INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Real-Time Parametric General-Purpose Mammalian Vocal Synthesiser

Roger K. Moore

Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield, UK
r.k.moore@sheffield.ac.uk

Abstract

Although R&D into 'speech synthesis' has received a considerable amount of attention over many years, there has been remarkably little effort devoted to constructing vocal synthesisers for non-human animals. Of course, interest in synthesising human speech has been driven by the demand for practical applications such as reading machines for the blind or voice-operated assistants. Nevertheless, there are potential uses for non-human vocal synthesis: *e.g.* in education, robotics or ecological fieldwork. The latter is of particular interest, since it is common practice to use 'playback' methods (based on recorded samples) that do not easily facilitate parametric control over key experimental variables. Therefore, this paper presents the design and implementation of a real-time parametric general-purpose mammalian vocal synthesiser. The approach taken has been to decompose the overall sound production system into the relevant anatomical components (such as the lungs, vocal folds, tongue and mouth), and to implement a real-time simulation in 'Pure Data' - an open-source dataflow programming language. The software was successfully used to design an appropriate mammalian voice for the *MiRo*® biomimetic robot, but there are potential applications in a number of areas. The software is available for free download at `http://www.dcs.shef.ac.uk/~roger/downloads.html`.

Index Terms: mammalian vocalisation, vocal synthesis, robotic voices, animal sound synthesis

1. Introduction

The idea that speech could be generated by some form of artificial device has a long and productive history. From von Kempelen's mechanical talking machine [1] to the voice of Apple's *Siri*, R&D into 'speech synthesis' has led to a wide variety of practical solutions ranging from simulating the human vocal apparatus [2], to concatenating segments of real speech [3], to building statistical models [4]. However, although the founding principles are well established, there has been relatively little effort devoted to constructing vocal synthesisers for *non-human* animals (see [5] for a recent review of vocal synthesis in toys).

Of course, research into the synthesis of human speech has been very much driven by the demand for practical applications such as reading machines for the blind or voice-based assistants. Nevertheless, there are potential uses for non-human vocal synthesis: for example in education, robotics or ecological fieldwork. The latter is of particular interest, since it is currently common practice to use 'playback' methods (based on recorded samples) [6], but such an approach does not easily facilitate parametric control over key experimental variables.

The work reported here was driven by two requirements. The first was an investigation into vocal interactivity between simple (robotic) agents not imbued with speech or language. Given the small size of the robots (in this case, modified e-pucks [7, 8]), it was deemed appropriate to provide them with the vocal abilities of a small rodent - see [9] for an example interaction. The second requirement was an invitation to supply the 'voice' for a commercial educational biomimetic robot called *MiRo*® [10] (see Figure 1) that was to be published by Eaglemoss Ltd. as a part-work construction project in a magazine. In this case, it was necessary to be able to investigate a range of alternative vocal designs (including robotic voices), so a more general-purpose solution was needed.

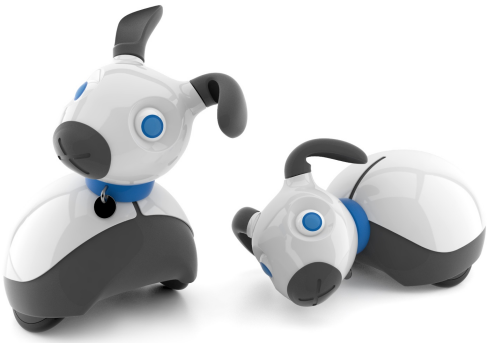

Figure 1: *The MiRo® biomimetic robot [10] (reproduced with permission from Consequential Robotics Ltd.).*

This paper presents the design and implementation of a real-time parametric general-purpose mammalian vocal synthesiser. The basic principles are outlined in Section 2 and additional features (such as emotion) are presented in Section 3. Section 4 describes the software implementation and Section 5 concludes with some remarks about future work.

2. Basic Principles

As is well established, different species make sound in different ways; many insects rub body parts together (a process known as 'stridulation'), birds create their songs using a vocal organ known as a syrinx, and mammals typically generate sound using a larynx [11]. The work described here was concerned with modelling the sound production mechanism in *land* mammals, with an emphasis on the ability to 'dial up' a range of different characteristics. The aim was thus to produce a general-purpose mammalian vocal synthesiser that could be configured to sound like any particular animal by selecting appropriate parameter settings. Also, since the target *MiRo*® robot was intended to be 'biomimetic' for educational reasons, it was important that the vocal synthesis should be grounded in the physical sound production apparatus (rather than based on animal recordings).

The most significant factors that influence sound production in the majority of mammals are physical characteristics such

http://dx.doi.org/10.21437/Interspeech.2016-841

as body size, lung capacity and the size and shape of the vocal tract. Of these, body size is the main determinant, since it impacts on the acoustic properties of the relevant anatomical components (lungs, vocal folds, tongue and mouth). Other factors are concerned with the dynamics of how the behaviour of the different components is organised and synchronised over time. The approach taken was thus to decompose the overall sound production system into these key elements, starting with the body.

2.1. Body

Clearly, mammals vary hugely in both physical size and shape. The basic body type has four limbs adapted for use on land, but some mammals are adapted for flying or swimming. The two latter categories of mammal tend to be the extremes in terms of size (the blue whale is over 30m in length, whereas the bumblebee bat is only 30mm long), and they also exploit different mechanisms for generating sound. The aim here was to target land mammals ranging in size from a small mouse to a large elephant, with the main interest focused on animals around the size of a small dog. Body mass for a selection of land mammals is shown in Table 1 [12].

Table 1: *Typical body mass for a selection of land-based mammals (data taken from [12]).*

Animal	**Body Mass** (kg)
African Elephant	6654
Asian Elephant	2547
Horse	521
Cow	465
Pig	192
Human	62
Sheep	55.5
Chimpanzee	52.16
Goat	27.66
Cat	3.3
Rabbit	2.5
Guinea pig	1.04
Mouse	0.023

2.2. Lungs

The main source of energy for mammalian vocalisation derives from the lungs, and the key factors are (a) the amount of air than can be stored in the lungs and (b) the rate at which it can be expelled. Again, there is a huge range across different animals: a blue whale's lung capacity is 5000 litres, whereas that of a mouse is only 500 microlitres. A human being has a typical lung capacity of 4-5 litres. In general, lung volume scales linearly with body mass [13, 14], and this has been characterised by [15] as follows:

$$C = 53.5 \times M^{1.06}, \tag{1}$$

where C is the lung capacity (in millilitres) and M is the body mass (in kilograms).

Airflow is related to breathing and, according to [13], breathing rate is given by:

$$B = 0.84 \times M^{-0.26}, \tag{2}$$

where B is the breathing rate (in Hertz).

However, breathing (and vocalising) uses only a proportion of the air in the lungs. Also, during vocalisation the flow rate is restricted by the actions of the vocal tract. For example, the average oral flow rate for voiced speech by an adult male human is 0.48 litres/sec [16] and the average breath group duration is 4.88 secs [17]. So, the average volume of air used during human vocalisation is 2.33 litres. For an 80 kg adult male with a lung capacity of 5.56 litres (from Equation 1), this corresponds to 0.42 of the total. Given that the normal exhalation time (derived from a breathing rate of 0.27 Hz) would be 1.86 secs, this means that vocalisation restricts airflow by a factor of 2.62.

So, assuming that similar principles hold across a range of different sized mammals, the volumetric flow rate Q (in litres per second) is given by:

$$Q = \frac{0.42 \times C}{2.62 \times \left(\frac{1}{2 \times B}\right)}, \tag{3}$$

which simplifies to:

$$Q = 0.32 \times C \times B. \tag{4}$$

2.3. Larynx

The main function of the larynx during vocalisation is to control the the length and tension of the vibrating vocal folds, thereby determining the pitch and timbre of the sound which excites the rest of the vocal tract. According to [18], the relationship between body mass and vocal excitation frequency for animals ranging in size from mice to elephants can be adequately modelled by:

$$F = M^{-0.4}, \tag{5}$$

where F is the fundamental excitation frequency (in kHz).

The timbre of a vocalisation is a function of the regularity of the vocal fold vibrations, the relationship between the fundamental frequency and its harmonics and the degree of turbulence in the airflow. Also, as well as fully voiced sounds, the mammalian larynx is capable of generating aspirated or noisy sounds (corresponding to whispering in human speech).

2.4. Vocal tract

The complete mammalian vocal tract consists of a larynx, a pharynx, an oral cavity and a nasal cavity. The pharynx lies immediately above the larynx, and this contains the epiglottis - an elastic cartilage that controls entry to the trachea (for breathing) or the oesophagus (for swallowing). Above the pharynx the airway splits into the oral cavity (containing the tongue and terminating at the mouth aperture) and the nasal cavity (terminating at the nasal apertures). The vocal tract structures above the larynx can be regarded as a set of interconnected acoustic tubes, each of which resonate at different frequencies depending on their size and shape. The resonances of the oral cavity are of particular significance since they can be changed by opening and closing the mouth and by moving the tongue.

Vocal tract resonances are often referred to as 'formants', and the frequencies of the different formants relate to the size and shape of the resonant cavities. This means that they are conditioned on the overall length of the vocal tract - the longer the vocal tract, the lower the formant frequencies. As one might expect, it has been found that the length of the vocal tract is correlated with body size (*e.g.* in dogs [19] and monkeys [20]), and [19] offers the following relationship:

$$L = 3.15 + (11.53 \times \log M), \tag{6}$$

where L is the vocal tract length (in cm).

Formant frequencies can be estimated by assuming that the vocal tract is a uniform acoustic tube which is closed at the glottis and open at the mouth. As the mouth closes, so the formants move down in frequency [21]. Hence, the resonant frequency of the nth formant R_n (in Hz) can be approximated by the equation:

$$R_n = (2n - (m+1)) \times \frac{c}{4 \times L}, \quad (7)$$

for $n = 1, 2, 3, :$, where m is the degree of mouth opening ($0 =$ open, $1 =$ closed) and c is the speed of sound (in cm/sec).

3. Additional Features

3.1. Emotion

Emotion is a complex physiological, cognitive and social phenomenon that is exhibited by both humans and animals. Formal study of the topic started with Charles Darwin [22]. It has been hypothesised that, despite different forms of expression in different species, there are certain common elements [23], and it is typical to refer to the six 'basic emotions': happiness, sadness, fear, anger, surprise and disgust [24]. However, more recent research favours a 'dimensional' approach based on *valence* (pleasure/displeasure), *arousal* and *dominance* [25].

What is certain is that the various 'affective states' that an animal can exhibit have the potential to influence vocalisation in predictable ways [1, 26, 27]. For example, high arousal may give rise to higher vocal pitch, amplitude and tempo, and low valence may give rise to a rougher/harsher voice quality.

3.2. Robotic vocalisation

Since one of the requirements for the work described here was to provide a voice for an educational robot (albeit based on a small mammal), there was interest in how the vocal characteristics could be made appropriate to a *non-living* entity. It has already been shown that aligning the look, sound and behaviour of robots [28] is a key to avoiding rejection by the users (the so-called 'uncanny valley' effect [29, 30]), so a robotic voice for a robotic animal was deemed important.

Of course there are many ways of post-processing a voice to make it sound robotic, one of the most famous being the use of a 'ring-modulator' to create the sound of the Daleks in the BBC's long-running TV series *Dr. Who*. Other techniques include modifying or replacing the residual signal in linear-predictive analysis-synthesis. These are rather artificial approaches based on audio special effects [31]. A more bio-inspired technique that has proved to be particularly effective in convincing listeners [32] is based on the premise that no living organism has more than one larynx. Hence, providing a robot with two or more sets of vocal folds elicits a strong perception of artificiality with no sacrifice in the quality of the vocalisation (somewhat similar to 'diplophonia', a medical condition in which two parts of the same set of vocal folds vibrate at different frequencies).

4. Implementation

4.1. Programming environment

The principles described above have been programmed in 'Pure Data' - referred to as "Pd" - an open-source visual programming language specifically designed to operate with real-time audio. Pd is a free alternative to Max™ - a programming language popular in the professional music industry. Both Pd and Max™ were authored by Miller Puckette from IRCAM in Paris. Pd is available for Windows, Mac OSX and GNU/Linux platforms, and `Pd-extended` is the recommended version to download [33].

Pd is an object-oriented dataflow programming language in which functions are created in a graphical design environment and which run immediately they are instantiated. A Pd program - known as a 'patch' - consists of objects, connections and data. Objects are functions such as [`print`], [`+`], [`fft~`] etc. and they connect with other objects via inlets and outlets. Connections between objects carry data in the form of messages or audio. Pd also provides various GUI (graphical user interface) objects such as sliders, graphs and buttons. Andy Farnell's book - *Designing Sound* [34] - provides an excellent introduction to Pd, and the advantages of using Pd for speech processing have been presented in [35].

4.2. Software structure

The overall structure of the synthesis software is based on a simulation of the flow of energy through a generic mammalian vocal apparatus in accordance with the principles outlined in Section 2. The key Pd objects correspond to the [`lungs`], [`larynx`], [`vocal tract`] and [`post-processing`] - see Figure 2. The command to vocalise initiates simulated airflow from the [`lungs`] with an amplitude that is calculated from the flow rate. A calculation is made of the duration of the vocalisation as a function of the flow rate and the lung capacity, and this is used to determine the period of the entire utterance (expressed in Hz).

These signals and messages are passed to the [`larynx`] which modulates the energy flow using the simulated action of one or two sets of vocal folds vibrating at a specified fundamental frequency, which is itself modulated by the utterance period. With default settings, this gives rise to a rise-fall intonation pattern. The voice quality, degree of aspiration (noise), level of quantisation and pitch difference between the two sets of vocal folds are all input parameters to the [`larynx`] and influence the signal that is output to the [`vocal tract`].

The [`vocal tract`] simulates three acoustic resonances (formants) using band-pass filters whose frequencies are determined by the vocal tract length and the degree of mouth opening (using Equation 7). A syllabic rate parameter controls the opening and closing of the mouth. Finally, the output from the [`vocal tract`] is sent to the [`post-processing`] object which contains an optional ring modulator and delay line (in order to introduce an echo effect).

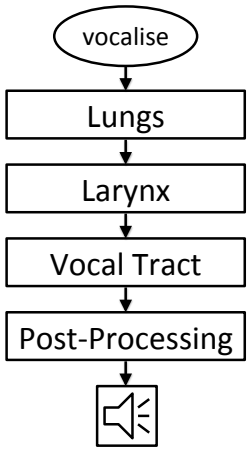

Figure 2: *Dataflow in the mammalian vocal synthesiser.*

4.3. Emotion

In order to facilitate the injection of some emotion into the vocalisations, parameters were included for *valence* and *arousal* (as discussed in Section 3.1). The arousal parameter modulates the airflow rate and, thereby, the amplitude and tempo of the vocalisations. High arousal leads to high airflow and *vice versa*. The valence parameter influences the fundamental frequency variance, the voice quality and, if a robotic voice is selected, the pitch difference between the two sets of vocal folds. The latter interpolates between a major chord for positive valence and a minor chord for negative valence.

4.4. Graphical user interface (GUI)

Since the aim was to produce a flexible *parametric* synthesiser, it was decided that it was important to allow key control parameters to be set via a GUI using appropriate buttons and sliders - see Table 2 and Figure 3. This facilitated real-time adjustment of the vocalisation, and greatly enhanced the process of designing different sound outputs. However, although in principle it is possible to set every parameter independently, in practice there are a number of potential dependencies (as described in Section 2). So, setting the body size to a particular value also sets:

- the lung capacity (using Equation 1),
- the breathing rate (using Equation 2),
- the flow rate (using Equation 4),
- the fundamental frequency (using Equation 5), and
- the vocal tract length (using Equation 6).

Table 2: *Control parameters provided by the program GUI.*

	Parameter	Values
Body	body mass	0.1 to 5000 (kg)
	body type	animal, robot
Lungs	lung capacity	0.1 to 1000 (lites)
	flow rate	0.005 to 500 (litres/sec)
Larynx	fundamental frequency	30 to 3000 (Hz)
	pitch quantisation	1 to 1000 (steps)
	pitch difference	1 to 1.5
	modulation frequency	0 to 20 (Hz)
	FM depth	0 to 1
	AM depth	0 to 1
	voice quality	1 to 10
	aspiration	0 to 1
Vocal Tract	uvula frequency	0 to 100 (Hz)
	vocal tract length	1 to 50 (cm)
	mouth close/open	0 to 1
	syllabic rate	0 to 20 (Hz)
	degree of closure	0 to 1
Emotion	arousal	0.7 to 7
	valence	-1 to +1
Post-Proc.	ring modulation	0 to 100 (Hz)
	delay line echo	0 to 100 (msec)

4.5. Preset animals and sound types

The software also provides a number of preset settings. For example, it is possible to select particular animals (such as a rat, cat, dog, sheep, dog or cow in the current version), and also select different types of vocalisation (such as normal, breathing, snoring, laughing/crying, sneezing and coughing). Selecting one of these presets simply moves relevant sliders to particular predetermined positions. After selecting a preset, it is still possible to vary any/all of the parameters as required in order to achieve a particular design objective.

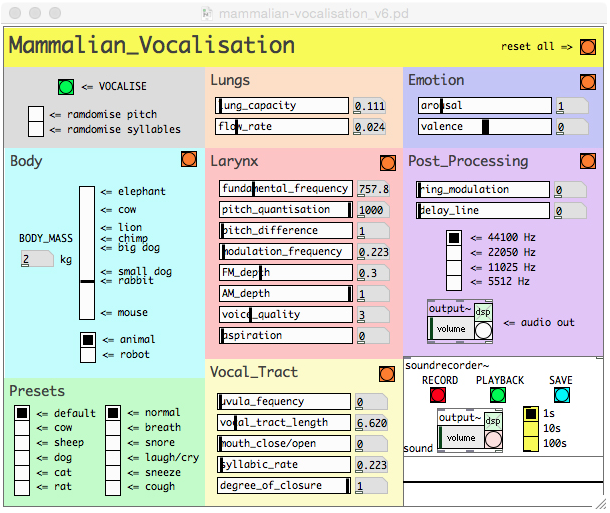

Figure 3: *Screenshot of the program GUI.*

5. Conclusion and Future Work

Although it has not been the subject of a formal evaluation, the software described herein was successfully used (a) to provide vocal output for an investigation into vocal interactivity between simple (robotic) agents, and (b) to design an appropriate mammalian voice for the *MiRo*® biomimetic robot (both described in Section 1). The output of the synthesiser certainly appears to be acceptable from an impressionistic perspective, with appropriate characteristics being exhibited for a range of different animals. A formal assessment will follow in due course.

A number of extensions and enhancements to the software are either planned or already underway. These include:

- implementing Fujisaki's model of intonation [36],
- providing independent control of the formants in order to better simulate movement of the tongue (and thereby facilitating more human-like vocalisations),
- allowing the static parameters (*i.e.* the presets) to be saved/recalled from file rather than hard-wired into the code, and
- driving the dynamic parameters by a continuous data stream (*e.g.* in the same way that the 'Holmes' parallel formant synthesiser is controlled [2]).

Finally, the latest version of the software is available for *free* download at `http://www.dcs.shef.ac.uk/~roger/downloads.html` (note that the program requires the `Pd-extended` programming environment to be installed from [33]).

6. Acknowledgements

This work was partially supported by the European Commission [grant numbers EU-FP6-507422, EU-FP6-034434, EU-FP7-231868 and EU-FP7-611971], and the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/I013512/1].

7. References

[1] W. R. von Kempelen, *Mechanismus der menschlichen sprache nebst der beschreibung seiner sprechenden maschine*. Wien: JV Degan, 1791.

[2] J. N. Holmes and W. Holmes, *Speech Synthesis and Recognition*. Taylor & Francis, 2002.

[3] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[4] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. W. Black, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS)," *6th ISCA Workshop on Speech Synthesis*, 2007.

[5] R. Hoffmann, "Voices for toys - First commercial spin-offs in speech synthesis," in *First International Workshop on the History of Speech Communication Research (HSCR 2015)*. Dresden, Germany: ISCA, 2015, pp. 60–70.

[6] P. K. McGregor, Ed., *Playback and Studies of Animal Communication*. Boston, MA: Springer US, 1992.

[7] F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J.-C. Zufferey, D. Floreano, and A. Martinoli, "The e-puck, a robot designed for education in engineering," in *9th Conference on Autonomous Robot Systems and Competitions*, Castelo Branco, Portugal, 2009, pp. 59–65.

[8] D. Floreano, S. Mitri, and J. Hubert, "E-puck: a robotic platform for studying the evolution of communication," in *Evolution of Communication and Language in Embodied Agents*, S. Nolfi and M. Mirolli, Eds. Berlin Heidelberg: Springer, 2010, pp. 303–306.

[9] "Squeaking E-Puck Robots." [Online]. Available: https://youtu.be/E4aMHK7AH5M

[10] "MiRo: The Biomimetic Robot." [Online]. Available: http://www.miro-robot.com/index.php

[11] S. L. Hopp and C. S. Evans, *Acoustic Communication in Animals*. Springer Verlag, 1998.

[12] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[13] J. Worthington, I. S. Young, and J. D. Altringham, "The relationship between body mass and ventilation rate in mammals," *Experimental Biology*, vol. 161, pp. 533–536, 1991.

[14] S. M. Tenney and J. E. Remmers, "Comparative quantitative morphology of the mammalian lung: diffusing area," *Nature*, vol. 197, no. 4862, pp. 54–56, jan 1963.

[15] W. R. Stahl, "Scaling of respiratory variables in mammals," *J. Applied Physiology*, vol. 22, no. 3, pp. 453–460, 1967.

[16] E. T. Stathopoulos, "Oral airflow during vowel production of children and adults," *Cleft Palate Journal*, vol. 21, no. 4, pp. 277–285, 1984.

[17] Y.-T. Wang, J. R. Green, I. S. B. Nip, R. D. Kent, and J. F. Kent, "Breath group analysis for reading and spontaneous speech in healthy adults." *Folia Phoniatrica et Logopaedica*, vol. 62, no. 6, pp. 297–302, 2010.

[18] N. H. Fletcher, "A simple frequency-scaling rule for animal communication," *Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2334–2338, 2004.

[19] T. Riede and T. Fitch, "Vocal tract length and acoustics of vocalization in the domestic dog (Canis familiaris)," *Journal of Experimental Biology*, vol. 202, no. 20, pp. 2859–2867, 1999.

[20] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[21] I. R. Titze, "Acoustic interpretation of resonant voice," *Journal of Voice*, vol. 15, no. 4, pp. 519–528, 2001.

[22] C. Darwin, *The Expression of the Emotions in Man and Animals*. London: John Murray, 1872.

[23] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research and Experience: Vol. 1. Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. New York: Academic Press, 1980, pp. 3–33.

[24] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds. New York: John Wiley, 1999, pp. 301–320.

[25] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology: Developmental, Learning, Personality, Social*, vol. 14, pp. 261–292, 1996.

[26] K. R. Scherer, "Vocal affect signaling: a comparative approach," in *Advances in the Study of Behavior*, J. Rosenblatt, C. Beer, M. C. Busnel, and P. J. B. Slater, Eds. New York: Academic Press, 1985, vol. 15, pp. 189–244.

[27] R. M. Seyfarth and D. L. Cheney, "Meaning and emotion in animal vocalizations," *Ann N Y Acad Sci.*, vol. 1000, pp. 32–55, 2003.

[28] R. K. Moore and V. Maier, "Visual, vocal and behavioural affordances: some effects of consistency," in *5th International Conference on Cognitive Systems - CogSys 2012*, Vienna, 2012, p. 76.

[29] M. Mori, "Bukimi no tani (the uncanny valley)," *Energy*, vol. 7, pp. 33–35, 1970.

[30] R. K. Moore, "A Bayesian explanation of the Uncanny Valley effect and related psychological phenomena," *Nature Scientific Reports*, vol. 2, no. 864, 2012.

[31] U. Zolzer, Ed., *DAFX: Digital Audio Effects*, 2nd ed. Chichester, UK: John Wiley & Sons, Ltd., 2011.

[32] R. K. Moore and A. Morris, "Experiences collecting genuine spoken enquiries using WOZ techniques," in *5th DARPA workshop on Speech and Natural Language*, New York, 1992.

[33] "Pure Data." [Online]. Available: https://puredata.info

[34] A. Farnell, *Designing Sound*. London: Applied Scientific Press Limited, 2008.

[35] R. K. Moore, "On the use of the Pure Data' programming language for teaching and public outreach in speech processing," in *INTERSPEECH*. Singapore: ISCA, 2014, pp. 1498–1499.

[36] H. Fujisaki and S. Nagashima, "A model for the synthesis of pitch contours of connected speech," *Annual Report of the Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.