
Proceedings Paper:
Moore, R.K. orcid.org/0000-0003-0065-3311 and Mitchinson, B. (2017) A biomimetic vocalisation system for MiRo. In: Biomimetic and Biohybrid Systems. Living Machines 2017. Conference on Biomimetic and Biohybrid Systems. Living Machines 2017, 26-28 Jul 2017, Stanford, CA. Lecture Notes in Computer Science, 10384 . Springer International Publishing , pp. 363-374. ISBN 978-3-319-63536-1

https://doi.org/10.1007/978-3-319-63537-8_30

A Biomimetic Vocalisation System for MiRo

Roger K. Moore[1] and Ben Mitchinson[2]

[1] Dept. Computer Science, University of Sheffield, UK
[2] Dept. Psychology, University of Sheffield, UK
{r.k.moore,b.mitchinson}@sheffield.ac.uk

Abstract. There is increasing interest in the use of animal-like robots in applications such as companionship and pet therapy. However, in the majority of cases it is only the robot's physical appearance that mimics a given animal. In contrast, *MiRo* is the first commercial biomimetic robot to be based on a hardware and software architecture that is modelled on the biological brain. This paper describes how *MiRo*'s vocalisation system was designed, not using pre-recorded animal sounds, but based on the implementation of a real-time parametric general-purpose mammalian vocal synthesiser tailored to the specific physical characteristics of the robot. The novel outcome has been the creation of an 'appropriate' voice for *MiRo* that is perfectly aligned to the physical and behavioural affordances of the robot, thereby avoiding the 'uncanny valley' effect and contributing strongly to the effectiveness of *MiRo* as an interactive device.

Keywords: biomimetic robot, MiRo, mammalian vocalisation, vocal synthesis

1 Introduction

Recent times have witnessed increasing interest in the use of animal-like robots in applications such as companionship and pet therapy. For example, *PARO* [1] is an interactive robotic seal that is particularly popular for therapeutic use in hospitals and care facilities where a live animal would be problematic. Like *PARO*, the majority of such 'zoomorphic' devices are engineered to support specific use-cases, and it is often only the robot's physical appearance that mimics a given animal. In contrast, *MiRo* [2] (designed and built by Consequential Robotics Ltd. in collaboration with the University of Sheffield) is the first commercial robot to be controlled by a hardware and software architecture that is specifically modelled on the biological brain [3,4].

MiRo is a highly featured, low-cost, programmable mobile developer platform, with a friendly animal-like appearance, six senses, a nodding and rotating head, moveable hearing-ears, large blinking seeing-eyes, and a responsive wagging tail. It has been designed to look like a cartoon hybrid of a generic mammal (see Fig. 1). A unique biomimetic control system allows *MiRo* to behave in a life-like way: for example, listening for sounds and looking for movement, then approaching and responding to physical and verbal interactions.

Fig. 1. *MiRo: the world's first commercial biomimetic robot (available from Consequential Robotics Ltd [2]).*

Of special interest here is *MiRo*'s ability to vocalise. In particular, it was considered important that the vocal output generator should be grounded in a biomimetic model of an appropriate physical sound production apparatus rather than based, for example, on pre-recorded animal sounds. As a consequence, *MiRo*'s voice was designed using a real-time parametric general-purpose mammalian vocal synthesiser [5] tailored to the particular physical and behavioural characteristics of the robot.

This paper presents the biomimetic principles underlying *MiRo*'s vocalisation system, and describes how they have been integrated into the robot's overall architecture. Section 2 reviews the principles underlying mammalian vocalisation, and Section 3 outlines *MiRo*'s overall control architecture. Section 4 then describes how the particular characteristics of *MiRo*'s voice were first designed using a general-purpose mammalian vocal synthesiser, and subsequently implemented on the robot platform itself. Finally, Section 5 concludes with some observations about the effectiveness of the derived solution.

2 Vocalisation in Mammals

The majority of animals make sound, and different species make sound in different ways. For example, many insects rub body parts together (a process known as 'stridulation'), birds create their songs using a vocal organ known as a 'syrinx', and *land*[1] mammals typically generate sound using a 'larynx' [6].

[1] Some mammals are adapted for the air or for water. However, such animals tend to be the extremes in terms of size (for example, the bumblebee bat measures only 30mm, whereas the blue whale is over 30m long) and exploit different mechanisms for generating sound than the majority of land mammals.

The vocal tract for a typical land mammal consists of a larynx, a pharynx, an oral cavity and a nasal cavity (see Fig. 2). These anatomical features evolved primarily for breathing and eating. However, over time they have been recruited to create and shape sound. The main sound source is the larynx, which contains a set of 'vocal folds' (sometimes referred to as 'vocal cords' or the 'glottis') - a pair of elastic membranes that are held apart while breathing, but brought together when eating (in order to stop unwanted material from entering the lungs). The consequence of this arrangement is that air from the lungs that is forced through the closed vocal folds causes them to vibrate. This creates acoustic energy in the form of a harmonic-rich buzzing sound with a distinct fundamental frequency (perceived as the 'pitch' of the voice). Muscles in the larynx control the length and tension of the vocal folds which, in turn, determine the pitch and timbre of the generated sound.

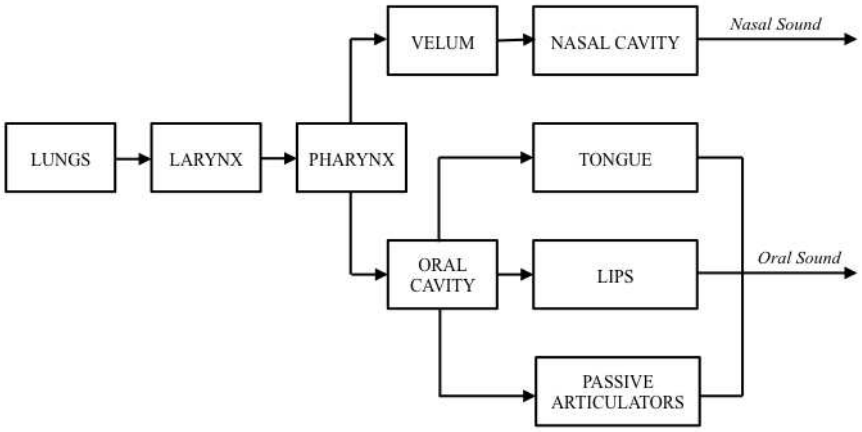

Fig. 2. *Schematic diagram of the mammalian vocal tract.*

The rest of the vocal tract can be regarded as a set of interconnected acoustic tubes. 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)[2]. Above the pharynx the acoustic tube splits into two: the main oral cavity (containing the tongue and terminating at the mouth and lips) and the nasal cavity (terminating at the nose). Airflow into the nasal cavity is controlled by the 'velum', a flap-like structure at the back of the mouth. All of these tubes resonate at different frequencies depending on their size and shape, hence they each filter the spectrum of the harmonic-rich

[2] It is this arrangement that enables mammals to choke!

excitation generated by the vibration of the vocal folds; energy is enhanced at some frequencies (the resonances) and suppressed at others. Vocal tract resonances are known as 'formants', and the formants arising from resonances in the oral cavity are of particular significance since they can be altered by moving the tongue and opening/closing the mouth[3].

In order to determine the characteristics of the sounds produced by a mammalian vocal tract, it is necessary to model (*a*) the 'airflow' (the rate at which air is expelled from the lungs), (*b*) the 'excitation' (the sound generated by the larynx) and (*c*) the 'filtering' (arising from the resonant cavities). Unsurprisingly, many of these processes are influenced by the size of the animal, since body mass has a direct impact on the physical and acoustic properties of the relevant anatomical components: the lungs, vocal folds, tongue and mouth.

2.1 Airflow

The main source of energy for mammalian vocalisation derives from the lungs, and the key factors are the amount of air that can be stored ('lung capacity') and the rate at which it can be expelled ('airflow'). In general, lung capacity C (in millilitres) scales linearly with body mass M (in kilograms) [7] as follows:

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

Obviously, airflow is related to breathing, and the breathing rate B (in Hertz) is given by [8] as:

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

However, breathing, and hence vocalisation, only uses a proportion of the air in the lungs (∼42%), and it also restricts airflow (by a factor of 2.62) [5]. This means that 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:

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

These parameters characterise the amplitude and duration of each vocalisation and, as can be seen, the predicted value for airflow (Q) is directly related to the size of the animal (M). However, these are mean values, and variation around the mean is possible. For example, a higher airflow would give rise to a shorter but louder vocalisation and *vice versa*.

[3] In human beings, these are the primary anatomical features used for speaking.

2.2 Excitation

As air is forced through the closed vocal folds, it escapes in bursts as the folds are momentarily pushed apart. After each bubble of air escapes, the Bernoulli effect causes the vocal folds to snap shut again, and this action generates a pulse of acoustic energy that propagates through the rest of the vocal tract. This sequence of events repeats at semi-regular intervals giving rise to a periodic excitation signal with energy at the fundamental frequency of vibration and its associated harmonics[4].

According to [9], the mean fundamental frequency F (in kHz) of the vocal fold vibration for animals ranging in size from mice to elephants is related to the body mass of the animal by the expression:

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

In other words, small animals have high-pitched vocalisations and large animals have low-pitched vocalisations.

The 'timbre' of a vocalisation is a function of (a) the regularity of the vocal fold vibrations, (b) the relationship between the fundamental frequency and its associated harmonics and (c) the degree of turbulence in the airflow. The latter means that, in addition to fully 'voiced' sounds, the mammalian larynx is capable of generating aspirated (breathy) 'unvoiced' sounds by holding the vocal folds close together and allowing a small amount of continuous airflow[5]. All of these aspects can be modelled by suitable shaping of the excitation waveform and by the injection of an appropriate level of random noise.

2.3 Filtering

The vocal tract resonances (formants) act as 'band-pass' filters which enhance acoustic energy at their resonant frequencies and suppress acoustic energy at other frequencies. This has a shaping effect on the harmonic-rich spectrum of the excitation signal emanating from the larynx[6].

The frequencies of the different formants can be estimated by assuming that the vocal tract is a uniform acoustic tube[7] which is closed at the vocal folds and open/closed at the mouth. As the mouth closes, so the formants move down in frequency [10]. Hence, the resonant frequency of the nth formant R_n (in Hz) can be approximated by the equation:

$$R_n = (2n - (m + 1)) \times \frac{c}{4 \times L}, \tag{6}$$

[4] To a first approximation, the signal generated by the glottis may be modelled as a 'sawtooth' waveform.

[5] Noisy unvoiced excitation at the vocal folds gives rise to whispering in human speech.

[6] It is the different placement of formants that gives rise to the production of different vowel sounds in human speech.

[7] This ignores the action of the tongue, which is an appropriate approximation for a non-human land mammal.

for $n = 1, 2, 3, :$, where m is the degree of mouth opening (0 = open, 1 = closed), c is the speed of sound (in cm/sec) and L is the length of the vocal tract (in cm) [5].

According to [11], vocal tract length is correlated with body size:

$$L = 3.15 + (11.53 \times \log M). \tag{7}$$

This means that large animals have long vocal tracts and thus low formant frequencies (and *vice versa*). It also means that the distribution of formant frequencies in a vocalisation provides information to a listener about the size of the animal[8].

2.4 Summary

The foregoing provides a complete specification of the minimum set of parameters necessary to simulate the vocalisation of a generic land mammal (as described in [5]). The novel contribution here is the mapping of this specification onto *MiRo*'s particular physical characteristics and control architecture.

3 MiRo's Control Architecture

MiRo's control architecture operates across three embedded ARM (Advanced RISC Machines) processors that mimic aspects of spinal cord, brainstem and forebrain functionality (including their relative speed and computational power) - see Fig. 3. One important feature is that the control latency of loops through the lowest reprogrammable processor can be as low as a few milliseconds. If required, *MiRo* can be operated remotely through WiFi or Bluetooth, and can also be configured as a Robot Operating System (ROS) node [13].

3.1 Actuators

MiRo is constructed around a differential drive base and a neck with three Degrees of Freedom (DoF). Additional DoFs include rotation for each ear, tail droop and wag, and eyelid open/close. All DoFs are equipped with proprioceptive sensors, and the platform also has an on-board loudspeaker.

3.2 Sensors

MiRo is equipped with stereo cameras in the eyes, stereo microphones in the base of the ears and a sonar range-finder in the nose. Four light-level sensors are placed at each corner of the base, and two infrared 'cliff' sensors point down from its front face. Eight capacitive sensors are arrayed along the inside of the body shell and over the top and back of the head (behind the ears). These provide an indication of direct human touch. Internal sensors include twin accelerometers, a temperature sensor and battery-level monitoring.

[8] Interestingly, animals such as the Red Deer are able to lengthen their vocal tract by lowering their larynx when vocalising, thereby giving the impression of being much larger than they really are [12].

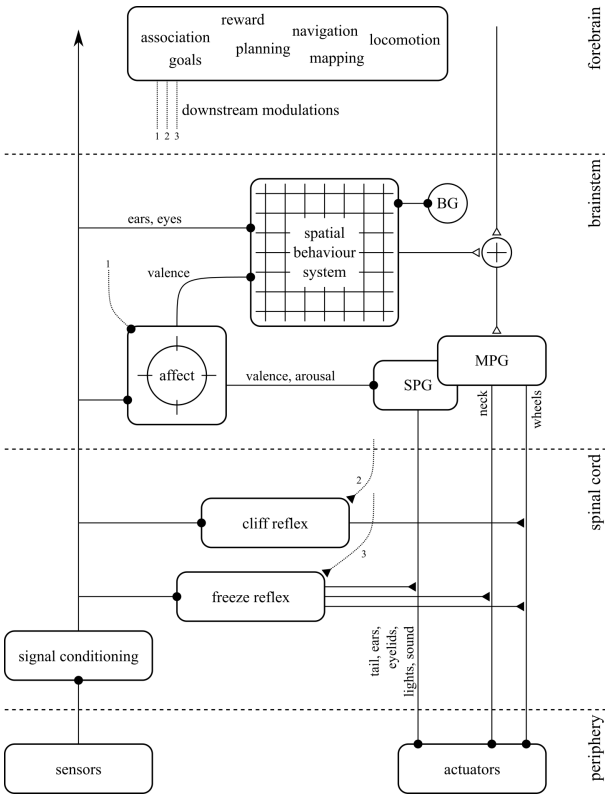

Fig. 3. *Illustration of MiRo's control architecture loosely mapped onto brain regions (spinal cord, brainstem, forebrain). Signal pathways are excitatory (open triangles), inhibitory (closed triangles), or complex (closed circles). BG is the Basal Ganglia. SPG and MPG are the Social and Motor Pattern Generators.*

3.3 Affect

MiRo represents its affective state (emotion, mood and temperament) as a point in a two-dimensional space covering 'valence' (unpleasantness-pleasantness) and 'arousal' (calm-excited) [14]. Events arising in *MiRo*'s sensorium are mapped into changes in affective state: for example, stroking *MiRo* drives valence in a positive direction, whilst striking *MiRo* on the head drives valence in a negative direction. Baseline arousal is affected by general sound/light levels as well as the time of day; *MiRo* is more active in the daytime. An individual event can cause an acute change: for example, a very loud sound might raise arousal and decrease valence. *MiRo*'s movements are modulated by its affective state, and it also expresses itself using a set of 'social pattern generators' that drive light displays, movement of the ears, tail, eyelids and - of particular relevance here - vocalisation.

4 MiRo's Vocalisation System

4.1 Vocal Design Environment

Prior to the development of *MiRo*, the first author had already constructed a real-time parametric general-purpose mammalian vocal synthesiser (in accordance with the principles outlined in Section 2) aimed at designing 'appropriate' vocalisations for a range of different animals and robots [5]. The design environment is implemented in 'Pure Data' (referred to as "Pd") - an open-source visual dataflow programming language specifically created to operate with real-time audio[9] [15]. The latest version is available for free download at `http://www.dcs.shef.ac.uk/~roger/downloads.html`.

The key Pd objects in the design software correspond to the `[lungs]`, [larynx], `[vocal tract]` and `[post-processing]`. The command to vocalise initiates simulated airflow from the `[lungs]` object with an amplitude that is calculated from the flow rate. The duration of the vocalisation is then calculated as a function of the flow rate and the lung capacity, and this is used to determine the period of the entire utterance.

These signals and messages are passed to the `[larynx]` object which modulates the energy flow using the simulated action of one or two[10] sets of vocal folds vibrating at a fundamental frequency determined by the body mass, 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]` object and influence the signal that is output to the `[vocal tract]` object.

The `[vocal tract]` object 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 6). A syllabic rate parameter controls the opening and closing of the mouth.

Control parameters are set via a GUI using appropriate buttons and sliders (see Fig. 4). This facilitates real-time adjustment of the vocalisation, and greatly enhances the process of designing different voices. In principle, it is possible to set every parameter independently. However, in practice, there are a number of potential dependencies (as described in Section 2). As a result, 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 7).

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

[9] Pd (and its professional counterpart: MAX-MSP) is commonly used in music studios.
[10] It is well established that two excitation signals slightly offset in fundamental frequency give the resulting vocalisation a distinct robotic timbre.

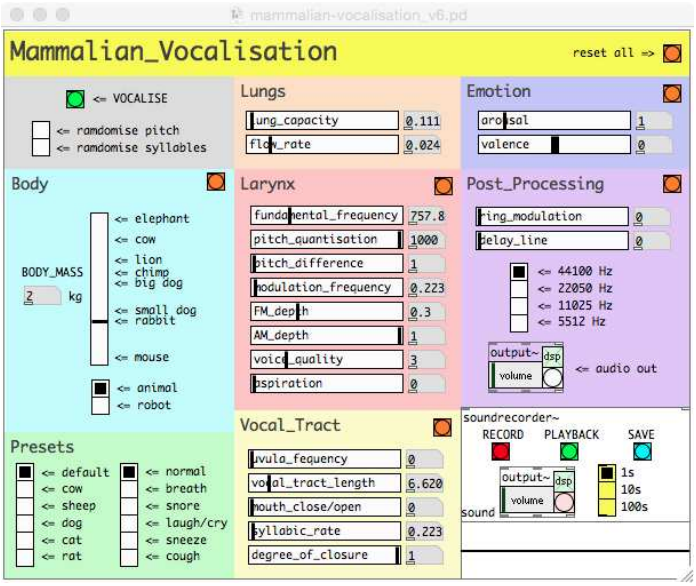

Fig. 4. *Screenshot of the Pd GUI for the real-time parametric general-purpose mammalian vocal synthesiser that was used to design MiRo's voice.*

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 all of the 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.

4.2 Implementation on MiRo

The real-time parametric general-purpose mammalian vocal synthesiser design environment described in Section 4.1 above was used (*a*) as a basis for implementing *MiRo*'s biomimetic vocalisation system on the robot platform outlined in Section 3 and (*b*) to determine the appropriate parameter settings. Accordingly, *MiRo*'s synthesis software (programmed in C) was structured to simulate the flow of energy through a mammalian vocal apparatus with body mass corresponding to a land mammal of an equivalent size (∼2 kg). The vocalisation modules were integrated into *MiRo*'s 'biomimetic core' (corresponding to the 'brainstem' in Fig. 3).

The robot has a breathing rhythm (∼0.7 Hz), the frequency of which is linked to arousal (see Section 3.3), and vocalisation is initiated stochastically during the exhalation phase. Breathing is simulated as cyclic airflow into and out of the lungs with an amplitude and duration that is calculated from the flow rate, lung capacity and body mass. When vocalising, the larynx modulates the airflow using the simulated action of a set of vocal folds vibrating at a

fundamental frequency (∼760 Hz) that is also determined by the body mass. The vocal tract then simulates three formants using band-pass filters whose frequencies are determined by the vocal tract length (∼6.6 cm) and the degree of mouth opening. A syllabic rate parameter controls the opening and closing of the mouth, and a vibrating uvula adds a 'cute' robotic timbre to the voice. It was decided *not* to employ two sets of vocal folds.

In order to allow the injection of emotion into the vocalisations, parameters were linked to *MiRo*'s two-dimensional affect map (as discussed in Section 3.3). Arousal modulates the rate of airflow and, thereby, the amplitude and tempo of the vocalisations; high arousal leads to high airflow and short vocalisations (and *vice versa*). Valence influences the variance of the fundamental frequency and the voice quality; high valence leads to expressive vocalisation whereas low valence produces more monotonic utterances.

4.3 Example Vocalisations

As an example of the vocalisation system in operation, Fig. 5 shows spectrograms[11] for two basic sounds - breathing and snoring. These *unvoiced* vocalisations are generated using noise as a excitation signal (as described in Section 2.2), and the spectrograms clearly illustrate the three-formant resonant structure of *MiRo*'s simulated vocal tract (Section 2.3). For these particular sounds, the vocal tract is fairly static throughout, hence there is little variation in the formant frequencies.

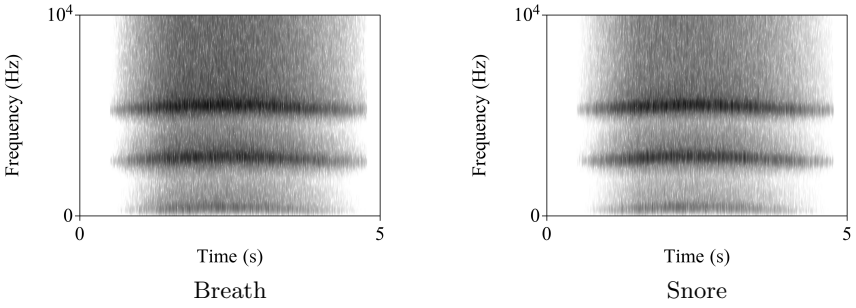

Fig. 5. Spectrograms of *MiRo*'s vocalisations for an exhaled breath and an inhaled snore. The dark bars indicate the concentration of energy at the formant resonances, and the vertical striations in the snore reflect the vibrating uvula.

In contrast, Fig. 6 shows spectrograms for *voiced* vocalisations with different affective states (as described in Section 3.3). As can be seen, these sounds are more dynamic than those shown in Fig. 5, mainly due to the opening and closing

[11] A time-frequency energy plot commonly used to analyse speech and audio signals.

of the mouth. In addition, the formants vary more with positive valence (due to larger mouth opening), and the durations are shorter with high arousal (due to higher airflow).

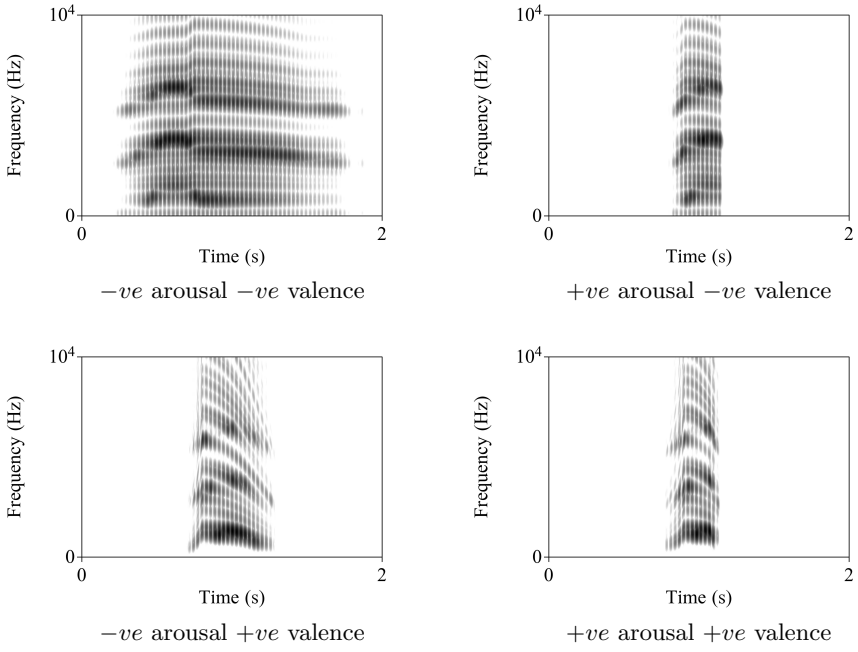

Fig. 6. Spectrograms of *MiRo*'s vocalisations resulting from different values for arousal and valence.

5 Summary and Conclusion

This paper has described the design and implementation of *MiRo*'s biomimetic vocalisation system. Based on the principles underlying vocalisation in land mammals, it has been shown how the key characteristics of *MiRo*'s vocalisations were selected using a real-time parametric general-purpose mammalian vocal synthesiser tailored to the specific physical characteristics of the robot. It has been explained how these design decisions were ported onto *MiRo*'s hardware/software platform and integrated into the robot's overall control architecture. The novel outcome has been the creation of an 'appropriate' voice for *MiRo* that is perfectly aligned to the physical and behavioural affordances of the robot [16]. As such, it successfully avoids the 'uncanny valley' effect caused by mismatched perceptual cues [17,18] and contributes strongly to the effectiveness of *MiRo* as an attractive interactive device.

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].

References

1. PARO Therapeutic Robot, http://www.parorobots.com
2. MiRo: The Biomimetic Robot, http://consequentialrobotics.com/miro/
3. Mitchinson, B., Prescott, T.J.: MiRo: A Robot Mammal with a Biomimetic Brain-Based Control System. In: Lepora, N., Mura, A., Mangan, M., Verschure, P., Desmulliez, M., Prescott, T.J. (eds.) Biomimetic and Biohybrid Systems. Living Machines 2016. LNCS, vol. 9793, pp. 179–191, Springer, Heidelberg (2016).
4. Collins, E.C., Prescott, T.J., Mitchinson, B., Conran, S.: MiRo: A Versatile Biomimetic Edutainment Robot. In: 12th International Conference on Advances in Computer Entertainment Technology (ACE'15). Iskandar, Malaysia, pp. 1–4, ACM Press (2015).
5. Moore, R.K.: A Real-Time Parametric General-Purpose Mammalian Vocal Synthesiser. In: INTERSPEECH. pp. 2636–2640. San Francisco, CA (2016).
6. Hopp, S.L., Evans, C.S.: Acoustic Communication in Animals. Springer Verlag, New York (1998).
7. Stahl, W.R.: Scaling of Respiratory Variables in Mammals. J. Applied Physiology, 22(3), 453–460 (1967).
8. Worthington, J., Young, I.S., Altringham, J.D.: The Relationship Between Body Mass and Ventilation Rate in Mammals. Experimental Biology, 161, 533–536 (1991).
9. Fletcher, N.H.: A Simple Frequency-Scaling Rule for Animal Communication. Journal of the Acoustical Society of America, 115(5), 2334–2338 (2004).
10. Titze, I.R.: Acoustic Interpretation of Resonant Voice. Journal of Voice, 15(4), 519–528 (2001).
11. Riede, T., Fitch, T.: Vocal Tract Length and Acoustics of Vocalization in the Domestic Dog (Canis familiaris). Journal of Experimental Biology, 202(20), 2859–2867 (1999).
12. Fitch, W.T., Reby, D.: The Descended Larynx is Not Uniquely Human. Proceedings of the Royal Society, B, 268(1477), 1669–1675 (2001).
13. Robot Operating System (ROS) http://www.ros.org
14. Collins, E.C., Prescott, T.J., Mitchinson, B.: Saying it with Light: A Pilot Study of Affective Communication Using the MiRo Robot. In: 4th International Conference on Biomimetic and Biohybrid Systems, 9222, pp. 243–55, Barcelona, Spain: Springer-Verlag New York (2015).
15. PureData https://puredata.info
16. Moore, R.K.: From Talking and Listening Robots to Intelligent Communicative Machines. In: Markowitz, J. (ed.), Robots That Talk and Listen, pp. 317–335, De Gruyter: Boston, MA (2015).
17. Mori, M.: Bukimi no Tani (The Uncanny Valley). Energy, 7, 33–35 (1970).
18. Moore, R.K.: A Bayesian Explanation of the 'Uncanny Valley' Effect and Related Psychological Phenomena. Nature Scientific Reports, 2(864) (2012).