Synthetic Ornithology: Machine learning, simulations and hyper-real soundscapes

Frederick Rodrigues
s222405968@deakin.edu.au
Deakin University
Melbourne, Victoria, Australia

Abstract

This paper presents *Synthetic Ornithology*, an interactive sound-based installation that uses machine learning to simulate sonic representations of localised Australian ecological futures, extending work in soundscape composition to engage in a speculative domain. Central to *Synthetic Ornithology* is a bespoke ML model, *Environmental Audio Generation for Localised Ecologies* (EAGLE), capable of generating high-quality, birdsong-focused soundscapes, up to 23 seconds in length. This paper outlines the development of the installation and how its design aims to influence audience perception of the sonic content of the work, extending established practices in NIME and sonic arts to a parafictional approach, and hyperreal aesthetics. Additionally, the paper examines the design and capabilities of the EAGLE model, and reflecting on how generative tools are positioned within a creative context, re-imagines the technical processes of training and configuring ML models as sites of artistic authorship in an expanded creative audio practice.

Keywords

soundscapes, machine learning, climate change, generative audio

1 Introduction

The rapid advancement of generative technology has sparked critical discussion among sonic artists about its role in creative processes, as well as motivations and deterrents to make use of it. Surveys indicate that generative audio in sonic arts remains underexplored, as musicians prefer tools that integrate with existing workflows, and avoid systems that generate complete compositions [17]. Generative tools like RAVE [4] and GANSpaceSynth [36], while available as plugins, require coding knowledge, specific resources for training, and offer limited audio quality and generation length, likely inhibiting their uptake [17]. Neutone FX[1], which acts as a host for generative audio models, while promising broad access to generative machine learning (ML) for sonic arts, has hard-coded limits on parameters for model control and conditioning data. Artists may be further deterred from generative tools by the "black box effect" of pre-trained models, a reluctance to deal with 'big data' models and datasets that induce issues of attribution, resource use, and unequal access to technology [15]. Developing bespoke models may address many of these issues, while also aligning with proposals for the subversive use of technology as outlined in the Critical Engineering Manifesto[2].

While such efforts become more common, there is little consensus on how this technology is seen within a creative process. In the NIME context, ML has been described as an 'autonomous instrument' [35] and even a collaborator [37]; however, as these tools continue to develop, each artist defines their role within their own practice.

Against this backdrop, *Synthetic Ornithology*, first presented publicly in March 2025, is an interactive installation that explores potential ecological futures by generating realistic birdsong-focused soundscapes using a bespoke ML model. The installation presents its sonic depictions of ecological futures as accurate; however, it is ultimately a work of speculative fiction. Visitors select a future time, location within Australia and climate conditions via a touchscreen interface, and the work generates a sonic representation of that scenario, along with generative text that contextualises the audio with social, environmental, activist and climate information. *Synthetic Ornithology* is one of a small but growing number of audio works where the entire sonic output comes from a generative model. As such, this research repositions the technical tasks required to design and train the model, including dataset curation, model architecture design, and conditioning data selection, as the loci for creative control. *Synthetic Ornithology* situates ML not as a tool, agent, or collaborator but as a mechanism for connecting audiences to inaccessible parts of the Earth system. This aligns with ideas from theorists like Latour who argues that technologies shape and transform human actions and perceptions [21] and Haraway, who proposes that technology mediates between worlds, and challenges traditional notions of agency [13]. This exploration into ML in creative practice adds to a growing body of research that will contribute to shaping future tools for sonic artists, as generative technology becomes more widespread.

Figure 1: A view of *Synthetic Ornithology* installed at Phoenix Gallery Melbourne, March 2025.

[1] https://neutone.ai/fx
[2] https://criticalengineering.org

NIME '25, June 24–27, 2025, Canberra, Australia

2 Related work

Synthetic Ornithology draws on a web of creative and technical research covering soundscape composition, previous NIME research utilising soundscapes and ML, as well as ML in a technical context.

Soundscape composition embraces environmental recordings as source material, and contextualises them using musical terminology, ultimately seeing soundscapes as discrete musical compositions. Soundscape composers often produce works that aim to engage with issues connected with the impacts of climate change [38]. The field has evolved significantly from documentation, through creative outputs using basic editing and playback [38], to engaging with immersive media [3]. Soundscapes reflect ecosystem processes and human activities [27] and as a medium 'can be employed to sense more than-human environmental change, bringing human and nonhuman bodies into proximity with events over vast geographical registers' [14][p.7], 'remapping the complex spatialities and temporalities of climate change' [14] [p. 5] and can facilitate the 'sonic transmission of meanings about place, time, environment' [40][p. 52]. This broad potential, supported by research into affective responses to features and changes in soundscapes [30], makes the acoustic environment an intriguing medium for many sonic artists .

Previous soundscape-focused works presented in NIME have investigated interactive analysis and composition of soundscapes [5], soundscape composition using concatenative synthesis [18] real-time musical synthesis from soundscape timbres [2] and the use of climatic data to influence granular synthesis for speculative sonic environments [26]. Cumulatively, these works, in their use of location and climate data, engagement with ecological issues and synthesis of soundscapes provide a lineage of developments in the NIME context that *Synthetic Ornithology* builds on.

ML has been engaged in numerous NIME studies, often focused on generating, recognising and manipulating control or gestural data [8, 28]. Music generation with ML has been explored in several approaches including score and note generation [22, 25], rhythm generation [24, 39], control of synthesis [16, 42] and the use of generative audio through existing tool-kits [32, 33]. As ML technology has advanced and been increasingly utilised, some artists have developed works where the entire sonic content is generated by ML. For example, *The Wandering Mind* [9] utilised Google's YAMNet[3] model, trained on 70,000 field recordings to create a work wholly based on generated content.

Configuring or modifying and training bespoke models for generative audio remains an emerging approach in sonic arts. The technical capabilities of an ML model impact the duration and quality of generated sound, methods of controlling output, time required for generation, and media and resources used for training. When a model is developed within and as artistic practice, these parameters influence the creative methodology, as much as creative needs influence the development of the model. While it may seem preferable to have faster models and higher quality audio, in creative practice, deeply engaging with technology and how it can contribute to a work overshadows technical limitations. For example, *Deepscape: Transversal* [31] used a modified version of the RAVE model to create a work that also featured entirely ML generated sonic content focused on planetary soundscapes. As discussed, RAVE has limited quality output but is well suited for the real time generation used by *Deepscape: Transversal*, and the work's artistic framework contextualises the output as an otherworldly exploration of computational infrastructures. Similarly *AI-terity* [35] utilised a bespoke GAN based model, *GanspaceSynth* [36], specifically developed for live performance as part of a NIME. *GanspaceSynth* generates only 4-second, 16khz audio, but at very high speed. *AI-terity* developed methods for the artist to traverse the latent space of the model in real time, and its implementation prevents continuous audio generation from a fixed latent, making the limited length of output irrelevant. Although 16khz audio is considered 'low' quality, post-processing of the output, part of the works methodology, ultimately rendered audio fidelity a non-issue.

This growing lineage of creative works using generative ML, in addition to providing a background for this research, illustrates how technology-driven innovations (for example GANsynth developed by Google), are adopted and reconfigured by pioneering artists (e.g. AI-terity's adaptation of GanSynth into GanSpaceSynth), then spread to a broader creative community (as evidenced by numerous works presented in NIME that utilise RAVE) and ultimately integrate into widely used audio software tools (exemplified by Neutone FX, a commercial plugin that hosts generative models).

Future developments in generative ML may result in flexible technology that integrates seamlessly with existing tools, lessening the need for custom model development. However, generative audio is still an emerging technique in sonic art. Consequently, continued publication and discussion of ML development in and as creative practice will uncover its artistic potentials and help shape future tools. Additionally, in the context of sonic arts, the conceptual 'universe' of a work plays an important role in contextualising a generative model's output, shaping how audiences interpret and engage with work, as can be seen in *Deepscape: Transversal* and *The Wandering Mind*. As such the following section first details the technical development of the EAGLE model and how the choices made in this process influence the creative work and follows with a discussion of how *Synthetic Ornithology* presents and contextualises its sonic output.

3 The *Synthetic Ornithology* System

Generative audio is a growing application of ML, and its task of generating sequences of cumulatively dependent values of a sound wave's amplitude is complex. Recent advances in the field have facilitated the efficient generation of high-quality, variable length audio that adheres well to conditioning data [7, 10, 29]. Despite these efforts, there is a gap in research generating soundscapes based on localised climate scenarios as required by *Synthetic Ornithology*. This unique gap is unlikely to be addressed by ML research, as it directly responds to the needs of a single creative work; it is therefore addressed through this practice-led research, configuring and training a bespoke ML model.

3.1 Machine learning Model Architecture

In *Synthetic Ornithology*, the EAGLE model is the sole source of sonic material; consequently, model architecture, conditioning data selection, dataset curation, and training parameters become the principal levers for influencing the sonic output. Different model architectures can drastically influence creative outcomes. GANs, for instance offer a diverse output potential but can be difficult to train. VAEs can be more stable and resource-efficient but have limited variety in their generated audio; the RAVE toolkit [4] exemplifies these trade-offs: its output may be less varied

[3] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

compared to a GAN-based model, yet its fast, low resource generation is ideal for live performance. This research opted to focus on long (23 second) high quality, 44.1 kHz audio (sacrificing stereo content to halve the resource use), and multiple numeric conditioning parameters, extending an existing approach [10] originally designed for text to audio generation.

The EAGLE model architecture has two key components: an audio encoder/decoder model, a Generative Adversarial Network with Residual Vector Quantisation (RVQGAN) based on the Descript Audio Codec [19], and a diffusion transformer, featuring numeric conditioning via a data embedder, and utilising cross-attention and Classifier-Free Guidance. This combination facilitates relatively long audio generation (23 seconds) at high-quality (44.1 kHz mono), with high temporal consistency, i.e. over the 23 seconds, the audio events and features resemble real-world recordings.

The RVQGAN acts like a data compression system by organising audio with similar features in closer proximity in the latent space. Given a 'location' in this latent space, the RVQGAN can output an audio file with the features that correspond to that location. During training, audio files are encoded into a latent representation (with a much smaller data footprint than raw audio, increasing efficiency), which is then sent to the diffusion transformer, along with the conditioning data. This multi-model architecture, where a diffusion model operates on latent representations of data instead of the raw data is referred to as latent diffusion. The diffusion mechanism iteratively adds a small amount of noise to the data and records the stepwise transformation between the input and each noised iteration; this process is repeated until only pure noise remains. In the generation process, starting with pure noise, the diffuser applies transformations that correspond to the given conditioning data in reverse, until the resulting latent representation is sent to the RVQGAN and 'decompressed' to audio. Long audio files are broken into sequences of smaller segments (conditioned with their start time in the source file and duration); the transformer part of the diffusion transformer allows these segments to be operated on in parallel, increasing efficiency as well as temporal coherence across long audio files.

EAGLE uses a cross-attention mechanism (ensuring the consistent application of conditioning data across diffusion steps and sequences) to apply conditioning data with a Classifier Free Guidance (CFG) scale. The CFG scale can modulate the influence of conditioning data during generation, lower CFG values allow the model greater freedom to deviate from generating audio that correlates from the conditioning data, while higher values enforce stricter adherence to it.

Figure 2 illustrates the EAGLE architecture during generation. On the left the conditioning parameters are applied to gaussian noise. The diffusion transformer then iteratively applies learned transformations aligned with the conditioning data, resulting in a latent representation, that is passed to the RVQGAN, where it is transformed to audio.

3.2 Dataset and Conditioning Data

This research required curating a bespoke dataset of birdsong-focused soundscapes from across Australia; each entry consists of an audio file accompanied by a metadata file. The final dataset consisted of 44,804 entries, approximately 710 hours of audio[4].

[4]The dataset is visualised online at https://audioweather.com

3.2.1 Dataset curation methodology. Empirical research using soundscapes preferences capture using high-quality fixed, and spatially aligned microphones, recording remotely for long periods and devoid of human presence [1]. This research sourced soundscape recordings from existing archives, whose content does not align with these requirements. Captured largely by citizen scientists, recordings in these archives, rather than unbiased soundscapes, are birdsong-focused recordings that also capture the surrounding sonic environments. Entries vary in length, are recorded using handheld devices, lack consistent microphone placement and frequently include evidence of human presence such as footsteps, talking and clothing rustles. The inclusion of artefacts of human presence and from variations in capture quality and techniques, in the dataset used for training, means that these artefacts also appear in the model's generated output.

Synthetic Ornithology embraces these artefacts primarily as they make the sound more relatable. Footsteps and clothing rustles place a human in the audio scenario, giving the listener a presence to substitute themselves into. This resonates with Feld's voicing, part of a reciprocal process that connects to the listener's sense of self and to an embodied experience of place [11]. Additionally, the colouring of audio from small recording devices connects to a more intimate experience of sound. Pristine soundscapes are found in cinema and documentary, far removed from everyday experiences. Smartphone or action-cam recordings are more likely found in our messages, social media or communication from family and friends. *Synthetic Ornithology*, aiming to connect with personal sonic experiences of place, embraces the mediation of small recording devices and non-professional techniques and equipment.

3.2.2 Sources. The dataset was sourced from xeno-canto[5] and the Macaulay Library[6]. xeno-Canto's library has a Creative Commons licence and recordings from the Macaulay Library were accessed through a negotiated agreement with the custodians. Both libraries have varying metadata attached to entries, for this research, only entries from Australia and with complete timestamps and locations were used.

3.2.3 Pre-processing. To minimise variation in format and recording levels, all recordings were transposed to 16-bit 44.1 kHz WAV format, and for all non-mono entries, only the first channel was retained. A DC offset removal filter, a high-pass filter at 60 Hz with a 24 dB/octave roll off, and normalisation to -0.1 dBFS were applied to all entries.

3.3 Metadata/conditioning data

The metadata file for each entry contains a timestamp, geographic coordinates, as well as climate conditions from the time and location of the recording. This metadata is used as conditioning data when training the model and is required to generate new audio. The selection of parameters was guided both by research detailing connections between changes in climate and variations in soundscape and birdsong [6, 27, 34], as well as the conceptual framework of the work. *Synthetic Ornithology* proposes to generate localised soundscapes from future climate scenarios, and the model was designed to reflect this capability. While a model that could generate highly accurate soundscapes from climate scenarios might use the same structure and set of conditioning

[5]https://xeno-canto.org
[6]https://www.macaulaylibrary.org

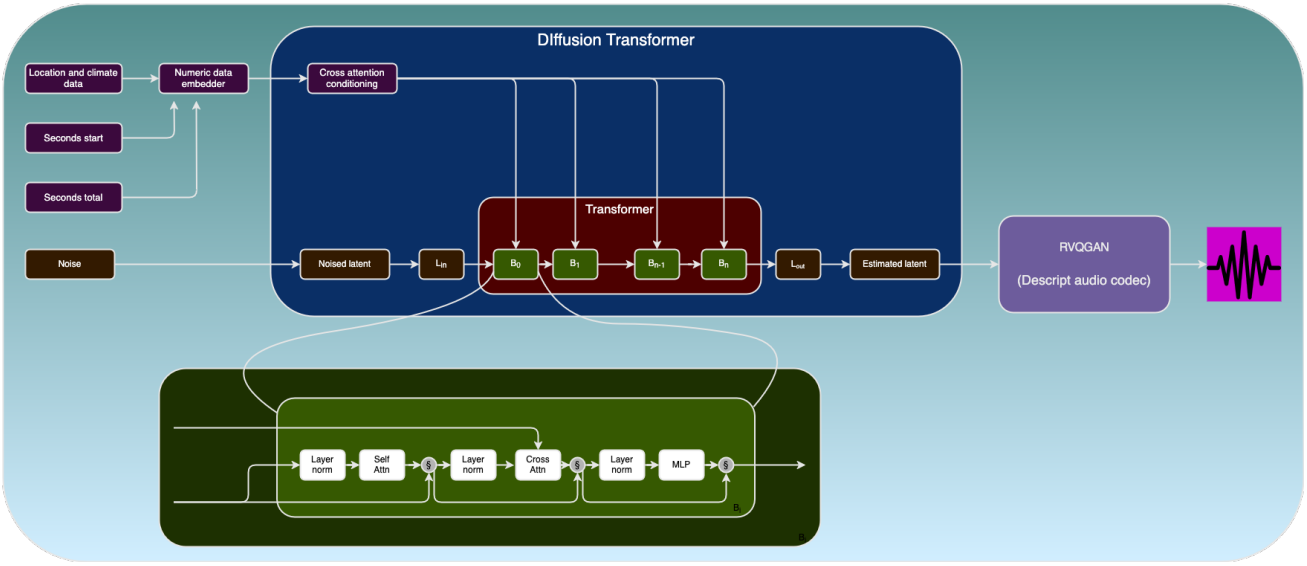

Figure 2: A flow diagram of the EAGLE model architecture for generating audio.

data, *Synthetic Ornithology* is a creative, speculative work. Although the EAGLE model does utilise correlations between audio features in the training data and the conditioning data, the influence of future climate change on localised soundscapes will likely depend on many more factors not accounted for by the EAGLE model. In satisfying the artistic requirements of the work, the final metadata selection was: latitude, longitude, temperature, humidity, wind speed, pressure, minutes of day and day of year (to represent seasonal and diurnal variations). Climate data for each entry was collected via the OpenWeatherMaps[7] API and time of day and day of year metadata was derived from each recording's timestamp.

Table 1 shows the final metadata selection, the minimum and maximum values of each parameter, as well as the mean and standard deviation based on their normalised values.

Table 1: The final chosen streams of metadata used to condition the model and their minimum, maximum, format, mean and standard deviation.

Parameter	Min	Max	Format	Mean	Std dev.
Latitude	-54.61	-10.13	double	0.569	0.192
Longitude	96.82	167.96	double	0.662	0.143
Temperature	-10.0	55.0	float	0.469	0.145
Humidity	0.0	100.0	float	0.722	0.189
Wind speed	0.0	50.0	float	0.122	0.076
Pressure	800	1200	float	0.537	0.086
Minutes of day	0.0	1439	float	0.398	0.206
Day of year	1	366	integer	0.594	0.280

3.4 Training and evaluation

EAGLE's multi-model architecture required separate training of the RVQGAN and the Diffusion Transformer. While pre-trained models of the RVQGAN are available, these use a fixed latent output size, reducing flexibility in subsequent steps designing the diffusion model. The RVQGAN was trained using randomly cropped 1-second audio segments for 200000 steps, approximately 320 GPU hours. Evaluation occurred every 2500 steps, where three audio segments were output each with the original audio and round-trip encoded and decoded audio for comparison. The Diffusion Transformer was trained for 640,000 steps, approximately 2800 GPU hours. Evaluation occurred every 2500 steps, generating 3 samples from a set of operator selected conditioning parameters, to assess quality and coherence. Notably, earlier training iterations produced outputs that, while lower fidelity, offered intriguing creative textures[8].

As *Synthetic Ornithology* relies on generating believable soundscapes to engage audiences, the output of the EAGLE model was evaluated to confirm its fidelity and perceptual realism. This evaluation was conducted using Mean Opinion Scores (MOS) surveys, a commonly used metric to understand the perceptual quality of a generative audio model.

The survey began by filtering participants based on their familiarity with birdsong or audio, to determine if professional perspectives differed from those of non-professionals. All participants were then played 15 randomly selected audio samples from a set of 9 real and 19 generated soundscapes. For each audio sample, participants responded to five questions using a Likert scale answer ranging from "Strongly Disagree" to "Strongly Agree." The questions were as follows:

(1) "The sounds in this recording appear natural and lifelike."
(2) "This audio recording creates a sense of being in a real environment."
(3) "This audio is real and not generated by an artificial intelligence model."
(4) "The audio in this recording is pleasant to listen to."
(5) "The audio in this recording is of high-quality and free from artefacts."

Likert responses were converted to numeric values between 1 to 5 for analysis; for all questions a higher number represented a more positive response to the audio. Responses from all listening events were separated based on whether they corresponded to a real or generated file, and then average scores for real and

[7]https://openweathermap.org

[8]https://fred-dev.github.io/Synthetic_ornithology_results/errors.html

generated samples were calculated. With 37 respondents, the survey collected data from 555 listening events.

3.4.1 Survey results. Table 2 presents the results of the MOS survey. The similarities in responses to real and generated samples suggest that the model's output is comparable to real recordings confirming that the model achieves both fidelity and realism. Professional respondents demonstrated similar responses to on professionals, with the generated audio consistently scoring very close to the real samples.

Table 2: Qualitative results from an MOS survey on the model's audio output.

	Section	Naturalness	Environment	Realness	Pleasantness	Quality	Overall
All	**Real**	4.14	4.14	3.58	3.37	3.58	3.76
All	**Generated**	3.94	3.95	3.47	3.58	3.30	3.65
Pro	**Real**	4.24	4.46	3.70	3.91	4.12	4.08
Pro	**Generated**	4.02	4.10	3.62	3.80	4.02	3.91

3.5 Model deployment

To use the model to generate audio for the interactive installation, the model and software stack need to be running on suitable hardware and wrapped in a suitable software layer. The model was deployed on a Hugging Face[9] virtual server, and wrapped with Gradio[10], a Python-based library that facilitates interaction with generative models. This allowed for the development of the web-based user interface, independently of the generation system, and also removed the need for specialised hardware on location to run the installation.

4 Installation and Interaction Design

Synthetic Ornithology, like many works of sonic art, has a strong conceptual framework that proposes how audiences might perceive the work; here the artistic concept proposed that the audio generated by the installation is an accurate prediction of future soundscapes under pressures from climate change. As discussed earlier, while the model architecture used in the work may be a plausible way to achieve this goal; the work is speculative, and plays on plausibility, rather than dealing with accuracy. As such the installation and the interface design focus on presenting the work as a plausible simulation, while also leaving several clues that the work is speculative. This approach resonates with hyperreal and parafictional artistic practices, most commonly seen in visual arts, discussed in more detail in the following section.

In practical terms, the installation's design evokes the aesthetics of a science exhibit: no visible cables, a clean space with unobtrusive grey couches, and minimal visual clutter. This is reinforced by a printed, academic-style text describing the use of soundscapes in climate change impact analysis, Australia's unique and threatened bird populations and the installation's ability to model the relationship between soundscapes and climate. The text positions the installation as a plausible scientific display rather than artistic. An AI-generated image of a bird flying over a burning Earth points to the speculative nature of the work.

4.0.1 Installation flow. The installation operates in a drift mode when left untouched for a few minutes. In this mode the system randomly plays pre-generated soundscapes from 1000 speculative scenarios, displaying the future dates, times, locations and weather conditions that were used to generate each file. In this uncanny audio non-existent people are heard walking through generated landscapes capturing the sounds of unknown futures, sometimes interrupted by unintelligible utterances, remnants from fragments of speech in the training dataset. Audiences are invited to interrupt the drift mode, and use the touchscreen to generate speculative soundscapes from locations and conditions of their choosing. Here, the interaction with *Synthetic Ornithology* unfolds through a step-by-step interface that lets users choose a location, date, and climate conditions to generate their own speculative soundscape. The design and flow of the interaction is aimed at maintaining the sense of plausibility, while an animated bird at loading steps hints at the work's speculative nature.

Figure 3: An image of the touch screen for user interaction and instruction text from the *Synthetic Ornithology* installation.

4.0.2 Interface design Methodology. Instrument interface design often focuses on facilitating embodied interaction or giving artists and audiences access to otherwise inaccessible systems. In contrast, *Synthetic Ornithology* foregrounds how interactive elements shape listeners' perception of the generated audio, an underutilised approach in the NIME context. *Synthetic Ornithology* uses known UI elements, representations and interaction patterns that the audience has likely been exposed to, to frame the installation as a plausible, accurate simulation, enhancing affective impact [12]. While such simple interface elements appear on many audio tools and even public interfaces, *Synthetic Ornithology's* adoption of these elements for its entire interface is an uncommon choice in sonic arts installations. A neutral map interface and an implementation of Apple's mobile UI date selector are used, leveraging their credibility, perceived neutrality and practical utility [23], as well as audience's previous experience of interacting with these elements resulting in factual outcomes. Through this interface, users choose a future scenario, which is further grounded by generated text that aggregates information about the location, date, time, and climate conditions that are used to seed the soundscape generation.

Familiarity and affect . In *Synthetic Ornithology*, listeners are prompted to hear the generated output less as purely artistic and more as a genuine forecast, engaging with it in relation to real-world experiences rather than comparing it to other creative works. This facilitates an affective response termed 'appropriateness'; Schulte-Fortkamp et al. explain that 'since an encountered situation is usually matched against existing cognitive schemes,

[9]https://huggingface.co
[10]https://www.gradio.app

appropriateness viewed as the level of congruency between a scheme and a real-world situation will influence positive affective responses. Inappropriate matches consequently lead to negative affective responses' [30][p. 36]. Such contemplative comparisons are further enhanced when the audience has an existing sonic experience relating to their chosen scenario. In practice, this hinges on the audience choosing personally significant parameters (e.g. a familiar location or meaningful date), and the likelihood of doing so, especially with a familiar user interface, is supported by Zajonc's 'mere exposure' principle [41].

Because place is a fundamental reference point for recalling past sonic experiences, the interface's main viewpoint is a map, built with LeafletJS[11]. Users pan, zoom, and scroll with familiar gestures to find their desired location. Long-pressing on a location drops a marker at that spot, storing the first parameter of the conditioning data, GPS coordinates. A popup bubble then appears and prompts the audience to select a future date (limited to a 20-year span). The interface for selecting the year, month, day, and time uses Apple iOS style rolling wheels, while familiar next and back buttons guide users through the steps. Once the date is selected, the user is then prompted to specify climate conditions (e.g. temperature, humidity, wind speed), starting with pre-filled suggestions based on real weather data from the past year for that location, date, and time, varied depending on the future year. These suggestions are presented in an interface that is based on common weather forecast widgets. Once finalised, this data is sent to the remote server to trigger soundscape generation, a process that takes roughly 23 seconds.

Generative text. While the soundscape generation is triggered, the user selected scenario is also sent to the ChatGPT API[12] to create a generative text element that serves multiple purposes: to engage the visitor while the audio is generated, and to provide an information scaffold that influences audience perception of the generated soundscape. The text is generated using a prompt that incorporates the audience selected data. The prompt gathers localised information on the social makeup of the location, local species, prominent, endangered and extinct, industrial activity such as mining or deforestation, environmental protests, and the extent to which the chosen climate conditions deviate from historical norms for that time and date. This information is then woven into a short, cohesive text bridging the present with the chosen future date, sometimes introducing fictional elements. The prompt directs that fictional elements are to be extrapolations of existing gathered information, for example where there have been protests in the past, a fictional element may be similar protests on a future date. By constructing a sequence of events that frame the audience's chosen scenario, the text contributes to the perceived plausibility of the work, the last part of the interactive experience before the soundscape is generated and automatically played.

Importantly, while ChatGPT and other language models are not always reliable sources of information, in keeping with the speculative nature of *Synthetic Ornithology*, the impact of this information was considered to be sufficiently useful despite inconsistencies. While a true evaluation of the accuracy of the text content is beyond the scope of this research, a substantial number of outputs were cross checked and found to be sufficiently reliable for the nature of the work. Due to the large variety of possible input conditions, it is likely that a proportion of the generated text will contain inaccuracies, however this is not considered detrimental to the work as whole.

Figure 4 shows the flow diagram of the user interaction from the initial map, through the date, time and climate selection to the resulting generated text and audi playback.

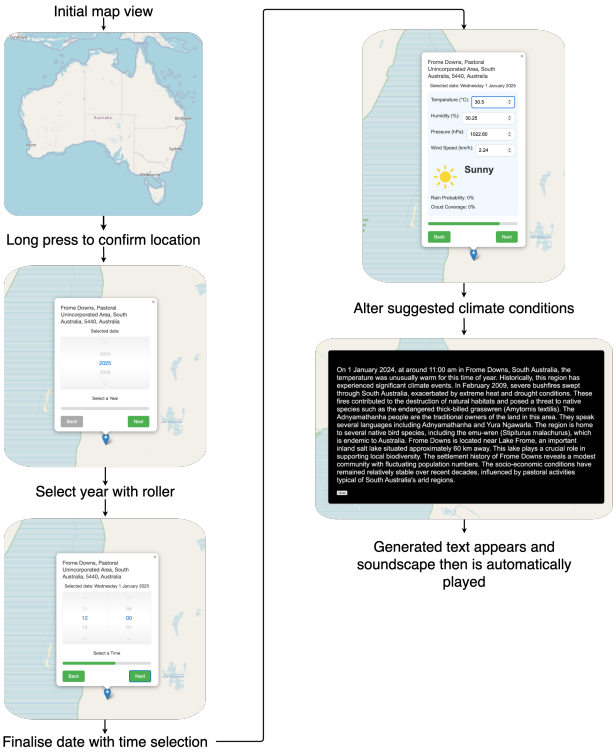

Figure 4: A flow diagram of the interaction process of the touch screen interface for *Synthetic Ornithology*.

5 Artistic and Theoretical Contributions

Synthetic Ornithology's use of a generative model that creates climate aligned speculative soundscapes fills an under-explored gap in soundscape composition and allows the field to operate in previously unavailable speculative and predictive modalities. The capabilities of the ML model also alter the materiality of the recorded soundscapes, from singular representations of a place and time to a malleable set of component audio features and their relationship to the landscape and climate. This grants the flexibility to orchestrate these features, akin to manipulating discrete musical sources, while retaining the authenticity of actual field recordings and the connection to real environmental events. Within the conceptual framework of *Synthetic Ornithology*, the model's facilitation of soundscape ecology to work in this speculative mode is viewed through the lens of posthuman philosophy. Here, ML, rather that adding to the capabilities of the artist or framework, it is seen to extend the 'being-ness' of the audience; allowing them to experience unheard futures and have a sensory experience of climate impacting discrete biotic responses to changing environments.

The EAGLE model, while not specifically designed to decompose soundscapes into discreet audio features, produces outputs that resonate with such capabilities. The composition of audio features by EAGLE in response to climate scenarios is the centre of this work's aesthetic potential. For example, recordings from a

[11] https://leafletjs.com
[12] https://openai.com/api

particular region at sunset in summer frequently include cockatoos and cicadas; accordingly, if the user selects a nearby location and a sunset-in-summer time frame, the generated soundscape often contains those species. While EAGLE readily captures these location and time-based patterns, its creative possibilities arise when the model diverges from expectations. Here the sonic vocabulary of the work is apparent; *biophonic* sounds not usually heard together appear in the same soundscape, one bird species sings the call of another, and familiar bird calls appear with unexpected variation. A series of comparative samples from the model output and the dataset can be heard online[13].

ML models, such as EAGLE, with their ability to deduce and reproduce complex patterns and correlations in data are often used for simulations. *Synthetic Ornithology* employs EAGLE to simulate the effects of speculative climate conditions on soundscapes, however, prioritises believability over scientific accuracy. This approach bridges scientific simulations and speculative fiction; where both embody the rules of a universe and present the outcomes of the application of those rules. While creative works that use simulation are not uncommon, audio-focused approaches are relatively unknown.

Synthetic Ornithology's highly realistic audio, confirmed through surveys, further aligns with hyperrealism, usually associated with visual art. Hyperrealist works like Patricia Piccinini's 'The Instruments of Life'[14] use realism to engage audiences and exaggerations to highlight external issues. *Synthetic Ornithology's* hyperrealist audio grounds the speculative scenarios in familiar experiences, creating an initial sense of plausibility. The subtle and sometimes impossible variations output by the model, like distortions of birdsong, and impossible combinations of biotic sounds, highlight possible futures of environmental degradation. These distortions, like the exaggerations of hyperrealist visual art, do not detract from realism but instead amplify a capacity to provoke reflection on humanity's relationship with the non-human world. However, unlike hyperrealist works, generally identified as 'fiction' through their presentation, *Synthetic Ornithology* is presented as ambiguously accurate, introducing a subtle misdirection, situating the work within parafictional art. Parafictional art engages with narratives that are presented as true, with an aim to 'softly' deceive the viewer (or listener) [20] and are realised through framing a work of fiction as plausibly true. *Synthetic Ornithology* is unique in engaging with hyperreal and parafictional approaches from an audio focused practice.

6 Ethical and environmental considerations

This work relies on a large number of recordings made by individuals that may not have considered the use case put forward by this research. While the licensing of all the material used in the dataset allows for its use in this work, many recordings were submitted when such a use case was not possible.

This research also exhibits a critical tension between the resource-intensive nature of ML development and the ecological concerns that this project addresses. Although the resources allocated to this research are negligible in comparison to commercial projects, they remain significant. The training and deployment of the model, if conducted on commercial cloud computing platforms, would have resulted in an estimated 1,536 kg of CO2 emissions, equivalent to the carbon footprint of a one-way economy flight from Sydney to London for a single passenger. While the computational resources provided by the National Computational Infrastructure (NCI) used in this research are powered by carbon-neutral renewable energy, and the NCI has a commitment to zero-emissions compute resources, this does not amount to this research being totally environmentally friendly.

While the number of training hours may seem high for this research, for some comparison, information on training a RAVE model (commonhly used in MIME research) from the *Institut de recherche et coordination acoustique/musique* (IRCAM)[15] suggests 2 phase training may take up to 600 hours with an unknown dataset size (an initial 3- 4 days for phase 1 and up to 3 weeks for phase 2). In total this is about 20% of the resources used for this research. However, the EAGLE model features significant advances in audio quality, generation length and artistic flexibility (through its conditioning capabilities). While the EAGLE model is certainly more complex and resource intensive than lightweight systems, a direct comparison is difficult to make without training the two models on the same dataset and accounting for the differences in output. The resource use of this research is justified by the artistic requirements of this project, the availability of carbon-neutral computing, and in reflection the relative use of resources used – the carbon equivalent of a one-way economy flight from Sydney to London for a single passenger – a commonly used route for academics and researchers presenting at international conferences to and from Australia.

7 Conclusion

In this paper, the key developments that underpin *Synthetic Ornithology* were introduced. The primary effort of configuring and training the EAGLE model resulted in its ability to produce realistic soundscapes of up to 23 seconds in length, and to be the sole source of audio in this work. This effectively repositions dataset curation, model design, and training parameters as components of the creative process, rather than technical tasks. The unprocessed generated audio foregrounds the system's creative possibilities, combining *biophonic*, *geophonic*, and *anthropophonic* features in unforeseen ways. This approach aligns with contemporary theories on post-anthropocentric perspectives in sonic arts, and positions ML as an apparatus for enhancing audience's sensorial experiences.

The interaction design, employing familiar user interface elements to shape the audience's perception of the generated audio, evoking trust and plausibility, situates the output in a speculative yet believable realm. The resulting positioning of the work as hyperreal and parafictional, under-explored contexts for sonic arts, presents the potential for machine learning based generative techniques to facilitate new artistic modalities.

The presentation of this research hopes to contribute to a growing collection of such reflections, and ultimately contribute usefully to future artistic work, and shape ML tools that will likely emerge as common in future creative audio tool-kits. Ultimately, this paper's methodology, findings, and reflections illustrate the potentials available to sonic artists in embracing the development of generative ML within creative practice.

8 Ethical Standards

Before the survey was released online this research passed ethics clearance at Deakin University Australia (reference HAE-22-077). Participants were self selecting and the survey began with an

[13] https://fred-dev.github.io/Synthetic_ornithology_results/comparison.html
[14] https://www.patriciapiccinini.net/a-show.php?id=2021-Tallinn
[15] See the warning under the 'Preparing the training' section on this tutorial page: https://forum.ircam.fr/article/detail/training-rave-models-on-custom-data

ethics and consent statement which participants viewed before completing the survey. The survey was fully anonymous.

Acknowledgments

This work was developed at Deakin University Australia and further supported by the Critical Digital Infrastructures and Interfaces research group. Computational resources were provided by the National Computational Infrastructure[16], part of the Australian National University.

References

[1] ISO/TC 43/SC 1. 2018. Acoustics—Soundscape Part 2: Data Collection and Reporting Requirements. https://www.iso.org/standard/75267.html

[2] Lior Arbel. 2021. Aeolis: A Virtual Instrument Producing Pitched Tones With Soundscape Timbres. In *NIME 2021* (Shanghai, China, 2021-06-01). PubPub. https://doi.org/10.21428/92fbeb44.64f66047

[3] Leah Barclay. [n. d.]. Sounding Extremes: Ecological Sound Art in the Anthropocene. 32, 2 ([n. d.]), 37–44. https://doi.org/10.7202/1091901ar

[4] Antoine Caillon and Philippe Esling. 2021. *RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis.* arXiv:2111.05011 [cs, eess] http://arxiv.org/abs/2111.05011

[5] Souhwan Choe and Kyogu Lee. 2011. SWAF: Towards a Web Application Framework for Composition and Documentation of Soundscape. In *Proceedings of the International Conference on New Interfaces for Musical Expression.* Oslo, Norway, 533–534. https://doi.org/10.5281/zenodo.1177985

[6] C.M. Coomes and E.P. Derryberry. 2021. High Temperatures Reduce Song Production and Alter Signal Salience in Songbirds. *Animal Behaviour* 180 (Oct. 2021), 13–22. https://doi.org/10.1016/j.anbehav.2021.07.020

[7] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. arXiv:2306.05284 [cs, eess]

[8] Marcel O DeSmith, Andrew Piepenbrink, and Ajay Kapur. 2020. SQUISHBOI: A Multidimensional Controller for Complex Musical Interactions Using Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Birmingham, UK, 2020-07), Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, 353–356. https://doi.org/10.5281/zenodo.4813412

[9] Gershon Dublon, Xin Liu, Nicholas Gillian, and Nan Zhao. [n. d.]. The Wandering Mind: Planetary Scale Dreaming in Latent Spaces. In *NIME 2021* (Shanghai, China, 2021-06-01). PubPub. https://doi.org/10.21428/92fbeb44.56a514bb

[10] Zach Evans, C. J. Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. 2024. *Fast Timing-Conditioned Latent Audio Diffusion.* arXiv:2402.04825 [cs, eess] http://arxiv.org/abs/2402.04825

[11] Steven Feld. 2012. *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression, 3rd Edition with a New Introduction by the Author.* Duke University Press. https://doi.org/10.2307/j.ctv113180h jstor:10.2307/j.ctv113180h

[12] Gary P Hampson and Matthew Rich-Tolsma. 2015. Transformative Learning for Climate Change Engagement: Regenerating Perspectives, Principles, and Practice. 11, 3 (2015).

[13] Donna Haraway, Thyrza Nichols Goodeve, and Lynn Randolph. 1997. *Modest_witness Second Millennium: FemaleMan®_meets_OncoMouse™$dfeminism and Technoscience* (second edition ed.). Routledge, Taylor & Francis.

[14] Harriet Hawkins and Anja Kanngieser. 2017. Artful Climate Change Communication: Overcoming Abstractions, Insensibilities, and Distances. *WIREs Climate Change* 8, 5 (2017), e472. https://doi.org/10.1002/wcc.472

[15] Théo Jourdan and Baptiste Caramiaux. 2023. Culture and Politics of Machine Learning in NIME: A Preliminary Qualitative Inquiry. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Mexico City, Mexico, 2023-05), Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Article 47, 332–338 pages. https://doi.org/10.5281/zenodo.11189202

[16] Hernán Kerlleñevich, Manuel C. Egua, and Pablo E. Riera. 2011. An Open Source Interface Based on Biological Neural Networks for Interactive Music Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Oslo, Norway, 2011). 331–336. https://doi.org/10.5281/zenodo.1178063

[17] Shelly Knotts and Nick Collins. 2020. A Survey on the Uptake of Music AI Software. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 499–504. https://doi.org/10.5281/zenodo.4813499

[18] Stratos Kountouras and Ioannis Zannos. 2017. Gestus: Teaching Soundscape Composition and Performance with a Tangible Interface. In *Proceedings of the International Conference on New Interfaces for Musical Expression.* Aalborg University Copenhagen, Copenhagen, Denmark, 336–341. https://doi.org/10.5281/zenodo.1176274

[19] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. *High-Fidelity Audio Compression with Improved RVQGAN.* arXiv:2306.06546 [cs, eess] http://arxiv.org/abs/2306.06546

[20] Carrie Lambert-Beatty. 2009. Make-Believe: Parafiction and Plausibility. 129 (2009), 51–84. https://doi.org/10.1162/octo.2009.129.1.51

[21] Bruno Latour. 1994. On Technical Mediation. 3, 2 (1994), 29–64. https://sciencespo.hal.science/hal-02057233

[22] Jeffrey A. T. Lupker. 2021. Score-Transformer: A Deep Learning Aid for Music Composition. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Shanghai, China, 2021-06). Article 59. https://doi.org/10.21428/92fbeb44.21d4fd1f

[23] Donald A. Norman. 2013. *The Design of Everyday Things* (revised and expandes editons ed.). The MIT Press.

[24] Thomas Nuttall, Behzad Haki, and Sergi Jorda. 2021. Transformer Neural Networks for Automated Rhythm Generation. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Shanghai, China, 2021-06). Article 33. https://doi.org/10.21428/92fbeb44.fe9a0d82

[25] Torgrim Rudland Næss and Charles Patrick Martin. 2019. A Physical Intelligent Instrument Using Recurrent Neural Networks. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Porto Alegre, Brazil, 2019-06), Marcelo Queiroz and Anna Xambó Sedó (Eds.). UFRGS, 79–82. https://doi.org/10.5281/zenodo.3672874

[26] Eleni-Ira Panourgia, Bela Usabaev, and Angela Brennecke. 2024. ClimaSynth: Enhancing Environmental Perception through Climate Change Sonic Interaction. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Utrecht, Netherlands, 2024-09), S M Astrid Bin and Courtney N. Reed (Eds.). Article 75, 516–520 pages. https://doi.org/10.5281/zenodo.13904937

[27] Bryan C. Pijanowski, Almo Farina, Stuart H. Gage, Sarah L. Dumyahn, and Bernie L. Krause. 2011. What Is Soundscape Ecology? An Introduction and Overview of an Emerging New Science. 26, 9 (2011), 1213–1232. https://doi.org/10.1007/s10980-011-9600-8

[28] Jan C. Schacher, Chikashi Miyama, and Daniel Bisig. 2015. Gestural Electronic Music Using Machine Learning as Generative Device. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Baton Rouge, Louisiana, USA, 2015-05), Edgar Berdahl and Jesse Allison (Eds.). Louisiana State University, 347–350. https://doi.org/10.5281/zenodo.1179172

[29] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. [n. d.]. *Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion.* arXiv:2301.11757 [cs, eess] http://arxiv.org/abs/2301.11757

[30] Brigitte Schulte-Fortkamp, André Fiebig, Joseph A. Sisneros, Arthur N. Popper, and Richard R. Fay (Eds.). 2023. *Soundscapes: Humans and Their Acoustic Environment.* Springer Handbook of Auditory Research, Vol. 76. Springer International Publishing. https://doi.org/10.1007/978-3-031-22779-0

[31] Hugo Scurto and Axel Chemla–Romeu-Santos. [n. d.]. Deeply Listening Through/Out the Deepscape. In *ISEA2023 PROCEEDINGS* (Paris, France, 2024-06-06). Ecole des arts decoratifs - PSL. https://doi.org/10.69564/ISEA2023-85-full-Scurto-et-al-Deeply-Listening

[32] Hugo Scurto and Ludmila Postel. 2023. Soundwalking Deep Latent Spaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Mexico City, Mexico, 2023-05), Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Article 33, 232–235 pages. https://doi.org/10.5281/zenodo.11189166

[33] Domenico Stefani, Matteo Tomasetti, Filiippo Angeloni, and Luca Turchet. 2024. Esteso: Interactive AI Music Duet Based on Player-Idiosyncratic Extended Double Bass Techniques. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Utrecht, Netherlands, 2024-09), S M Astrid Bin and Courtney N. Reed (Eds.). Article 72, 490–498 pages. https://doi.org/10.5281/zenodo.13904929

[34] Jérôme Sueur, Bernie Krause, and Almo Farina. 2019. Climate Change Is Breaking Earth's Beat. 34, 11 (2019), 971–973. https://doi.org/10.1016/j.tree.2019.07.014

[35] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2021. AI-terity 2.0: An Autonomous NIME Featuring GANSpaceSynth Deep Learning Model. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Shanghai, China, 2021-06). Article 80. https://doi.org/10.21428/92fbeb44.3d0e9e12

[36] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2021. GANSpaceSynth. In *Proceedings of the 2nd Joint Conference on AI Music Creativity* (online, 2021-07-18). Zenodo. https://doi.org/10.5281/zenodo.5137902

[37] Notto J. W. Thelle and Bernt Isak Wærstad. [n. d.]. Co-Creatives Spaces: The Machine as a Collaborator. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Mexico City, Mexico, 2023-05), Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Article 35, 244–250 pages. https://doi.org/10.5281/zenodo.11189170

[38] Barry Truax. 2002. Genres and Techniques of Soundscape Composition as Developed at Simon Fraser University. 7, 1 (2002), 5–14. https://doi.org/10.1017/S1355771802001024

[39] Nick Warren and Anıl Çamcı. [n. d.]. Latent Drummer: A New Abstraction for Modular Sequencers. In *NIME 2022* (The University of Auckland, New Zealand, 2022-06-28). PubPub. https://doi.org/10.21428/92fbeb44.ed873363

[40] Hildegard Westerkamp. 2002. Linking Soundscape Composition and Acoustic Ecology. 7, 1 (2002), 51–56. https://doi.org/10.1017/S1355771802001085

[41] Robert Zajonic. 1968. The Attitudinal Effects of Mere Exposure. 9 (06 1968). https://doi.org/10.1037/h0025848

[16] https://nci.org.au

[42] Shuoyang Zheng, Bleiz M Del Sette, Charalampos Saitis, Anna Xambó, and Nick Bryan-Kinns. 2024. Building Sketch-to-Sound Mapping with Unsupervised Feature Extraction and Interactive Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (Utrecht, Netherlands, 2024-09), S M Astrid Bin and Courtney N. Reed (Eds.). Article 86, 591–597 pages. https://doi.org/10.5281/zenodo.13904959