A SYNTHESIS MODEL FOR MAMMALIAN VOCALISATION SOUND EFFECTS

WILLIAM WILKINSON[1], JOSHUA D. REISS[1]

[1] *Centre for Digital Music, Queen Mary University of London*
w.wilkinson@se13.qmul.ac.uk

In this paper, potential synthesis techniques for mammalian vocalisation sound effects are analysed. Physically-inspired synthesis models are devised based on human speech synthesis techniques and research into the biology of a mammalian vocal system. The benefits and challenges of physically-inspired synthesis models are assessed alongside a signal-based alternative which recreates the perceptual aspects of the signal through subtractive synthesis. Nonlinear aspects of mammalian vocalisation are recreated using frequency modulation techniques, and Linear Prediction is used to map mammalian vocal tract configurations to waveguide filter coefficients. It is shown through the use of subjective listening tests that such models can be effective in reproducing harsh, spectrally dense sounds such as a lion's roar, and can result in life-like articulation.

INTRODUCTION

There is increasing demand in the film and video game industries for adaptable, interactive sound effects whose output can be adjusted through user action, such as video game players or a sound designer creating effects for film. This has motivated much research into synthesis models based on procedural audio [1]. Attempts have been made to reproduce a large number of sound effects with a small number of synthesis models using high-level parameters that control the output intuitively [2]. Animal sound synthesis offers high potential in this regard, since they are commonly used in video games and often adapted to produce sounds for fictitious creatures, but the process of recording real-life animal vocalisations to fit the requirements of multiple video game situations can be very difficult. Research into the animal vocal system [3] suggests that human speech synthesis techniques such as the source-filter methodology used in [4] can also be applied to animals. This paper examines this approach, focusing on the reproduction of complex vocalisations by mammals such as lions, tigers and wolves.

Source-filter synthesis models for human and animal vocalisation sound effects fall into two main categories: physically-inspired models, such as [2], and signal-based models such as those outlined in [1].

Physically-inspired models are closely linked to the physics of sound. Their parameters are of a physical nature, and as such can be controlled in a way that corresponds to the real world. Such models can be accurate and respond intuitively to parameter changes, but also involve much complexity which is undesirable if processing capabilities are limited.

For these reasons, signal-based approaches, which aim to recreate the perceptual quality of a signal through non-physical techniques, are often favoured for their simplicity. While physical vocalisation models involve an excitation signal based on the movement of the vocal cords, signal-based models don't concern themselves with the physical movement itself, but aim to recreate the signal through other means, such as waveshaping. Similarly, the vocal tract, through which the excitation signal resonates, is often modelled in a physical system using a waveguide filter whose delay segments represent the propagation of sound waves through the vocal tract, and whose filter coefficients directly relate to the shape of the tract [4]. A signal-based system reproduces the filtering effect of the vocal tract without modelling wave propagation, often using a series of band-pass filters [1].

This research hopes to advance the state of the art of mammalian vocalisation synthesis through modelling of nonlinearity in the larynx and modelling of vocal tract losses. It also aims to show that accurate vocal tract configuration through Linear Prediction can result in realistic articulation, motivating further research into comprehensive documentation of system parameters enabling synthesis of a vast number of animal sounds.

In this paper, section 1 discusses existing research in the field of animal vocalisation synthesis, and the physics of the mammalian vocal system. Section 2 describes the software design and synthesis implementations, which were performed in Pure Data. Sections 3 and 4 are analysis and testing of example sound effects.

1 BACKGROUND AND LITERATURE

1.1 The Physics of Mammalian Phonation

Most animals share a common method of producing sound [3], similar in many aspects to the human vocal system, in which phonation occurs due to the existence of a muscular diaphragm pumping air from the lungs through the larynx. The existence of elastic membranes in the larynx of mammals, called ‘vocal folds’, enables complex phonation.

The term ‘vocal tract’ is used to describe the pharyngeal, oral and nasal cavities through which the sound must pass to reach the mouth. The length and shape of the vocal tract, which mammals are able to alter, determines the resonant frequencies at which the air travelling though it will vibrate. These frequencies are called formants.

Through analysis of the spectral dynamics of mammalian vocalisation it has been found that the first two formants, F1 and F2, have significant potential for variation during phonation, 23% and 61% respectively [5]. Additionally, the formant dispersion (the spacing between F1 and F2) is thought to be the most salient indication of body size.

1.2 Nonlinearity in the Mammalian Vocal System

Figure 1 highlights the position of the vocal membrane, which is attached to the vocal folds but can oscillate with very different behaviour, resulting in nonlinearity [6].

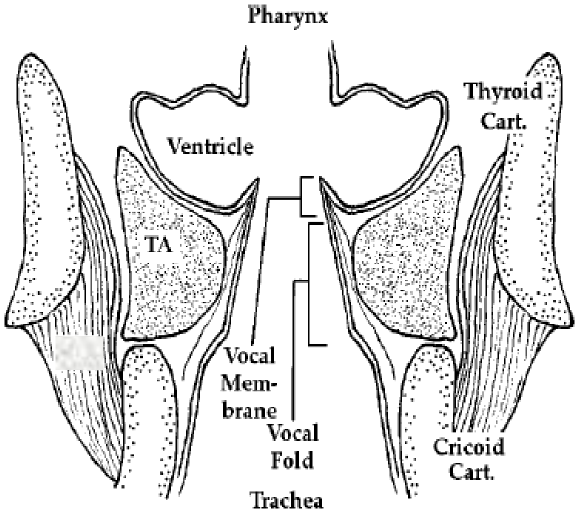


Figure 1: Diagram taken from [7]. The middle of the larynx. The vocal membrane extends upwards from the vocal fold, which covers the thyroarytenoid muscle (TA)

The vocal tract length and jaw cross-sectional areas of a lion change with respect to time [8], and such changes to the vocal tract result in dynamic formant structure. Studies of the functional morphology of tiger and lion vocal folds conclude that the square-like shape of the folds, which are very capable of stretching and shearing at low energy levels, allow for phonation (including nonlinearities) to occur at very low frequencies [9].

1.3 Existing Vocalisation Synthesis Techniques

Various attempts have been made to model such potential causes of nonlinearities present in the mammalian vocal system. For instance, [10] uses a mass-spring model for the vocal folds to implement nonlinearity using a van der Pol oscillator.

A signal-based model for animal vocalisations is outlined in [1], which consists of a parametric waveshaper used to recreate broadband excitation signals. A regular pulse sinusoid is modulated with another cosine function representing “vocal cord ripple”, creating sidebands around the signal’s harmonic components. This signal is passed through multiple parallel band-pass filters to create the desired sound texture.

Existing physical implementations of a human vocal tract system, such as [4], are based on the Kelly-Lochbaum waveguide filter [11], which models left- and right-travelling waves propagating through sections of the vocal tract modelled as acoustic tubes (cylinders) of varying size, separated by scattering junctions whose coefficients directly correspond to the vocal tract segment size in relation to neighbouring segments.

Figure 2 shows a small section of a waveguide filter system similar to those used in human speech synthesis. The vocal tract scattering coefficient, k_i, for segment *i* is calculated as follows, where R_i is the radius of vocal tract segment *i*:

$$k_i = \frac{R_{i-1} - R_i}{R_{i-1} + R_i} \tag{1}$$

2 MODEL DESIGN

2.1 The Glottal Flow Model

The system described here was implemented in Pure Data, with the exception of Linear Prediction analysis, which was performed in MATLAB. We model the glottal pulse using Equation (2) below. This is based on [4] and [12], with some changes aiding the usability of the system parameters. Parameter restrictions were applied to prevent the glottal flow from resulting in unrealistic values. The motivation for such a model is to ensure that the resultant glottal flow waveform closely matches the physical movement of the vocal cords, as outlined in [13], without introducing too much model complexity.

A noise factor is applied to the time-domain waveform, ensuring that noise only occurs during opening and closing. This obeys the observations in [4], where it was shown that noise was most prominent at these points. The glottal flow is then produced through implementation of wavetable synthesis.

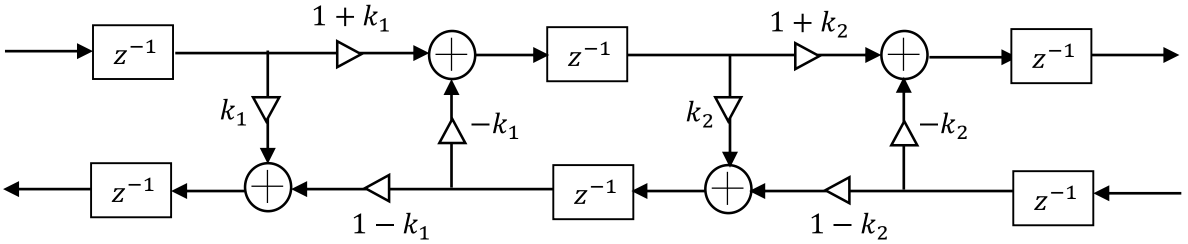


Figure 2: A waveguide filter system including scattering junctions

$$g[n] = \begin{cases} \frac{e_2 - e_1}{\pi}(1 - \cos[\frac{\pi n}{e_1 N}]) & n < e_1 \\ \frac{2(e_1 - e_2)}{\pi}\cos(\frac{\pi}{2(e_2 - e_1)}\left[\frac{n}{N} - 3e_1 + 2e_2\right]) & e_1 \leq n \leq e_2 \\ 0 & n > e_2 \end{cases} \tag{2}$$

The glottal flow, $g[n]$, for samples n = 0, 1, ..., N where N is the total number of samples in each period,$e_1$is the time throughout the pulse period at which the vocal folds fully open, and e_2 is the time at which they fully close, and $0 \leq e_1 \leq e_2 \leq N$.

2.2 Modelling Nonlinearity in the Mammalian Vocal System

Multiple nonlinear phenomena can be observed in the mammalian vocal system [6]. Causes of such phenomena include vocal folds oscillating at different frequencies or oscillating asynchronously, oscillation of the vocal membrane and stretching and shearing of the surface of the tract. In the physically-inspired models presented in this paper, such complexities are replicated using amplitude and frequency modulation (FM) synthesis. A user control parameter was built containing 6 levels of complexity, implemented as follows:

1. Limit Cycles. No amplitude modulation or frequency modulation is introduced, representing a biological system in which the oscillators (vocal cords) are synchronised.

2. Shimmer. Amplitude modulation is introduced with a low modulator frequency to recreate the effect of natural variation in vocal fold vibration.

3. Jitter and Subharmonics. Recreate the effect of synchronous vocal folds vibrating at frequencies with an integer ratio to each other.

4. Inharmonic Sidebands. Frequency modulation replicates the effect of desynchronised vocal folds vibrating at frequencies that do not have an integer ratio to each other.

5. Multiple-Nonlinearities. FM with a series of 2 modulators, creating a complex wave with many sidebands replicating the interaction of the vocal folds with the vocal membrane.

6. Deterministic Chaos. Introduce a third modulator to the FM series (3) so that the signal replicates deterministic chaos occurring due to stretching and shearing of the vocal membrane, the vocal folds and the vocal tract walls simultaneously:

$$\tilde{g}[n] = g[\omega_g \frac{n}{M} + A_{m_1}\cos\left(\omega_g \frac{n}{M} + A_{m_2}\cos\left(\omega_g \frac{n}{M} + A_{m_3}\cos\left(\omega_g \frac{n}{M}\right)\right)\right)] \tag{3}$$

M is the synthesis lookup table size, A_{m_i} is the amplitude of modulator *i*, ω_g is the fundamental frequency of the glottal waveform, $g[n]$.

For comparative purposes, a simpler method of nonlinear synthesis is also implemented and tested in which signal complexity is introduced by wrapping an amplified phasor back on itself to create small packets of waves whilst maintaining the fundamental signal frequency. This creates a rippling effect which is a simplistic representation of how nonlinearity may occur in the vocal system of a mammal. $\bar{p}(i)$ in (4) is the phasor used to drive wavetable synthesis:

$$\bar{p}(i) = \sqrt{mod[(\sqrt{k+1}\,p(i))^2, 1]} \tag{4}$$

where $p(i)$ is the original linear phasor and $k \in [1,6]$ is the user controlled complexity level.

2.3 Modelling the Vocal Tract

The physically-inspired vocal tract shown in Figure 3 is implemented using a waveguide filter model as in [4] with fifteen segments separated by scattering junctions

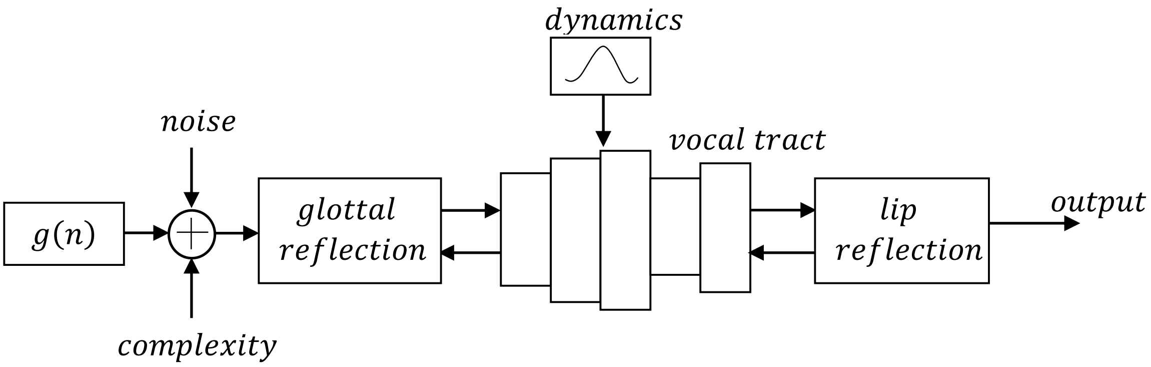


Figure 3: Block diagram of the physically-inspired source filter model. The glottal flow passes through the vocal tract which is modelled as a series of cylindrical tubes with varying radii. Waves are reflected back into the system at the glottis and the lip.

whose coefficients are controlled by the user and relate directly to the radii of the vocal tract segments. The input to the waveguide filter is the glottal flow waveform (2). The lip reflection is modelled explicitly by a high pass filter which reflects low frequencies back into the vocal tract and transmits high frequencies as output. In addition, the left travelling waves are also reflected back at the glottis. The controllable dynamics include amplitude, fundamental frequency variation, complexity control (3), noise control and duration.

The proposed model of the vocal tract is a totally lossless system, resulting in formants with very narrow bandwidth. Attempts have been made to model losses, such as [14] which incorporates factors like damping due to vibrating tract walls, fluid conduction and heat conduction losses. In this paper, a simplified version of such an effect is achieved by crossfading between the unfiltered and filtered signals. This enables user controlled changes to the signal-to-noise ratio, and the prominence of the formants.

In the signal-based model, the source signal is fed into a vocal tract model consisting of multiple parallel band-pass filters representing the signal formants, each with user controllable centre frequency, resonance and gain. Frequency and amplitude curves are exposed to the user, allowing for independent variation of the formants.

2.4 Parametrisation through Linear Prediction Analysis

The vocal tract characteristics of animals during phonation is still relatively unknown. For human speech synthesis, vocal tract configurations have been obtained via Linear Predictive Coding (LPC) analysis. In [4], a digital filter is designed based on a desired formant structure, the coefficients of which can be used to recursively calculate the vocal tract radius via an algorithm such as the Levinson-Durbin recursion [15], [16].

Mammalian vocalisations are less periodic and less predictable in their formant structure than human speech, which makes the design of a dataset of vocal tract configurations difficult. LPC was performed to demonstrate that such techniques can also work for a mammalian model, and some configurations are stored as examples to be used and edited.

The LPC analysis is carried out as follows: The input audio file is downsampled, for a fixed frame size, to achieve greater filter resolution at low frequencies. A high-pass filter is applied to account for the low-pass signal behaviour. The Levinson-Durbin recursion is then implemented to calculate the filter coefficients, which are subsequently used to calculate the vocal tract segment radii. Note that the delay parameter in the synthesis model must be set equal to the downsample rate to achieve formant accuracy.

In order to represent the changing shape of the vocal tract throughout phonation, LPC analysis was performed on a sample file split into multiple sections. The filter coefficients for these sections are then stored and sent to the vocal tract model during phonation. The bifurcation between states is handled by linearly interpolating the coefficients.

In order to implement sample-accurate feedback delay for the physical vocal tract model in Pure Data, a re-blocking technique was required, in which the block size for the vocal tract model was reduced based on the delay value chosen by the user.

3 ANALYSIS

Two example vocalisations, a lion’s roar and the growl of a wolf, were analysed via spectral examination and listening tests. In total, four synthesis models were used, with the intention of testing the effectiveness of physically-inspired modelling vs signal-based modelling and the effectiveness of the FM nonlinearity implementation (3) vs Phasor Wrapping (4):

Model 1: FM synthesis of $g(n)$ / Physical Vocal Tract
Model 2: Phasor Wrapping synthesis of $g(n)$ / Physical Vocal Tract
Model 3: FM synthesis of $g'(n)$ / Signal-Based Vocal Tract
Model 4: Phasor Wrapping synthesis of $g'(n)$ / Signal-Based Vocal Tract

3.1 Example 1: A Lion's Roar

The spectrogram analysis in Figure 4 shows more colouration in the low frequency bins of the synthesised lion's roar compared to the real version (sound files obtained from http://soundbible.com), suggesting that the lip reflection filter, designed to reduce the prominence of the glottal flow fundamental frequency and its low frequency partials, needs to be improved to create a vocalisation whose spectrum matches more closely that of the real signal.

A notable difference between the two spectrograms in Figure 4 is the existence of an additional formant in the real signal at around 220Hz. This can be attributed to the close proximity of the two formants at 220Hz and 320Hz. The filter order of 15 was not high enough to accurately represent these neighbouring peaks in the spectrum at the start of the roar.

3.2 Example 2: A Wolf Growl

Figure 5 shows that general formant dispersion appears to match well between the real wolf vocalisation and the

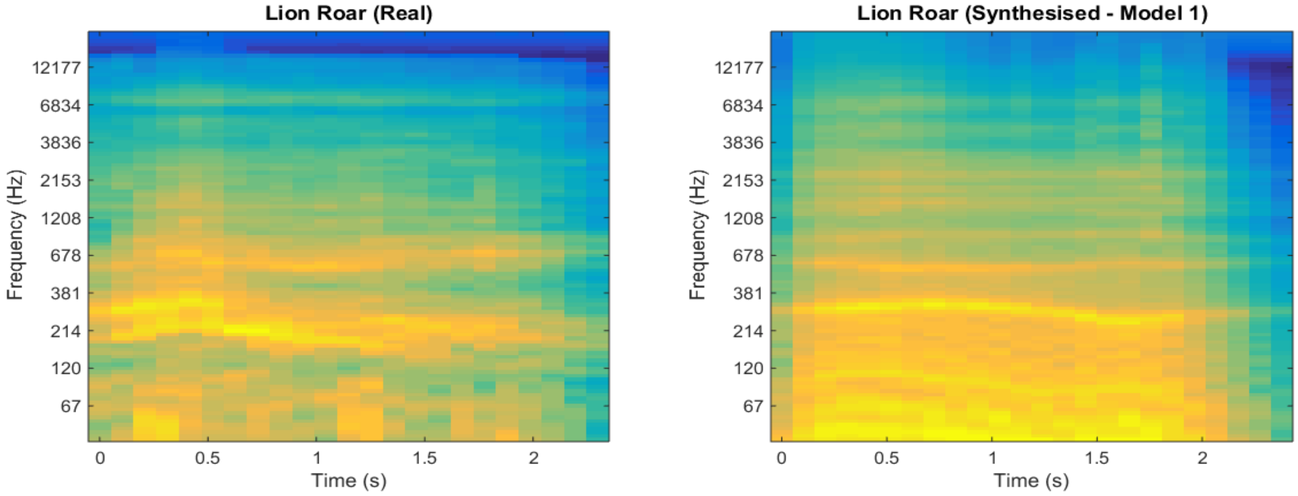


Figure 4: Log-frequency spectrogram of a recording of a lion's roar and a synthesised lion's roar produced using a physically-inspired synthesis model.

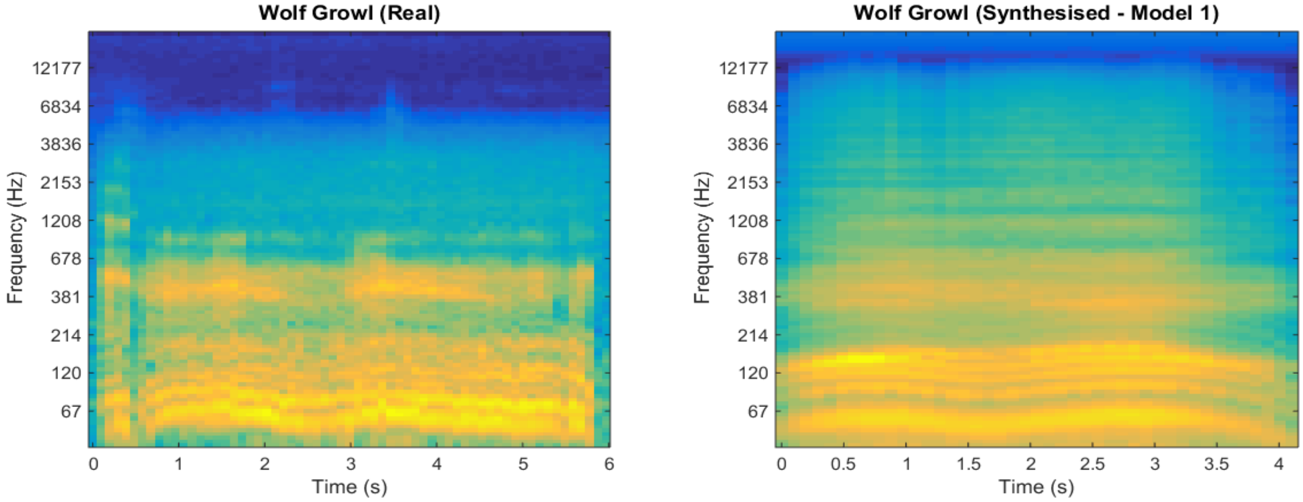


Figure 5: Log-frequency spectrogram of a real wolf growl and a synthesised wolf growl produced using a physically-inspired synthesis model

physically-inspired synthesis. However the relative amplitudes of these formants appear different. Through analysis of the output signals, the formant positions for the wolf growl sound effect were obtained for comparison.

The table of formants below suggests that the formant positions for the signal-based model are easier to control. This is to be expected since these are directly chosen and controlled by the user. However, the physically-inspired model appears to introduce more complexity to the signal and produces realistic articulation due to the bifurcation between vocal tract states. This hypothesis is supported by the listening tests performed in section 4.

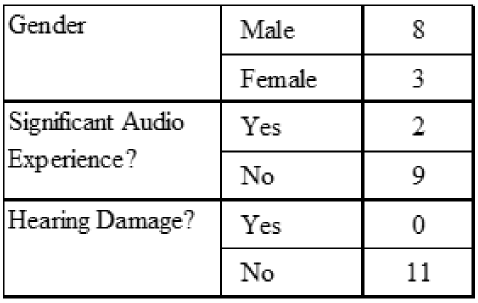
	Real	Model 1 (Physical)	Model 4 (Signal-Based)
F1	90 Hz	138 Hz	90 Hz
F2	441 Hz	400 Hz	455 Hz
F3	565 Hz	634 Hz	620 Hz

Figure 6: Formant positions in Hz for the real signal, Model 1 and Model 4

4 LISTENING TESTS

4.1 Testing Methodology

Both sound examples listed in section 3 were tested using a methodology based on the MUSHRA listening test [17] (a similar approach to evaluating synthesis techniques was used in [18]). A bespoke testing environment was developed in Pure Data for this project in which participants were asked to rate multiple lion roars (two real, six synthesised) and wolf growls (one real, five synthesised) by realism on a scale of 0 to 100. The tests were all performed over headphones (Sennheiser HD 25 II) and a total of eleven participants took part in each listening test with varying levels of audio experience.



Gender	Male	8
	Female	3
Significant Audio Experience?	Yes	2
	No	9
Hearing Damage?	Yes	0
	No	11

Figure 7: Listening test participant details.

4.2 Test 1: A Lion's Roar

Figure 8 shows that the physically inspired models designed for this project (Models 1 & 2) produced roaring sounds that often ranked higher than their signal-based counterparts. The poor performance of Model 4 suggests that the physical assumptions around nonlinearity and lip reflection techniques used in this work improve the effectiveness of synthesising a dense, harsh vocalisation such as a lion roar.

It is clear through listening to the output signals that the physical vocal tract succeeds in producing more realistic sounding articulation. Rather than a generic rise and fall of formant frequency, there is perceptually more expression in the roar. This can be attributed to the vocal tract configurations calculated via Linear Prediction analysis.

4.3 Test 2: A Wolf Growl

Figure 9 shows Model 2 performed poorly in Listening Test 2. In general, the FM nonlinear synthesis technique ranked well. Listening to the output samples suggests that this is due to the lack of variation present in the alternative phasor wrapping technique. In particular, FM synthesis introduces jitter and shimmer as well as natural-sounding variation.

5 CONCLUSION

A physically-inspired source-filter system was implemented based on detailed modelling of the glottal waveform and a waveguide filter representing the vocal tract. It was also demonstrated that Linear Prediction analysis could be used to create articulation that is crucial to producing life-like sounds. Such articulation was observed in the synthesised signals, resulting in the model performing well during subjective evaluation, often being perceived as more realistic than signal-based methods when synthesising a lion's roar. It is hoped that this will motivate further research into mapping of mammalian vocal tract parameters to Linear Prediction filter coefficients.

The use of FM synthesis techniques for replicating nonlinearity in a physical system also performed well in the listening tests, which should provide a platform for further analysis of the ways to model complexity in animal phonation. It should however be noted that analysis of the various synthesis techniques highlighted issues in the accuracy of the formant locations over time for the physically-inspired models. Spectral analysis suggested that the synthesis would benefit greatly from a more detailed lip reflection filter, and more detailed modelling of the losses throughout the vocal tract.

6 FURTHER WORK

Since the parameters built into these models are physically relevant, it is possible for synthesis to be embedded into a physical video game engine. This could be facilitated by parameter development and grouping, enabling the synthesis model to respond to game controls

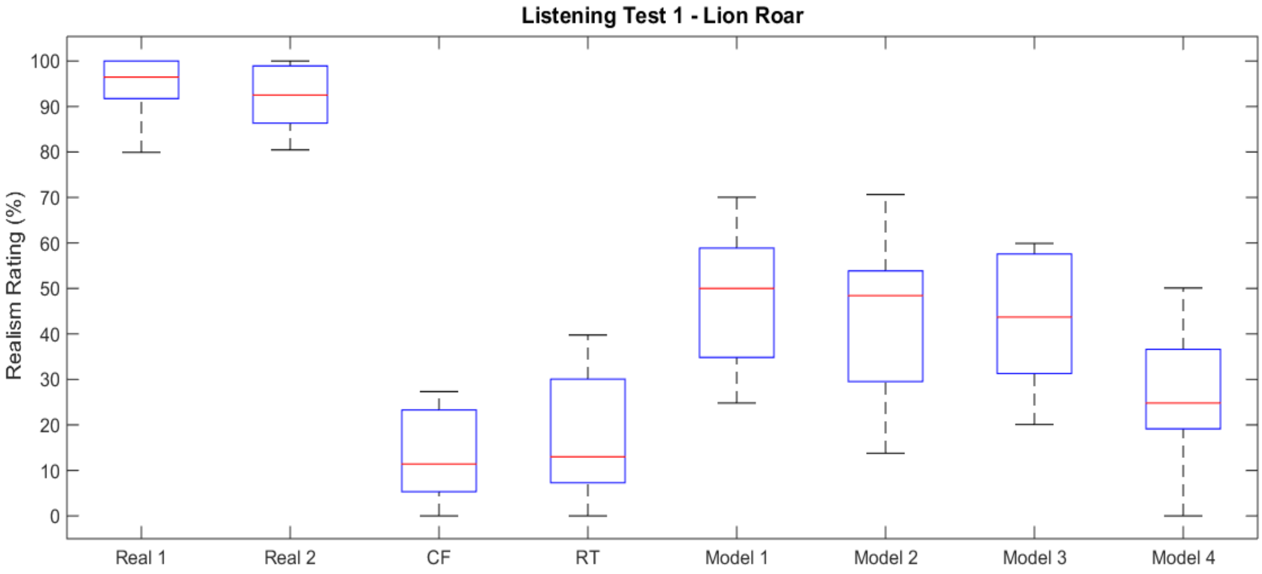


Figure 8: Listening test results for a lion's roar. CF represents the 'Creature Factory' signal-based model from [1]. RT represents synthesis from an online Lion Roar Synthesis Tutorial also created by Andy Farnell http://www.obiwannabe.co.uk/tutorials/html/tutorial_roar.html

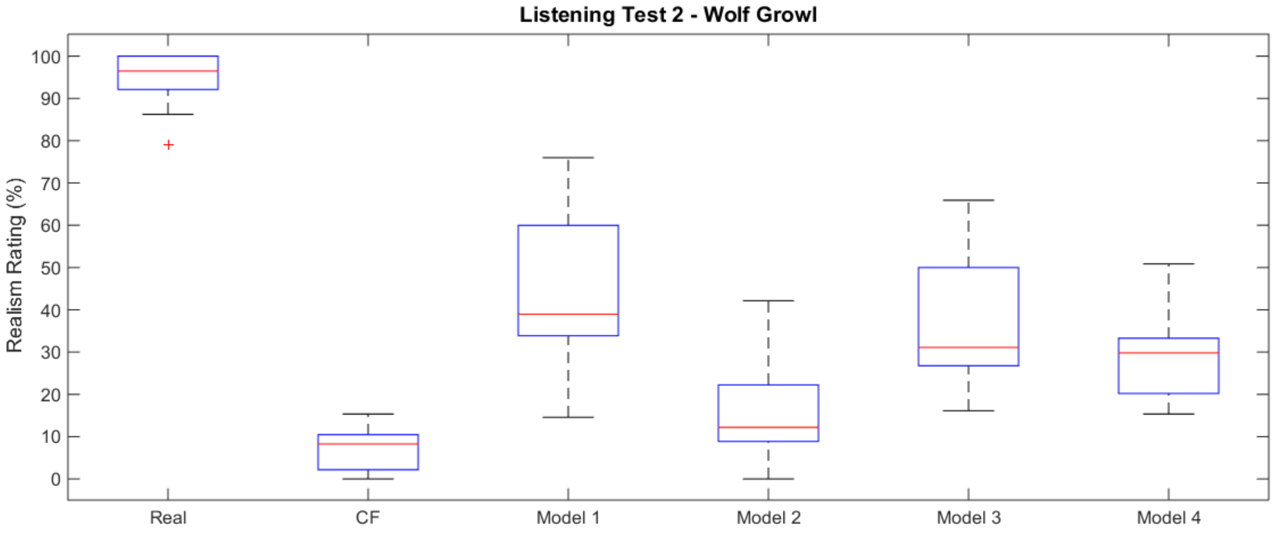


Figure 9: Listening test results for a wolf growl. CF represents the 'Creature Factory' signal-based model from [1].

such as animal size, intended emotion and listener distance. Since CPU resource may be scarce in this scenario, some of the realism of a physical model may have to be substituted for the computational efficiency of a signal-based model.

This paper has outlined a foundation for research into physically-inspired mammalian vocalisation synthesis. Further work should build on this foundation to analyse in detail the nonlinearity present in animal vocal systems, such as advanced modelling of the vocal folds and vocal membrane as coupled oscillators.

More physically-inspired features based on human speech synthesis research could also be added to the model. These should include the implementation of a realistic lip reflection filter, and a 3-way scattering junction in the vocal tract to model sound waves propagating through various cavities in the vocal system. The losses outlined in [14] should also be fully implemented to more closely reproduce formants with varying bandwidths.

The creation of a comprehensive dataset of vocal tract configurations, like those used to create vowel sounds in human speech synthesis would benefit this field of research. Linear Prediction techniques as outlined in this paper could be used to achieve this. Such a dataset would need to incorporate a wide variety of formant dispersions.

7 REFERENCES

[1] Farnell, Andy. *Designing sound*. Cambridge: MIT Press (2010).

[2] S. Hendry and J. D. Reiss, "Physical Modeling and Synthesis of Motor Noise for Replication of a Sound Effects Library," 129th AES Convention, San Francisco, Nov. 4-7 (2010).

[3] Fitch, W. T. "Production of vocalizations in mammals." *Visual Communication* 3 (2006): 145.

[4] Cook, Perry. "Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing." (1991).

[5] Taylor, Anna M., and David Reby. "The contribution of source–filter theory to mammal vocal communication research." Journal of Zoology 280.3 (2010): 221-236.

[6] Fitch, W. Tecumseh, Jürgen Neubauer, and Hanspeter Herzel. "Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production." *Animal Behaviour* 63.3 (2002): 407-418.

[7] Mergell, Patrick, W. Tecumseh Fitch, and Hanspeter Herzel. "Modeling the role of nonhuman vocal membranes in phonation." *The Journal of the Acoustical Society of America* 105.3 (1999): 2020-2028.

[8] Ananthakrishnan, Gopal, et al. "An acoustic analysis of lion roars. II: Vocal tract characteristics." *Fonetik 2011, Fonetik 2011. Royal Institute of Technology, Stockholm, Sweden, 8–10 June*. (2011).

[9] Klemuk, Sarah A., et al. "Adapted to roar: functional morphology of tiger and lion vocal folds." *PloS one* 6.11 (2011): e27029.

[10] Lucero, Jorge C., and Jean Schoentgen. "Modeling vocal fold asymmetries with coupled van der Pol oscillators." *Proceedings of Meetings on Acoustics*. Vol. 19. No. 1. Acoustical Society of America (2013).

[11] Kelly, John L., and Carol C. Lochbaum. "Speech synthesis." *Proc. Fourth Int. Congr. Acoustics* (September 1962): 1-4

[12] Rosenberg, Aaron E. "Effect of glottal pulse shape on the quality of natural vowels." *The Journal of the Acoustical Society of America* 49.2B (1971): 583-590.

[13] Doval, Boris, Christophe d'Alessandro, and Nathalie Henrich. "The spectrum of glottal flow models." *Acta acustica united with acustica* 92.6 (2006): 1026-1046.

[14] Story, Brad Hudson. "Physiologically-Based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract." (1995).

[15] Levinson, N. "The Wiener RMS error criterion in filter design and prediction." J. Math. Phys., v. 25 (1947): 261–278.

[16] Durbin, J. "The fitting of time series models." Rev. Inst. Int. Stat., v. 28 (1960): 233–243.

[17] International Telecommunication Union, "Multiple Stimuli with Hidden Reference and Anchor", ITU-R BS.1534-1 (2003).

[18] G. Durr, L. Peixoto, M. Souza, R. Tanoue and J. D. Reiss, Implementation and evaluation of dynamic level of audio detail, AES 56th Conf.: Audio for Games, London, UK, February 11–13 (2015).