Instrumentation

In the evaluation of Audio Metaphor, we conduct three experiments. The first experiment aims to test if there is a significant difference between the perceived pleasantness of soundscape compositions generated from the three conditions. The second experiment asks a similar question in terms of perceived eventfulness. The last experiment asks participants about the correlation between the utterance and the sounds heard, and if the soundscape composition is believable.

We use Figure Eight (Morris, McDuff, & Calvo, 2014), a crowdsourcing company that lets users access an

[1] www.audiometaphor.ca/

online workforce of millions of individuals to label data, to recruit participants. The research outlined by Fan, Thorogood, and Pasquier (2017) using a similar listening study has used this crowdsourcing method, with participants accessing the survey using a standard computer system and web-browser. To the best of our abilities, we avoid the limitations of crowdsourcing by marking the necessary precautions in the tutorial, training participants before the study, inserting 8 test questions to ensure the quality of results, and use outlier detection techniques in the post-processing stage.

Before commencing the study, we provide participants with information about the purpose of the experiment and the properties of a soundscape. In the tutorial, a participant is asked to use stereo headphones to hear the audio. Although the sounds in our study are best described as ambient, we ask that participants adjust the volume level of their system so that the loudest recording in the experiment dataset is at a comfortable listening level. The participant is requested to follow a short tutorial to be familiar with both the study interface and listening to soundscape compositions.

During the study, a participant is presented with the set of three soundscape compositions for a particular specification: one generated by Audio Metaphor, one generated by a human, and one randomly generated. The soundscape compositions are presented in a randomised order, and can be listened to in any order and listened to repeatedly. After listening to a set of soundscape compositions, a participant responds to a 10-point Likert-scale question for each condition.

The study interface has a simple audio player with start, stop and listen buttons. They can listen to the audio repeatedly. They're presented with a set of Likert-scale questions. Upon listening to the soundscape compositions and entering their response, the system logs the data and presents the next sample.

6.5. Participants

To be recruited for this study, participants need to be able to use a computer and listen to audio with headphones. They must be able to read English, as text-based utterances are presented in this language. People with Figure Eight accounts can view and decide if they want to participate in the survey for payment.

Before entering the actual study, participants are presented with a quiz, where five gold standard questions are provided. For example, in the first experiment that tests if there is a significant difference between the perceived pleasantness of soundscape compositions generated from the three conditions (Audio Metaphor, human composed, and randomly generated), we provide audio recordings that are easily distinguishable regarding the perceived pleasantness under these three conditions. These audio recordings are carefully selected by experts. For the experiments testing the perceived eventfulness and semantics, we also set gold standard questions. To ensure that a participant has a firm understanding of the tasks, we exclude participants who score less than 75% on the quiz questions.

Table 2. Geographic distribution of study participants.

Study	Venezuela	Egypt	US	other
Pleasantness	33%	13%	5%	49%
Eventfulness	30%	10%	10%	50%
Semantics	30%	8%	8%	54%

Each study takes a total time of 22.5 min to complete. This time includes listening to 7.5 min of audio with roughly 5 min for answering the questions, and 10 min for listening and reading in the tutorial stage. A participant can leave the study at any time without penalty.

There were a total of 418 participants for the pleasantness study, 474 participants for the eventfulness study and 633 participants for the semantics study. The geographic distribution of the participants is highlighted in Table 2.

6.6. Questionnaire

We run the study in three parts and design a separate 10-point Likert-scale questionnaire for each.

Emotion characteristics of the soundscape composition. The following three questions are designed to determine if a significant difference exists between the perceived emotion of the soundscape compositions generated by the 3 conditions.

The first question asks a participant to respond to the perceived pleasantness of the soundscape composition. After listening to a set of 3 recordings, a participant is asked to rate the perceived pleasantness of each. We place the tag *unpleasant* on one extreme of the scale and *pleasant* on the other.

Similar to the previous question regarding pleasantness, eventfulness is a term derived from the soundscape literature, used by people to describe the amount of perceived activity in a soundscape composition. After listening to a set of 3 recordings, a participant is asked to rate the eventfulness of each. We place the tag *uneventful* on one extreme of the scale and *eventful* on the other.

Semantic characteristics of the soundscape compositions. The next two questions seek to determine if the soundscape composition represents the one described in the text and if it sounds believable.

After listening to a set of 3 recordings, a participant is asked to rate their level of agreement with the statement

that a particular soundscape composition is believable. We place the tag *strongly agree* on one extreme of the scale and *strongly disagree* on the other.

The final question asks that a participant listen to the set of 3 recordings, and then rate their level of agreement with the statement that a particular soundscape composition represents the one described in the sentence. We place the tag *strongly agree* on one extreme of the scale and *strongly disagree* on the other.

6.7. Data analysis plan

We analyse psychometric data obtained from the study by applying standard statistical methods. Regarding our research questions, we wish to know whether machine-generated soundscape compositions are perceived as different from the human-generated and the random-generated. We conduct separate statistical tests for each of the properties of eventfulness, pleasantness, believability, and if the soundscape composition is representative of the utterance. For each of these properties, we test for rejection of the null hypothesis that there is no significant difference between the average ratings for the three conditions.

7. Results and discussion

We compute the error between the perceived and intended pleasantness and eventfulness. For each of the tests, an ANOVA shows there is significant difference ($p < .05$) between the average ratings for the three conditions. We therefore performed a Tukey post hoc test in each case to check for the significant differences between the results for each pair of conditions. A summary of these results is demonstrated in Tables 3 and 4.

Table 3. Mean and standard deviation results from an analysis of variance between the perceived affect of soundscapes generated by the three conditions. A smaller mean value demonstrates a greater accuracy of the soundscape to the specification.

	Pleasantness	Eventfulness
Machine	2.160, 2.181	3.079, 2.377
Human	2.680, 2.545	3.205,2.573
Random	4.696, 2.647	3.855, 2.797

Table 4. Mean and standard deviation results from an analysis of variance between the semantic properties of soundscapes generated by the three conditions. A smaller mean value demonstrates a greater agreement with the semantic property.

	Believable	Represents
Machine	2.727, 2.569	3.671, 3.217
Human	2.366, 2.569	3.924, 3.217
Random	3.234, 2.695	5.460, 3.568

7.1. Pleasantness

An ANOVA was conducted to compare the mean difference between ratings and target values of the perceived pleasantness of a soundscape composition. It shows that the effect of mixing engine type on the perceived pleasantness of a soundscape composition at the $p < .05$ level for the three conditions machine, human, and random was significant $F(2, 899) = 88.5656$, $p = 0.000$. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the Machine condition ($M = 2.160$, SD = 2.181), the Human condition ($M = 2.68$, SD = 2.545) and the Random condition ($M = 4.696$, SD = 2.647) were all significantly different.

7.2. Eventfulness

An ANOVA was also used to compare the mean difference between ratings and target values of the perceived eventfulness of a soundscape composition. It shows that the effect of mixing engine type on the perceived eventfulness of a soundscape composition at the $p < .05$ level for the three conditions machine, human, and random was significant $F(2, 641) = 5.534$, $p = 0.004$. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the Machine condition ($M = 3.079$, SD = 2.377) was not significantly different than the Human condition ($M = 3.205$, SD = 2.573). However, the Random condition ($M = 3.855$, SD = 2.797) was significantly different than the Machine condition and the Human condition.

7.3. Believable

Similarly, we conducted an ANOVA to compare the mean difference between ratings and target values of the perceived believability of a soundscape composition. It shows that the effect of mixing engine type on the perceived believability of a soundscape composition at the $p < .05$ level for the three conditions machine, human, and random was significant $F(2, 638) = 6.6225$, $p = 0.001$. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the Machine condition ($M = 2.727$, SD = 2.569) was not significantly different than the Human condition ($M = 2.366$, SD = 2.569). However, the Random condition ($M = 3.234$, SD = 2.695) was significantly different than the Machine condition and the Human condition.

7.4. Representation

An ANOVA test evaluating the mean difference between ratings and target values on how closely a soundscape

composition represented an utterance reveals the effect of mixing engine type on the representative level of a soundscape composition at the $p < .05$ level for the three conditions machine, human, and random was significant $F(2, 638) = 17.473, p = 0.000$. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the Machine condition ($M = 3.671$, SD = 3.217) was not significantly different than the Human condition ($M = 3.924$, SD = 3.217). However, the Random condition ($M = 5.460$, SD = 3.568) was significantly different than the Machine condition and the Human condition.

7.5. Discussion

Taken together, the results demonstrate that responses show variability between soundscapes generated by the machine, human experts, and randomly concerning the properties of pleasantness, eventfulness, believability, and representation. Moreover, a post-hoc analysis reveals that the Audio Metaphor system is human-competitive. In all four of the properties we test, Audio Metaphor always performs better than the baseline randomly generated soundscape compositions. Furthermore, Audio Metphor does at least as well as a human composer in representing the soundscape specification in terms of pleasantness, eventfulness, believability, and semantics. Examples of the system outputs are available at the project site.[2]

8. Conclusion

Sound designers have many tasks when generating soundscape compositions. Some of these tasks, such as sound file search/retrieval and segmentation are repetitive and time-consuming. The mixing task provides other challenges when working in generative contexts such as video games and virtual reality. Having a computer automate these tasks in a sound design workflow is an advantage. In answering the question of whether a machine can autonomously generate soundscape compositions, we have described a system for generating soundscape compositions. Our system utilises acoustic features, semantic information, and emotional characteristics to produce, from a database of existing sound files, soundscape compositions specific to an utterance.

To position our research, we surveyed the state of the art in soundscape generation systems, identifying the significant characteristics of these types of systems. These systems aim at automating one or more sound design tasks. Our research addresses the problem of automating sound file search, retrieval, segmentation, and arrangement for sound design production using the semantic and sentiment criteria.

We describe the Audio Metaphor system that automates the generation of a soundscape composition from an utterance and returns a soundscape representing semantics and emotion. Further, we describe a method to segment and classify sound files based on background, foreground, pleasantness, and eventfulness classes. Lastly, we detail a mixing engine that generates a layered soundscape composition. Through a crowdsourcing listening experiment, we evaluated the Audio Metaphor system for the properties of perceived pleasantness and eventfulness, the correlation between what is heard in the soundscape and read in the utterance, and the believability of the artificial soundscape. The results indicate that Audio Metaphor is human-competitive in all the properties under analysis.

Sound design is a creative process, and there are many nuances of the processes that must be taken into account when modelling for a generative system. Our research has identified and addressed the essential elements of soundscape generation systems, paying particular attention to details such as the perception of background/foreground and pleasantness/eventfulness. However, there is still much to explore in this domain. We have demonstrated how non-trivial tasks, traditionally, undertaken by sound designers, can be automated in a computational system. Soundscape and sound design literature have been the point of departure in recognising and modelling these tasks. Building on our research, we see further integration of human creative practice into autonomous systems used in sound based arts. As demonstrated here, one approach to this goal is to align knowledge from human perception, cultural factors, and situated sound design practice for building intelligent systems.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We acknowledge the support of the Insight program of the Social Sciences and Humanities Research Council of Canada.

References

Akkermans, V., Font, F., Funollet, J., de Jong, B., Roma, G., Togias, S., & Serra, X. (2011). *Freesound 2: An improved platform for sharing audio clips*. International society for music information retrieval conference, Miami, Florida.

Ariza C. (2009). The interrogator as critic: The turing test and the evaluation of generative music systems. *Computer Music Journal*, *33*(2), 48–70.

[2] https://digitalmedia.ok.ubc.ca/projects/aume/mixingExamples/

Audiokinetic. (2000). Wwise. Retrieved from http://www.audiokinetic.com.

Augoyard, J.-F., & Torgue, H. (2006). *Sonic experience: A guide to everyday sounds*. Montreal: McGill-Queen's University Press.

Bazil E. (2008). *Sound mixing tips and tricks*. Pc Publishing Series. PC Publishing.

Berglund, B., Nilsson, M. E., & Axelsson, Ö. (2007). *Soundscape psychophysics in place*. InterNoise, Istanbul.

Birchfield, D., Mattar, N., & Sundaram, H. (2005). *Design of a generative model for soundscape creation*. International computer music conference, Catalunya, Spain.

Boden, M. (2004). *The creative mind: Myths and mechanisms*. London: Taylor & Francis.

Botteldooren D., Coensel B. D., & Muer T. D. (2006). The temporal structure of urban soundscapes. *Journal of Sound and Vibration*, *292*(1-2), 105–123.

Brandon A. (2005). *Audio for games: Planning, process, and production*. New Riders Games Series. New Riders Games.

Brocolini, L., Waks, L., Lavandier, C., Marquis-Favre, C., Quoy, M., & Lavandier, M. (2010). *Comparison between multiple linear regressions and artificial neural networks to predict urban sound quality*. Proceedings of 20th international congress on acoustics, Sydney, Australia.

Bruce, N. S., Davies, W. J., & Adams, M. D. (2009). *Development of a soundscape simulator tool*. Internoise 09, Ottawa, Canada.

Candy L., & Edmonds E. A. (1997). Supporting the creative user: A criteria-based approach to interaction design. *Design Studies*, *18*(2), 185–194.

Cano, P., Fabig, L., Gouyon, F., & Loscos, A. (2004). *Semi-automatic ambiance generation*. Proceedings of 7th international conference on digital audio effects, Naples, Italy (pp. 1–4).

Casu, M., Koutsomichalis, M., & Valle, A. (2014). *Imaginary soundscapes: The soda project*. Proceedings of the 9th audio mostly: A conference on interaction with sound, Aalborg, Denmark (p. 5). ACM.

Cherry E., & Latulipe C. (2014, June). Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction*, *21*(4), 21:1–21:25.

Cook C., Heath F., Thompson R. L., & Thompson B. (2001). Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, *61*(4), 697–706.

Davies, W., Adams, M., Bruce, N., Cain, R., Carlyle, A., Cusack, P., ... Plack, C. (2007). *The positive soundscape project*. 19th international congress on acoustics, Madrid (pp. 2–7).

Davies, W. J., Adams, M. D., Bruce, N. S., Carlyle, A., & Cusack, P. (2009). *A positive soundscape evaluation tool*. Euronoise, Edinburgh.

DeBeer, G. (2012). *Pro tools 10 for game audio*. Ontario: Nelson Education.

Dubois, D., & Guastavino, C. (2006). *In search for soundscape indicators: Physical descriptions of semantic categories*. Internoise, Honolulu, Hawaii.

Eigenfeldt, A., & Pasquier, P. (2011). *Negotiated content: Generative soundscape composition by autonomous musical agents in coming together: Freesound*. Proceedings of the second international conference on computational creativity, Mexico City (pp. 27–32).

Fan J., Thorogood M., & Pasquier P. (2016). Automatic soundscape affect recognition using a dimensional approach. *Journal of the Audio Engineering Society. Audio Engineering Society*, *64*(9), 646–653.

Fan, J., Thorogood, M., & Pasquier, P. (2017). *Emo-soundscapes: A dataset for soundscape emotion recognition*. International conference on affective computing and intelligent interaction, Alamo, TX.

Fan, J., Thorogood, M., Tatar, K., & Pasquier, P. (2018, July). *Quantitative analysis of the impact of mixing on perceived emotion of soundscape recording*. Proceedings of the 15th sound and music computing, Limassol, Cyprus.

Fan, J., Tung, F., Li, W., & Pasquier, P. (2018, July). *Soundscape emotion recognition via deep learning*. Proceedings of the 15th sound and music computing, Limassol, Cyprus.

Farnell A. (2010). *Designing sound*. University Press Group Limited.

Finney, N., & Janer, J. (2010). *Soundscape generation for virtual environments using community-provided audio databases*. W3C workshop: Augmented reality on the web, Cambridge, MA.

Flexer A. (2006). Statistical evaluation of music information retrieval experiments. *Journal of New Music Research*, *35*(2), 113–120.

Firelight Technologies. (2002). FMOD. Retrieved from http://www.fmod.org/.

Freeman J., DiSalvo C., Nitsche M., & Garrett S. (2011). Soundscape composition and field recording as a platform for collaborative creativity. *Organized Sound*, *16*, 272–281.

Garland R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing Bulletin*, *2*(1), 66–70.

Gaver W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, *5*, 1–29.

Hall D. A., Irwin A., Edmondson-Jones M., Phillips S., & Poxon J. E. W. (2013). An exploratory evaluation of perceptual, psychoacoustic and acoustical properties of urban soundscapes. *Applied Acoustics*, *74*(2), 248–254.

Janer, J., Kersten, S., Schirosa, M., & Roma, G. (2011). *An online platform for interactive soundscapes with user-contributed audio content*. Audio engineering society conference: 41st international conference: Audio for games, London, UK.

Janer, J., Roma, G., & Kersten, S. (2011). *Authoring augmented soundscapes with user-contributed content*. ISMAR workshop on authoring solutions for augmented reality, Basel, Switzerland.

Jianyu, F., Miles, T., & Philippe, P. (2015). *Automatic recognition of eventfulness and pleasantness of soundscape*. Proceedings of the 10th audio mostly, Thessaloniki, Greece.

Jordanous, A. (2011). *Evaluating evaluation: Assessing progress in computational creativity research*. Proceedings of the second international conference on computational creativity, Mexico City.

Kallinen K., & Ravaja N. (2006). Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, *10*, 191–213.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010). *Music emotion recognition: A state of the art review*. Proceedings of the international symposium on music information retrieval, Utrecht, Netherlands (pp. 255–266).

Lamere P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, *37*(2), 101–114.
Malandrakis N., Potamianos A., Evangelopoulos G., & Zlatintsi A. (2011). *A supervised approach to movie emotion tracking*. Proceedings of the international conference on acoustics, speech, and signal processing, Prague, Czech Republic.
Matell M. S., & Jacoby J. (1971). Is there an optimal number of alternatives for likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.
McCartney, A. (2002). *Soundscape compositions and the subversion of electroacoustic norms*. The radio art companion (pp. 14–22). New Adventures in Sound Art. Retrieved from https://naisa.ca/radio-art-companion/soundscape-composition-and-the-subversion-of-electroacoustic-norms/
Minton S., Johnston M. D., Philips A. B., & Laird P. (1992, December). Minimizing conflicts: A heuristic repair method for constraint satisfaction and scheduling problems. *Artificial intelligence*, *58*(1-3), 161–205.
Moffat, D., & Kelly, M. (2006, August). *An investigation into people's bias against computational creativity in music composition*. The third joint workshop on computational creativity, Trento, Italy. ECAI 2006. Universita di Trento.
Morris, R., McDuff, D., & Calvo, R. (2014). Crowdsourcing techniques for affective computing. *The Oxford handbook of affective computing* (pp. 384–394). Oxford: Oxford University Press.
Niessen, M., Cance, C., & Dubois, D. (2010). *Categories for soundscape: Toward a hybrid classification*. InterNoise 2010, Lisbon, Portugal.
Pearce, M., & Wiggins, G. (2001). *Towards a framework for the evaluation of machine compositions*. Proceedings of the AISB'01 symposium on artificial intelligence and creativity in the arts and sciences, York, UK (pp. 22–32).
Pearse N. (2011). Deciding on the scale granularity of response categories of likert type scales: The case of a 21-point scale. *The Electronic Journal of Business Research Methods*, *9*(2), 159–171.
Pease, A., & Colton, S. (2011). *On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal*. Proceedings of the AISB symposium on AI and Philosophy, York, UK.
Pedersen, T. (2008). *The semantic space of sounds: Lexicon of sound-describing words*. Delta.
Freesound. (2012). Retrieved from http://www.freesound.org/
Roma, G., Herrera, P., & Serra, X. (2009). *Freesound radio: Supporting music creation by exploration of a sound database*. Computational creativity support workshop CHI09, Boston, MA.
Roma G., Herrera P., Zanin M., Toral S. L., Font F., & Serra X. (2012). Small world networks and creativity in audio clip sharing. *International Journal of Social Network Mining*, *1*(1), 112–127.
Roma G., Janer J., Kersten S., Schirosa M., Herrera P., & Serra X. (2010). Ecological acoustics perspective for content-based retrieval of environmental sounds. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 1–11.
Rossignol M., Lafay G., Lagrange M., & Misdariis N. (2014). Simscene: A web-based acoustic scenes simulator. *¡hal-01078098¿*.
Russell J. A., Weiss A., & Mendelsohn G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502.
Salamon, J., MacConnell, D., Cartwright, M., Li, P., & Bello, J. P. (2017). *Scaper: A library for soundscape synthesis and augmentation*. IEEE workshop on applications of signal processing to audio and acoustics, New Paltz, NY.
Schafer R. M. (1977). *The soundscape: Our sonic environment and the tuning of the world*. Destiny Books.
Scherer, K., Bänziger, T., & Roesch, E. (2010). *A blueprint for affective computing: A sourcebook and manual*. Oxford: Oxford University Press.
Serafin, S., & Serafin, G. (2004). *Sound design to enhance presence in photorealistic virtual reality*. ICAD, Sydney, Australia.
Shepard, M. (2007). *Tactical sound garden toolkit*. ACM SIGGRAPH 2007 art gallery, New York, NY, USA. SIGGRAPH '07 (p. 219). ACM.
Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice and sound effects in cinema*. Studio City: Michael Wiese Productions.
Symonds P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*(6), 456–461.
Thomas, N. G., Pasquier, P., Eigenfeldt, A., & Maxwell, J. B. (2013). *A methodology for the comparison of melodic generation models using meta-melo*. ISMIR, Curitiba, Brazil (pp. 561–566).
Thorogood, M., Fan, J., & Pasquier, P. (2015). *Bf-classifier: Background/foreground classification and segmentation of soundscape recordings*. Proceedings of the 10th audio mostly, Thessaloniki, Greece.
Thorogood M., Fan J., & Pasquier P. (2016). Soundscape audio signal classification and segmentation using listeners perception of background and foreground sound. *Journal of the Audio Engineering Society. Audio Engineering Society*, *64*(7/8), 484–492.
Thorogood, M., & Pasquier, P. (2013a). *Computationally generated soundscapes with audio metaphor*. Proceedings of the 4th international conference on computational creativity, Sydney, Australia (pp. 1–7).
Thorogood, M., & Pasquier, P. (2013b). *Impress: A machine learning approach to soundscape affect classification for a music performance environment*. Proceedings of the international conference on new interfaces for musical expression, Daejeon, Republic of Korea, May 27–30 (pp. 256–260).
Thorogood, M., Pasquier, P., & Eigenfeldt, A. (2012). *Audio metaphor: Audio information retrieval for soundscape composition*. Proceedings of the 6th sound and music computing conference (pp. 372–378).
Truax B. (1996). Soundscape, acoustic communication and environmental sound composition. *Contemporary Music Review*, *15*(1-2), 49–65.
Truax, B. (2001). *Acoustic communication* (2nd ed). New York, NY: Ablex Publishing.
Truax B. (2009). Island. In Soundscape Composition DVD. DVD-ROM (CSR-DVD 0901). Cambridge Street Publishing.
Truax B. (2012, November). Sound, listening and place: The aesthetic dilemma. *Organised Sound*, *17*, 193–201.

Valle, A., Schirosa, M., & Lombardo, V. (2009). *A framework for soundscape analysis and re-synthesis*. Proceedings of the SMC, Porto, Portugal (pp. 13–18).

Ventura, D. A. (2008). *A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems*. Proceedings of the 5th International Joint Workshop on Computational Creativity. Association for Computational Creativity, Madrid, Spain.

Wiggins G. A. (2006). Searching for computational creativity. *New Generation Computing*, *24*(3), 209–222.