http://dx.doi.org/10.14236/ewic/EVA2020.15

YomeciLand x Bunjil Place: The sounding body as play

Matthew Riley
RMIT University
Melbourne VIC 3000
matthew.riley@rmit.edu.au

Uyen Nguyen
RMIT University
Melbourne VIC 3000
uyen.nguyen@rmit.edu.au

Phuong Duy Nguyen
RMIT University, Vietnam
Ho Chi Minh City, Vietnam
duy.nguyenphuong@rmit.edu.vn

Max Piantoni
RMIT University
Melbourne VIC 3000
max.piantoni@rmit.edu.au

The advancing capabilities of computational systems to learn and adapt autonomously with datasets have provided new opportunities for designers and artists in their creative practice. This paper examines *YomeciLand x Bunjil Place* (Nguyen 2019), a playable sound-responsive installation that uses audio recognition to capture, recognise and categorise human sounds as a form of input to evolve a virtual environment of 'artificial' lifeforms. The potential of artificial intelligence in creative practice has recently drawn considerable interest, however our understanding of its application in sound practice is only emerging. The project is analysed in relation to three key themes: artificial intelligence for sound recognition, the 'sounding body' as play, and digital audiovisual composition as performance. In doing so, the research presents a framework for how artificial intelligence can aid sound recognition in a sound-responsive installation with *YomeciLand x Bunjil Place* shared as a case study to demonstrate this in practice.

Play. Sounding. Performance. Interactive installation. Sound recognition.

1. YOMECILAND X BUNJIL PLACE

YomeciLand x Bunjil Place (Nguyen 2019*)* is a playable sound-responsive installation exhibited at Bunjil Place, a cultural and community precinct in the City of Casey, Melbourne. Part of a series of works initiated by Uyen Nguyen titled *Yomeci (You, Me and the City)* that have investigated the playful potential of sound in animation, interactive installation and pervasive games, *YomeciLand x Bunjil Place* features a real time 3D environment of invented artificial flora and fauna that evolves through the sonic performances of visitors to the gallery space. To interact with the work a participant presses a button situated on a central stand, which triggers the work to 'listen' to sounds that are made into its microphone. Non-verbal expressions such as whistling, laughing and humming are recognised and categorised as a form of audio input and visualised in its virtual environment as an animated 'lifeform'. Each 'lifeform' has their own sonic and visual identity, and when the various digital entities are called into the world, a diverse digital ecology of movement, animation and sound is established. As the world becomes increasingly populated, unique relations and assemblages between the digital entities are formed. Sounds made by one type of entity layer and converge with others to form different compositions and atmospheres. Participants build the work's sonic relationships through either their improvised choices or in a more deliberate manner, each sound input changing the audio-visual composition of *YomeciLand*. These configurations are formed continuously in real time with the gallery activated into a dynamic and playful performance space.

Figure 1: *YomeciLand x Bunjil Place.*

2. THE SOUNDING BODY AS PLAY

Play has long been a mode of engaging with the world and of constructing meaning and experience. More recently play scholars and designers (Gaver et al. 2004, Costello 2009, Polaine 2012) have examined how interactive work can evoke play in novel and expressive ways. With situated, ubiquitous and pervasive media embedded in our material world, the potential of embodied interactions to foster enjoyment and engagement has been a notable consideration in designing for play in this context. These experiences invite physical involvement where our bodies bring the work into being through interaction with digital systems.

In this paper we use the term 'sounding body' to describe human participants' embodied interaction with *YomeciLand x Bunjil Place*. With the body in action through making sounds with vocalisation and gestures, participants engage directly with the work. This sound-making is performative and enacts a playful engagement, aiming to elicit a delight and pleasure when participants explore how their physical actions relate and affect representations in the virtual environment.

Figure 2: *A participant of YomeciLand x Bunjil Place making sounds into the microphone.*

With the work responding to non-verbal sounds (clapping, humming, stomping, whistling, chewing, singing, chuckling) it encourages anyone to play. This accessibility was an important feature given almost a third of City of Casey residents speak English as a second language, with over 140 languages represented in that community. Young children readily engaged with *YomeciLand x Bunjil Place* in particular, without being timid in making sounds and enjoying the social interactions that emerged through discovering its modes of engagement.

YomeciLand x Bunjil involves the participant's 'sounding body' as both a maker and listener of sound. With each audio input altering the sounds of its world, this feedback cycle creates an emergent ensemble of both human and non-human sounds. The 'sounding body' is continually constructed and transformed in relation to the virtual world and the situated nature of the participants in the gallery space. Although the audio inputs are finite, the simple interactions with the work give rise to a diverse and dynamic experience each time the work is engaged with.

3 ARTIFICIAL INTELLIGENCE FOR SOUND RECOGNITION

Early prototypes of *YomeciLand* as an interactive installation utilised pitch and volume to categorise audio input, however *YomeciLand x Bunjil Place* implemented an audio classifier model to recognise audio input as sounds. The intention of this system was to encourage participants to perceive that they were being 'listened' to with the artificial world understanding them and the sounds they made with more accuracy.

Audio classification involves classifying sounds through informative features extracted from them. Its performance thus depends on the algorithms used for feature extraction and grouping. Although classifying sound and image might seem disparate, there has been an increased use of computer vision in connection with audio in recent years. This paradigm relies on the representation of audio signals as images, such as spectrogram, and subsequently methods for image classification derived from convolutional neural networks (CNN) of deep learning can now be applied to carry out classification for audio.

In support for audio classification, Google provides AudioSet (Gemmeke et al. 2017), a collection of more than 2 million human-labelled 10-second sound clips from YouTube videos, as a comprehensive audio dataset. In addition, of the recent milestones of image classification, VGGNet (Simonyan and Zisserman 2014) yielded higher performance than any other contemporaries, and Google provides VGGish (Hershey et al. 2017), a weighted variation of this CNN, trained using another large YouTube dataset, for feature extraction of audio classification systems, outputting an embedding of 128 dimensions representing the audio input. As VGGish is a CNN that takes images as inputs, the input sound must first be converted into a log mel spectrogram using Short-Time Fourier Transform followed by a mel scale conversion and a logarithm of the mel spectrogram.

The audio classification system used for *YomeciLand x Bunjil Place* was an adaptation of IBM developer model asset exchange. The system

reads a signed 16-bit PCM wav file, uses VGGish to generate embeddings, applies Principal Component Analysis transformation/quantization, forwards the output to a pre-trained multi-attention classifier (Yu et al. 2018) and outputs 5 highest correlated predictions. The classifier is trained using AudioSet, with 527 classes. The audio classifier is hosted on a network server. To enable sound recognition for *YomeciLand x Bunjil Place*, a cross-platform network socket was developed to send a wav file recorded from the interactive application to the network server. The audio file was fed into the classification process, and the predicted result was sent back to the application.

As 90% of the samples used for training the audio classification model of *YomeciLand x Bunjil Place* were categorised as Music or Speech, and the defects in recorded sounds during live demonstration caused by external factors, incorrect predictions were noticed to form a defined pattern. Thus, an empirical solution was developed to address this issue: a logic table served as a complementary filter supporting the audio classification system, as shown below:

Table 1: *The logic table used as a complementary filter to post-process the output of the classifier. The sound input represents the actions of the participant. The classifier output is the output of the audio classification model. Through this complementary table, all respective classifier outputs are post-processed as the desired value sent to the interactive application.*

Sound Input *(Participant's action)*	**Audio Classifier output**	**Desired Outcome**
Hum	*acapella, vocal music, lullaby, fly, housefly, bee, wasp, insect, chant*	*Spawn 'Hum' entity*
Clap	*finger, gunshot, cap gun, pulse, gun fire, whip*	*Spawn 'Clap' entity*
Step / Stomp	*shuffle, crunch, run, walk, footsteps, running, walking, bouncing, knock, door, tap, heart murmur, heart sounds, heartbeat, door, basketball bounce, percussion, drum, bass drum, music, musical instrument, hammer, drum machine, firecracker, firework*	*Spawn 'Step / Stomp' entity*
Whistle	*bird, coo, pigeon, dove, whistling, car alarm, owl, hoot, alarm, siren, fire alarm*	*Spawn 'Whistle' entity*
Chew	*typing, crumpling, crinkling, sizzle, frying (food)/ patter, tick-tock, computer keyboard, tick, mouse*	*Spawn 'Chew' entity*
Chuckle	*chortle, laughter, belly laugh, whimper, snort, baby laughter, whimper*	*Spawn 'Chuckle' entity*
Sing	*singing, lullaby, female singing*	*Spawn 'Sing' entity*
Whisper	*whispering*	*Spawn 'Whisper' entity*
Other	*other sounds*	*Spawn 'Other' entity*

4. DIGITAL AUDIOVISUAL COMPOSITION AS PERFORMANCE

With the 'artificial' lifeforms of *YomeciLand* manifested as various sounds, images and movements the work invites participants to playfully construct a complex audiovisual composition when populating its world. Consisting of invented flora and fauna, each 'species' has unique characteristics, which relate to each other in varying ways and play a consequential role in the composition with their different rhythms and qualities.

The sonic experience of *YomeciLand x Bunjil Place* builds through combinations of audio loops, which are activated when a digital entity is spawned into the virtual environment. Participants construct and influence this composition through their own sound making. For example, if a participant hums, it will grow a low-lying, grass-like form, which plays a calm and soothing sound. If a participant whispers, a fire-fly like form is called high into the sky playing light melodic notes. Taking advantage of the sound recognition accuracy, the system was designed to support participants control an agency in shaping the *YomeciLand* world. Some were drawn to creating different aural intensities at either end of the sound spectrum – from light and melodic ambiance to low-pitched deep sounds. However, most participants created diverse compositions and arrangements in experimenting with the varied interactions that were possible.

The composition of *YomeciLand x Bunjil Place* is also shaped through its visual forms and animation. It's virtual ecology of artificial fauna and flora animates in rhythmic movements and patterns, which form an ever-changing spectacle. Entities' distinct visual attributes (colour, shape, size, animation and so on) act as aesthetic effects and

Figure 3: *The 'lifeforms' of YomeciLand x Bunjil Place.*

events that can be rearranged and juxtaposed in different ways. Some participants were attached to specific 'lifeforms' and purposefully called them into the artificial world, to personalise and construct the scene to their interest. Participants interacted with the work both individually and collaboratively, harmonising in the gallery space as they made the same sound together. Through these relations, improvisations and collaborations *YomeciLand x Bunjil Place* acted as a performance between humans and non-humans, an assemblage of sound, movement and visuals enacted both by the embodied interactions of participants and the responses and behaviours of the artificial lifeforms.

5. CONCLUSION

This paper has shared a framework for using artificial intelligence to aid sound recognition in a playable installation. The modes of interaction and design of *YomeciLand x Bunjil Place* was shared to demonstrate this in practice. The research is intended to be of significance to designers and artists interested in exploring the potential of this intersection in their creative practice.

6. ACKNOWLEDGEMENTS

Uyen Nguyen – artist and director, Matthew Riley – producer, Max Piantoni – interactive software development, engineer - Duy Phuong Nguyen, Rod Price – sound design, thank you to Georgia Cribb, Angela Barnett, Catherine Bennetts-Cash, Sarah Lyons, the front of house team and staff at Bunjil Place, Ika Jumali (cut-out design), Nick Margerison (poems), School of Design RMIT University and the staff and students of RMIT Master of Animation, Games and Interactivity.

7. REFERENCES

Costello, B. M. (2009) Play and the experience of interactive art (Doctoral dissertation).

Gaver, W. W., Boucher, A., Pennington, S., and Walker, B. (2004) Cultural probes and the value of uncertainty. interactions, 11(5), 53-56.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., . . . Ritter, M. (2017) Audio Set: An ontology and human-labeled dataset for audio events. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5-9 March 2017.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., . . . Wilson, K. (2017) CNN architectures for large-scale audio classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5-9 March 2017.

Nguyen, U. (2019) *YomeciLand x Bunjil Place* (interactive installation).

Polaine, A. J. (2010) *Developing a language of interactivity through the theory of play* (Doctoral dissertation).

Simonyan, K., and Zisserman, A. (2014) *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.

Yu, C., Barsim, K. S., Kong, Q., and Yang, B. (2018) Multi-level attention model for weakly supervised audio classification. arXiv preprint arXiv:1803.02353.