Singing Synthesis With an Evolved Physical Model

Crispin Cooper, Damian Murphy, David Howard, and Andy Tyrrell, *Senior Member, IEEE*

***Abstract*—A two-dimensional physical model of the human vocal tract is described. Such a system promises increased realism and control in the synthesis of both speech and singing. However, the parameters describing the shape of the vocal tract while in use are not easily obtained, even using medical imaging techniques, so instead a genetic algorithm (GA) is applied to the model to find an appropriate configuration. Realistic sounds are produced by this method. Analysis of these, and the reliability of the technique (convergence properties) is provided.**

***Index Terms*—2-MUSI, digital waveguide mesh, finite difference, genetic algorithms (GAs), optimization methods, singing synthesis, speech synthesis, voice analysis.**

I. Introduction

ALTHOUGH digital speech synthesis has existed since the 1960s, research still continues into creating natural, humanlike sounds—both for those who cannot speak themselves and for human–computer interaction. Singing synthesis is a much newer problem, with additional difficulties and different applications. The world of music would benefit from singers who can achieve articulations beyond human ability, 24-h availability and also, singers who do not tire of "trying out" all of a composer's successive refinements to their composition! Increased understanding of the mechanism of singing aids teachers in both the music and speech therapy professions. As in the real world, it should be possible to unify speech and singing synthesis in the same model.

A. Speech and Singing Synthesis

Synthesis methods can be divided into two broad categories: spectral and physical models [1]. The earliest spectral model is Dudley's Vocoder [2], which approximates the voice source and vocal tract filter in the source-filter model [3] to enable later reconstruction. Linear predictive coding (LPC) [4] predicts the next sample of a speech signal based on past samples. Formant synthesizers (e.g., [5]) directly invoke the source-filter model of speech production. These methods assume linearity which can cause a resynthesized signal to sound artificial.

Some commercial speech systems splice together recorded sounds (often diphones—the transitions between phonemes) to synthesize speech. This concept has recently been extended to singing.[1]

Although some spectral synthesis methods have reached a level of practical everyday usage, it is not possible to recreate the voice of a speaker or singer who has not already undertaken a lengthy studio analysis session. A true physical model would provide this.

The first digital physical model was created by Kelly and Lochbaum [6], who simulated the vocal tract as a series of one-dimensional tubes. More recent extensions of their model, for example Cook [7], add greater control and more sophisticated modeling of the vocal tract wall, based on digital waveguide synthesis [8].

Such one-dimensional (1-D) models are often considered sufficient for speech synthesis. However, it has been argued that a two-dimensional (2-D) model provides more realism [9] and shown that it provides increased control of formant bandwidths over a 1-D model [10]. A 2-D mesh can also demonstrate additional modes of resonance, depending on closed wave paths between the two pairs of opposing boundaries or the four bounding surfaces combined. In the 1-D case, resonant modes are only supported between the two boundaries to the system formed at the glottis and lip ends. The additional frequencies generated in the 2-D model are well within the range of human hearing. Finally, 2-D and 3-D models may provide more data for the analysis of real singing. A 2-D model has been implemented using a 2-D rectilinear waveguide mesh [11]. However, the question remains as to how we choose the parameters that determine the shape of the mesh structure.

B. Speech and Singing Analysis

Obtaining data on the shape of the vocal tract while in use is an ongoing area of research. The subject has teaching and medical applications as well as being essential to a physical model as detailed in Section I-A. However, such data is hard to obtain.

Functional magnetic resonance imaging (fMRI) techniques can provide an estimate (e.g., [12]), but there are issues concerning the acoustic noise levels associated with the machine, the supine position required of the subject and too low a time resolution to enable dynamic changes to be accurately tracked. Improved time resolution as well as imaging of the teeth can be achieved with X-ray computed tomography (CT), but there are issues concerned with safe radiation dose levels [13].

Instead of direct measurement, it would be desirable to be able to infer the exact shape of a vocal tract from the sound it produces. For example, LPC analysis can provide access to the cross-sectional area functions of an acoustic tube model of the

Manuscript received January 1, 2004; revised July 21, 2005. This work was supported by the Future and Emerging Technologies Programme (IST-FET) for the European Community under Grant IST-2000-28027 (POEtic). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Davies.

The authors are with the Department of Electronics, University of York, Heslington YO10 5DD, U.K. (e-mail: crispin@cantab.net; dtm3@ohm.york.ac.uk; dh@ohm.york.ac.uk; amt@ohm.york.ac.uk).

Digital Object Identifier 10.1109/TSA.2005.860844

[1]Yamaha Vocaloid Technology. http://www.vocaloid.com.

oral tract from a speech pressure waveform [14]. This has been used to provide tools for the training of singers [15]. However, the mapping of sounds to potential tract shapes is one-to-many, so the LPC system does not provide a unique solution. It is conceivable that a unique solution *achievable by a human vocal tract* exists, however, it is not clear how to constrain the LPC process to exclude the other "unwanted" solutions. In addition, being an all-pole model, LPC can not model the acoustic effects of the nasal cavity.

Another approach is to wrap a synthesis model in an optimization loop [16]—automatically tweaking the parameters until a sound is produced which matches a human voice. We must of course, constrain the parameters to those that a human can achieve. This is not an easy optimization problem as it is not convex.[2] However, if successful, especially with a physical model, highly realistic synthesis and accurate analysis could be achieved.

The contribution of this paper is a system consisting of a 2-D physical model of the vocal tract, evolved using a genetic algorithm (GA) to find a shape suitable for the production of a given sound. While shapes derived in this way are not yet constrained to those achievable by a real vocal tract, a GA is capable of finding a shape which produces a synthesized output close to the original.

This method offers the potential for vocal tract area estimation that: 1) does not involve the presence of the informant for direct physical vocal tract measurement and 2) has the potential to provide a unique solution, as GA evolution can readily be constrained to model the known articulatory limits of the human vocal tract.

The remainder of the paper is structured as follows. Section II provides some background on Genetic Algorithms. Section III describes our system. The audio and shape data produced by the system is presented and discussed in Sections Sections IV–VI concludes.

II. Genetic Algorithms

GAs are now a standard technique for multivariable optimization problems, as described in numerous textbooks, e.g., [18]. They can be regarded as a smart (although partly random) search of the space of all possible solutions, which is infeasibly large to explore in its entirety. In this case the space consists of the range of possible shapes for the human vocal tract; the variable being optimized is the "realism" of the output, in other words its similarity to a sung sound recorded in the studio.

The search works by mimicking the biological process of evolution. Some random potential solutions are encoded into a population of *genotypes* (also referred to as *individuals*). The genotypes are evaluated for *fitness* (optimality). These are then copied to the next *generation*, but with preference given to the fitter individuals who are likely to receive more "offspring" while unfit individuals may receive none. *Mutation* and *crossover* operators are defined which operate on one or two genotypes respectively, and these operators are applied to the new generation. The process is iterated until the population converges to a solution.

GA searches can also be viewed as a quest for the highest (or another suitably high) peak on a "fitness landscape." The landscape may have many local optima, analogous to minor peaks. Any GA must strike a good balance between selective pressure (a tendency to climb the nearest peak) and diversity (a tendency to explore the landscape more), else a suitable solution will not be found. A good algorithm demonstrates the ability to converge to good solutions regardless of the random number seed.

When evolving 2-D model shapes, the solution comes in two parts: the fittest genotype contains a set of model parameters, defining its geometry, but we must also consider the sound they produce. As fitness is evaluated in terms of the latter, the produced sounds should all be similar; however similar sounds can be produced by dramatically different shapes so it will be interesting to note, over several algorithm runs, how similar our evolved shapes are.

III. System Design

A. Synthesis Engine

In our synthesis engine, a two dimensional rectilinear waveguide mesh[3] model is excited at one end with either: 1) white noise or 2) the signal from an electrolaryngograph (Lx) [19]. Output is recorded from the opposite end of the model. The LX signal represents the current flowing between two electrodes placed superficially on the neck at the level of the vocal folds, and it is usually interpreted as representing the change in vocal fold contact area. Whilst this is not directly representative of the acoustic excitation during voicing, it embodies many dynamic features associated with a voiced output.

The 2-D model is chosen as computation is relatively cheap—but some of the properties of a full three-dimensional (3-D) model are preserved. In particular, it has the ability to contain complex standing waves and allows greater control of boundary conditions, thus potentially offering more realistic synthesis than is achievable with LPC. It can also be considered as proof of concept for a 3-D model, which would allow for a direct mapping of vocal tract articulations to synthesis parameters.

A spatial resolution of 1.1 cm has been chosen, to model the vocal tract with a sufficient degree of accuracy. The sampling rate is dependant on the spatial resolution (see the Appendix) and is accordingly set to 44.1 kHz. Undersampling effects reduce this to 22 kHz. However, note that by by comparison, human speech typically has a bandwidth of 8 kHz.

The human vocal tract is 17 cm long in the average male, and is thus simulated by a model 16 nodes in length. However, 18 nodes are used in order to include a partial region outside the lips. The width is variable, but 9 nodes, not including the reflective edge, is chosen as a maximum. This gives a diameter of 9.9 cm, which is not exceeded in nature.

The walls of the mesh model have a coefficient of reflection $\lambda = 9/10$ while the mouth has $\lambda = -9/10$: these figures

[2]In many cases [17] a genetic algorithm with excessive selective pressure, i.e., a hill-climbing algorithm, is found to converge prematurely to a local optimum. This would not be possible on a convex problem.

[3]The digital waveguide mesh is described in the Appendix.

are chosen as they are close to those believed to be natural. A hardware implementation of an evolvable 2-D digital waveguide mesh is also under development as part of the POEtic project [20], using $\lambda = \pm 7/8$ which produces similar results but is more efficiently implemented.

B. Genetic Representation

A genetic representation is required to encode potential solutions into an evolvable genotype. A population of genotypes can then be evolved as described in Section II.

For simplicity, symmetry along the long axis is assumed, as is the presence of at least one normal node in the centre of the mesh model (otherwise no sound would pass through). Thus, the genotype is defined to be a string of integers g_i such that at each point, the width in nodes w_i of the mesh is

$$w_i = 1 + 2g_i. \tag{1}$$

To reduce the amount of information stored in the genome and thus the size of the search space, and also to create a smoother fitness landscape to aid convergence, not all of these widths are stored (we store w_0 then $w_1, w_3, \ldots, w_{15}, w_{17}$). The remaining widths are restored by linear interpolation.

C. Fitness Evaluation

A fitness function is needed to evaluate the effectiveness of evolved solutions.

In the algorithm presented, some two-track recordings are created in the studio, in which one track contains an ordinary recording of a sung vowel sound, while the other track is connected to an Lx—thus providing data on the excitation to the vocal tract used to produce the same sound.

The fitness is measured by exciting the 2-D model with the Lx signal and comparing the spectral content of the output to the desired vowel sound.

The spectral comparison is similar to that described in [21]. The individual being evaluated and the target sound are both normalized and transferred to the Fourier domain, where a mean absolute difference between the two spectra is computed (excluding dc components). Phase information is discarded. The optimal individual is thus defined to be the one with the lowest fitness.

Two minor variations on this were tried.

- A penalty was added to overly quiet sounds, this being a constant factor multiplied by the difference in peak levels of the target and individual sounds (before normalization);
- Phase information was included.

However, neither of these showed better results.

The region of signal evaluated was 2400 samples long, starting at the 10 000th sample. This was chosen as a representative part of the recording. It also means all frequencies above 18 Hz were analyzed. While the lower frequencies may not be audible or even present in the excitation signal, this was found to improve results purely because a greater portion of the output signal is sampled. Simulation and fitness evaluation took approximately 1.5 s per individual.

D. Selection and Mutation Operators

The GA used was generational; 50 generations each of 50 individuals produced good solutions. Thus, 0.2% of the search space was explored; giving GA runtimes in the region of one hour.

TABLE I
GA CONVERGENCE

Vowel	Base Fitness	Model Fitness	$\sigma_{spectra}$	σ_{shapes}
ah	0.0011	0.0006	0.0001	0.37
ii	0.0010	0.0005	0.0003	0.61
uu	0.0008	0.0006	0.0001	0.36

Universal Stochastic Sampling [22] with rank selection was used. The stochastic part of the process can be viewed as a lottery: a number of *tickets*, each representing an equal chance of being selected for the next generation, are allocated. The fittest individual I_1 receives 25 tickets, I_2 receives 24, and so on until I_{25} which receives one ticket. I_{26} to I_{50} all receive one ticket each. This system was found to produce a good tradeoff of selective pressure and diversity.

The mutation rate was governed by the 1/5 rule, as described in, e.g., [23]. However it was also limited to a maximum (probability) of 0.08 per gene per generation, giving an expected 0.8 mutations per individual. The procedure for mutating a gene was to select a different random integer for the width of the vocal tract at that point.

The crossover rate (probability) was 0.2 per individual. Crossover was implemented by splicing together two parents at a random point, thus producing a vocal tract consisting of a part from each parent.

IV. EXPERIMENTAL VALIDATION

A. Convergence

The results would be of reduced usefulness for synthesis, and of no use for analysis, if GA runs did not converge i.e., successive runs of the GA, with different seedings of the random number generator, produced dramatically different results. Thus, the convergence properties are given in Table I.

All results concerning spectral fitness and convergence are measures of the similarity of two signals, as defined by the fitness function. This metric is zero for identical signals. The following are given:

- *Base Fitness* or similarity of 2-D model input to target signal—the fitness of a mesh which does nothing.
- *Model Fitness* or similarity of the best output to target signal—a measure of what the GA has achieved.
- *Spectral Standard Deviation* $(\sigma_{spectra})$, or similarity of five successive program outputs to one another—a measure of the stability of the GA.[4]

The standard deviation of evolved 2-D model shapes (σ_{shapes}) is also given. As the range of model radii is from 0 to 4, this metric would be approximately 1 for uncorrelated shapes.

B. Evolved Results

Figs. 1–3 give spectral plots of three vowels, and their counterpart real vowels recorded in the studio. To give a clearer picture of the differences between real and evolved vowels, Fig. 4

[4]Experiments involving 15 runs of a single vowel have also been performed, with similar results.

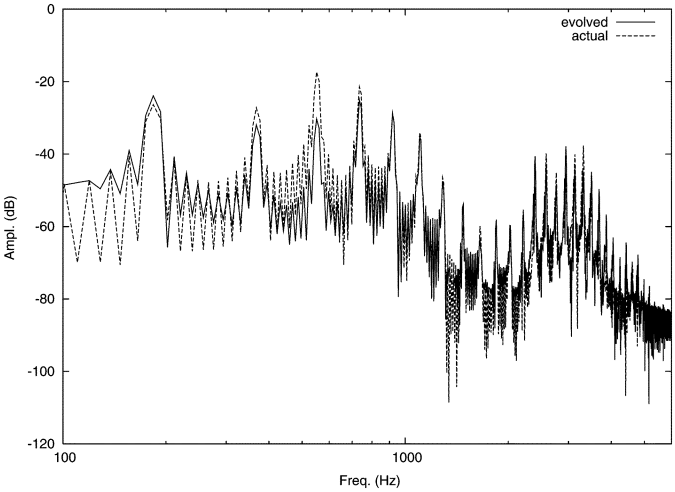

Fig. 1. Real and evolved spectra for an "ah" vowel.

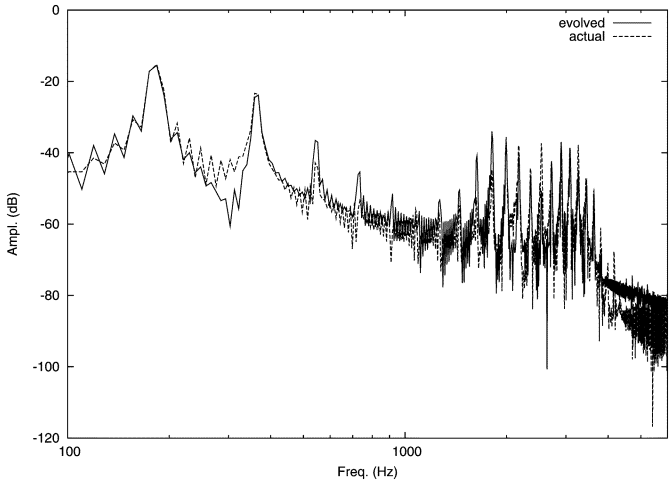

Fig. 2. Real and evolved spectra for an "ii" vowel.

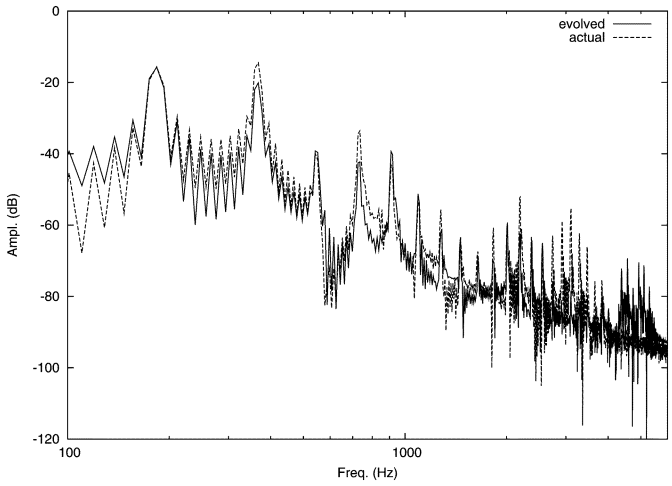

Fig. 3. Real and evolved spectra for an "uu" vowel.

shows the associated error for the "ii" vowel. Figs. 5–7 show the shapes evolved to produce these vowels.

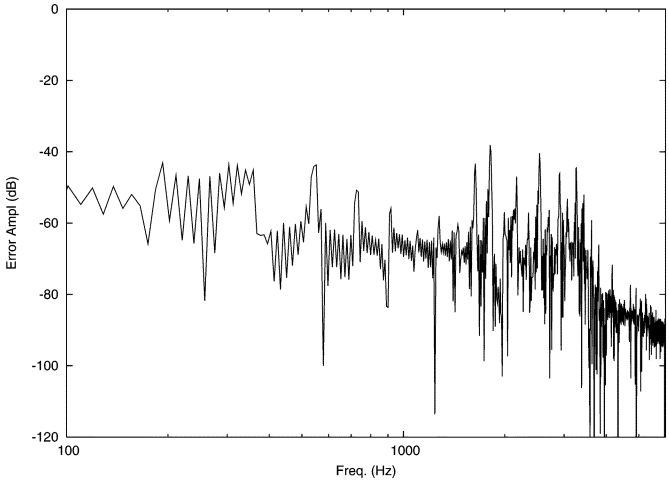

Fig. 4. Plot of the difference between a real and evolved "ii" vowel.

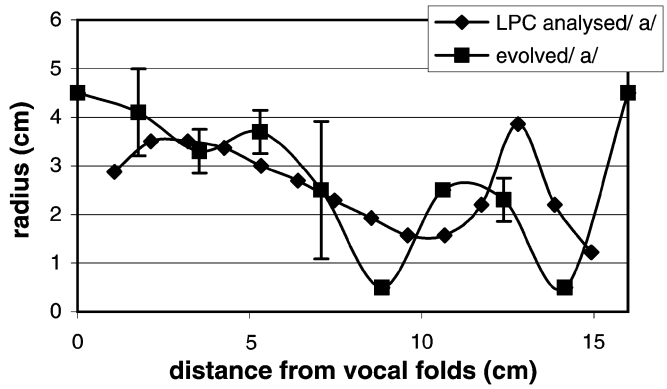

Fig. 5. Evolved 2-D model shape for the "ah" vowel. The error bars extend to one standard deviation either side of the mean.

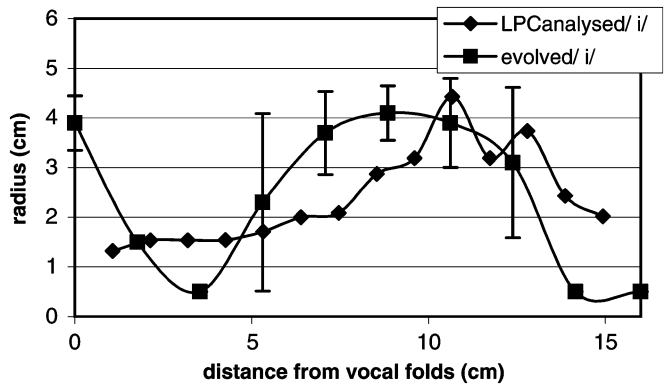

Fig. 6. Evolved 2-D model shape for the "ii" vowel.

V. Discussion

A. Synthesized Sounds

The sounds produced in this experiment can be heard online [24]. As can be seen from the spectra plotted in Figs. 1–3, the sounds are very similar (although not identical) to real recordings.

There are a number of ways in which this similarity might be improved. First, the physical model could be matched more accurately to reality in a number of ways, including:

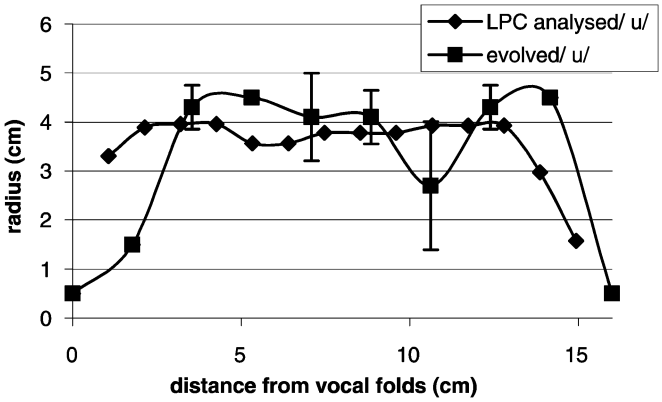

Fig. 7. Evolved 2-D model shape for the "uu" vowel.

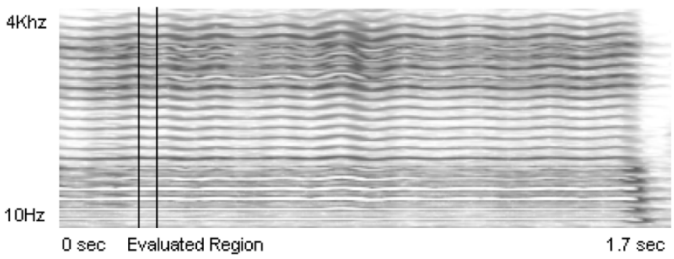

Fig. 8. Spectrogram of a real, sung "ah" vowel.

- implementing the model in 3-D not 2-D;
- deriving a more accurate model of the vocal folds, so that the exact signal produced by them can be known (rather than using the Lx signal which is at best an approximation);
- implementing a more sophisticated 2-D model—alternative decompositions of the 2-D plane have been proposed offering improvements over the rectilinear topology [25], as have improved boundary models [27], [28];
- implementing a higher-resolution 2-D model, but with more interpolation between the genome and mesh so as not to increase the search space;[5]
- implementing the nonlinear loss which is known, in reality, to occur at the boundaries of the vocal tract.

Also, while the spectral content of the recorded vowel sounds remains relatively constant over the duration of the recording, it may be necessary to model small changes in the sound to improve the model. Fig. 8 shows the spectrogram of the "ah" vowel used in the experiment; note how as time progresses the formant peak around 3 kHz rises slightly. Formant peaks are also known to vary when vibrato is present, as in this signal. The modeling of such changes may not be necessary to produce a vowel sound acceptable to most listeners (although including them will doubtless improve realism)—but given that all time slices of the signal are not equal, an evolved sound may be perceived as unnatural if it has been evolved to match the "wrong" part of the spectrogram.

Ideally, such improvements would be provided by a model which changes in shape over time. This is also a natural extension to the project and indeed a prerequisite for producing a complete singing synthesizer as we need to be able to morph from one vowel to another and produce consonants. However, the question of how to model dynamic changes in shape is an ongoing area of research.

[5]Better interpolation could be achieved with curve fitting, and also variable resolution/accuracy of the stored data points, weighted toward greater accuracy where vocal tracts vary more from individual to individual.

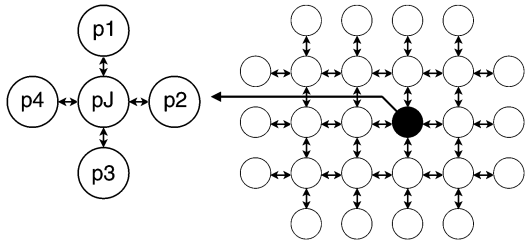

Fig. 9. Two-dimensional rectangular-grid mesh, showing edge nodes and the finite difference equation variables for a single node.

The final feature needed to complete a singing synthesizer is intuitive control of pitch. Pitch in our system is determined by the pitch of the Lx waveform, so this could easily be adjusted either by using traditional multisampling techniques, or by use of a physical model for the larynx.

B. Evolved Shapes

One thing known for certain since early experiments [17] is that there are many dramatically different shapes which produce similar sounds—for the purposes of our GA, lots of local optima—and while it may be possible to evolve "natural" shapes, the possible nonuniqueness of any solution should be considered when performing shape analysis experiments with this system. However, we deduce from the level of convergence shown in Table I that the algorithm finds reasonably similar optima (perhaps close to the global optimum) on each run: thus, analysis of the evolved shapes is meaningful.

It is clear that the shapes (Figs. 5–7) are not close to those produced in real human vocal tracts, as they exhibit features known not to be possible in reality. For example, the "ah" vowel displays a radius of 5 cm at the larynx. As mentioned in the introduction, convergence to the same solution found in nature is not guaranteed. However, in this case, the differences can be attributed to differences between the model and reality as detailed above. In particular, the use of Lx as an excitation is not a direct substitute for glottal flow, but an approximation. The GA will have evolved the shapes to compensate for the Lx-glottal flow differences and thus differ from natural data.

VI. Conclusion

A 2-D physical model based on the digital waveguide mesh is a promising tool for the synthesis and analysis of speech and singing. However, real data on the shape of the human vocal tract while in use is hard to obtain. Evolution has been shown to be an effective alternative design method for the shape of such models. Realistic sounds are produced through synthesis, even though the synthesis model used is fairly simple.

If the synthesis model can be improved to better reflect reality, it should be possible in the future to use such models to recreate the exact shape of the vocal tract of a given singer—for both improved synthesis (including articulations) and analysis. Both

of these application areas require the evolved model to match the actual vocal tract, so the possible nonuniqueness of solutions is a concern: it is not proven that two different vocal tracts cannot produce *exactly* the same sound. However, limiting the evolved models to shapes which we know humans can definitely achieve, and sampling the output signal in great detail for a long period of time, may reduce the search space sufficiently that a correct solution is overwhelmingly likely.

Finally, it is suggested that evolution is an effective design technique not only in this and other applications where we wish to produce an accurate simulation of reality, but also for the synthesis of new sounds not available acoustically. A multi-dimensional waveguide mesh, for example, could easily simulate a 10-dimensional object,[6] but designing the actual shape of such an object would be extremely difficult. Evolution allows us to skip the design phase, instead exploring the search space until we find a sound we like.

APPENDIX
DIGITAL WAVEGUIDE MESH

A. Introduction

The digital waveguide mesh (DWM) [31] is a discrete-time simulation used to model acoustic wave propagation in an enclosed system. It can be considered as an extension of the 1-D digital waveguide commonly used to model string and wind instruments [8], an approach similar to the Kelly–Lochbaum 1-D transmission line simulation of the vocal tract [6]. Both of these 1-D models are founded in a discretized formulation of the d'Alembert solution to the wave equation through the use of bi-directional digital delay lines and scattering junctions. However, a direct numerical solution to the wave equation using second-order finite differences leads to the alternative implementation of the DWM as a finite difference time domain (FDTD) simulation. Both approaches have been employed in DWM research and recent work has explored the equivalence between these two models [32]. The FDTD approach is computationally efficient in terms of memory and processing time, although is exact only at dc, whereas the direct implementation of the DWM will propagate bandlimited solutions to the wave equation without error [33]. However, the implementation of boundary conditions in both cases is quite different, with the FDTD approach in particular being susceptible to problems relating to instability, although this can be solved in some part through the use of hybrid mesh structures [32], [34]. A thorough comparison of the equivalences between these schemes is presented in [32] and [33]. The implementation used in this paper is the FDTD approach although it is still referred to as a digital waveguide mesh as its background lies in this tradition in previously published studies, e.g., [29].

B. Two-Dimensional DWM

The 1-D digital waveguide is a discretized formulation of the d'Alembert travelling wave solution to the 1-D wave equation

$$\frac{d^2y}{dt^2} = c^2\frac{d^2y}{dx^2} \Rightarrow y(x) = \psi_L(x+ct) + \psi_R(x-ct) \qquad (2)$$

which shows that a 1-D wave in a medium of constant impedance can be decomposed into two separate signals travelling in opposite directions.

This can be implemented using bidirectional delay lines such that the sound pressure of a propagating wave signal can be defined as the sum of these travelling waves. A 2-D digital waveguide mesh is constructed from a regular array of such 1-D digital waveguides connected via scattering junctions. By determining that for a lossless junction J, the sum of the input velocities is equal to the sum of the output velocities (flows add to zero), and that the sound pressures in all crossing waveguides are equal (continuity of pressure or force), the sound pressure p_J at J for N connected neighbors at unit distance can be derived as the following difference equation:

$$p_J(t) = \frac{2}{N}\sum_{i=1}^{N} p_i(t-\Delta t) - p_J(t-2\Delta t). \qquad (3)$$

A number of different 2-D mesh topologies have been proposed, corresponding to different decompositions of the 2-D plane, and hence the number of neighbors, N. This work uses the 2-D rectilinear digital waveguide mesh, such that $N = 4$. Note that (3) is valid for a mesh of any topology or dimensionality. For instance, the same expression for the 2-D rectilinear mesh with $N = 4$ can be used for the 3-D tetrahedral mesh [26], the only difference is in terms of the implementation and spatial arrangement of the neighboring junctions. Unlike the 1-D case, this 2-D implementation is not an exact approximation to wave propagation in the continuous domain as a signal will not propagate equally in all directions [31]. However, a high-resolution structure can provide a good approximation.

C. Boundary Conditions

A multidimensional mesh structure is typically terminated at a boundary via a single 1-D connection and will act to reflect an incident sound wave via a change in the impedance of the different waveguide elements connected at the boundary scattering junction [35]. The simplest case can be considered by connecting a dummy junction J on the other side of the boundary junction i, essentially within the boundary itself. This leads to the following formulation for a boundary junction based on a FDTD implementation of a DWM:

$$p_J(t) = (1+\lambda)p_i(t-\Delta t) - \lambda p_J(t-2\Delta t). \qquad (4)$$

The amount of energy reflected at the boundary is determined by setting $-1 \leq \lambda \leq 1$, with $\lambda = 1$ giving total reflection and $\lambda = 0$ approximating total absorption.

D. Sampling Rate

The sampling rate of a waveguide mesh is determined by (5) where c is the speed of wave propagation in the medium, N is the number of dimensions, and Δx is the internodal distance

$$\Delta t = \frac{\Delta x}{c\sqrt{N}}. \qquad (5)$$

However, the valid bandwidth of a digital waveguide mesh is actually much lower than the limit suggested by (5). Dispersion error, where the velocity of a propagating wave is dependent upon both its frequency and direction of travel, leads to

[6]This need not be complex, indeed could contain as few as 11 elements, but could be interesting as it breaks the constraints of 3-D space.

wave propagation errors and a mistuning of the expected resonant modes.

The degree of dispersion error is highly dependant upon mesh topology and has been investigated in [26]. Interpolated [29] and triangular [34] mesh topologies demonstrate dispersion characteristics that are reduced substantially to a function of frequency only. Additional pre- and post-processing of results using frequency warping techniques [29], [30] gives further significant improvements, increasing the overall valid bandwidth of the model. Oversampling the mesh also offers improvements in this regard, such that the required bandwidth lies within accepted limits, typically $0.25 \times 1/\Delta t$[31].

ACKNOWLEDGMENT

The information provided is the sole responsibility of the authors and does not reflect the Community's opinion. The Community is not responsible for any use that might be made of data appearing in this publication.

Data analysis and the production of plots for this paper was performed with GNU Octave, an open-source high level numerical computation language.

REFERENCES

[1] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Comput. Music J.*, vol. 20, no. 3, pp. 38–46, 1996.
[2] H. Dudley, "The vocoder," *Bell Lab. Rec.*, Dec. 1939.
[3] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[4] B. Atal, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 47, pp. 65(A)–65(A), 1970.
[5] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, 1987.
[6] J. Kelly and C. Lochbaum, "Speech synthesis," in *Proc. 4th Int. Congr. Acoustics*, 1962, pp. 1–4.
[7] P. Cook, "SPASM: A real-time vocal tract physical model editor/controller and singer; the companion software synthesis system," *Comput. Music J.*, vol. 17, no. 1, pp. 30–44, 1992.
[8] J. O. Smith, "Physical modeling using digital waveguides," *Comput. Music J.*, vol. 16, no. 4, pp. 74–87, 1992.
[9] J. Mullen, D. M. Howard, and D. T. Murphy, "Acoustical simulations of the human vocal tract using the 1D and 2D digital waveguide software model," in *Proc. DAFx-04*, Naples, Italy, Oct. 5–8, 2004, pp. 311–314.
[10] ——, "Waveguide physical modeling of vocal tract acoustics: Improved formant bandwidth control from increased model dimensionality," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, Jul. 2006, to be published.
[11] ——, "Digital waveguide mesh modeling of the vocal tract acoustics," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 119–122.
[12] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 104, no. 1, pp. 471–487, 1996.
[13] B. H. Story, "Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics," in *Proc. Stockholm Music Acoustics Conf.*, 2003, SMAC-03, pp. 435–438.
[14] Markel and Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.
[15] D. P. Rossiter, D. M. Howard, and M. Downes, "A real-time LPC-based vocal tract area display for voice development," *J. Voice*, vol. 8, no. 4, pp. 314–319, 1995.
[16] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, pt. II, vol. 2, no. 1, pp. 133–150, Jan. 1994.
[17] C. Cooper, D. Howard, and A. Tyrrell, "Using GA's to create a waveguide model of the oral vocal tract," in *Proc. Applications of Evolutionary Computing, EvoWorkshops*, 2004, pp. 280–288.
[18] T. Baeck, D. Fogel, and Z. Michalewicz, *Evolutionary Computation*. New York: Inst. Phys, 2000.
[19] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: A tutorial," *Clin. Linguist. Phon.*, vol. 3, pp. 281–296, 1989.
[20] A. Tyrrell, E. Sanchez, D. Floreano, G. Tempesti, D. Mange, J.-M. Moreno, J. Rosenberg, and A. E. P. Villa, "POEtic tissue: An integrated architecture for bio-inspired hardware," in *Lecture Notes in Computer Science*, 2003, vol. 2606, pp. 129–140.
[21] Garcia and A. Ricardo, "Automating the design of sound synthesis techniques using evolutionary methods," in *Proc. DAFX*, Limerick, Ireland, 2001, pp. 1–6.
[22] J. E. Baker, "Reducing bias and inefficiency in the selection algorithm," in *Proc. 2nd Int. Conf. Genetic Algorithms*, 1987, pp. 14–21.
[23] FAQ for Usenet Group comp.ai.genetic, Part 6. [Online]. Available: http://www-2.cs.cmu.edu/Groups/AI/html/faqs/ai/genetic/part6/faq-doc-6.html
[24] Audio Clips. [Online]. Available: http://www.bioinspired.com/users/cc26/poeticaudio
[25] G. Campos and D. M. Howard, "A paralell 3-D digital waveguide mesh model with tetrahedral topology for toom acoustic simulation," in *Proc. DAFX*, Verona, Italy, 2000, pp. 73–78.
[26] S. A. Van Duyne and J. O. Smith, "The 3D tetrahedral digital waveguide mesh with musical applications," in *Proc. Int. Computer Music Conf.*, Hong Kong, 1996, pp. 9–16.
[27] A. Kelloniemi, D. T. Murphy, L. Savioja, and V. Välimäki, "Boundary conditions in a multi-dimensional digital waveguide mesh," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, QC, Canada, May 17–21, 2004, pp. IV-25–IV-28.
[28] A. Kelloniemi, L. Savioja, and V. Vlimki, "Spatial filter-based absorbing boundary for the 2-D digital waveguide mesh," *IEEE Signal Process. Lett.*, vol. 12, pp. 126–129, Feb. 2005.
[29] L. Savioja and V. Valimaki, "Interpolated rectangular 3-D digital waveguide mesh algorithms with frequency warping," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 783–789, Nov. 2003.
[30] ——, "Reducing the dispersion error in the digital waveguide mesh using interpolation and frequency warping techniques," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 184–194, Mar. 2000.
[31] S. A. Van Duyne and J. O. Smith, "Physical modeling with the 2-D digital waveguide mesh," in *Proc. Int. Computer Music Conf.*, Tokyo, Japan, 1993, pp. 40–47.
[32] M. Karjalainen and C. Erkut, "Digital waveguides versus finite difference structures: Equivalence and mixed modeling," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 978–989, 2004.
[33] J. O. Smith. (2004) On the Equivalence of Digital Waveguide and Finite Difference Time Domain Schemes. [Online]. Available: http://arxiv.org/abs/physics/0407032.
[34] M. J. Beeson and D. T. Murphy, "Roomweaver: A digital waveguide mesh based room acoustics research tool," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFX)*, Naples, Italy, Oct. 5–8, 2004, pp. 268–273.
[35] L. Savioja, T. J. Rinne, and T. Takala, "Simulation of room acoustics with a 3-D finite difference mesh," in *Proc. Int. Computer Music Conf.*, 1994, pp. 463–466.

Crispin Cooper received the B.A. (Hons.) degree in computer science from the University of Cambridge, Cambridge, U.K., in 2002.

In 2003, he joined the Department of Electronics at the University of York, Heslington, U.K., as a Research Associate on the EU-funded POEtic project. The project aims to create a hardware platform supporting evolution, growth, and learning; he has been applying its capabilities to audio technology.

Mr. Cooper won a Best Paper Award at EvoIASP2004, Portugal, for his work on modeling the vocal tract.

Damian Murphy received the B.Sc. (Hons.) degree in mathematics in 1993, the M.Sc. degree in music technology in 1995, and the D.Phil. degree in music technology in 2000, all from the University of York, York, U.K.

In 1999, he was Lecturer in music technology in the School of Engineering, Leeds Metropolitan University, Leeds, U.K., and in 2000 was appointed as Lecturer in the Department of Electronics, University of York. He has worked as an independent audio consultant and since 2002 has been a Visiting Lecturer in the Department of Speech, Music and Hearing, KTH, Stockholm, Sweden. His research is in the areas of physical modeling and spatial sound, with particular interests in applications of the multidimensional digital waveguide mesh. He is an active composer in the fields of electroacoustic and electronic music, where sound spatialization forms a critical aspect of his musical works.

Dr. Murphy is a member of the Audio Engineering Society. In 2004, he was appointed as one of the U.K.'s first AHRB/ACE Arts and Science Research Fellows, investigating the compositional and aesthetic aspects of sound spatialization and acoustic modeling techniques.

David Howard received the First Class Honors degree in electrical and electronic engineering from University College, London, U.K., in 1978 and the Ph.D. degree in cochlear implants from the University of London in 1985.

He became a Lecturer in speech and hearing sciences at University College in 1979, and he moved to the University of York in 1990. He became Personal Chair in Music Technology in 1996. His research interests include the analysis and synthesis of singing, music, and speech.

Dr. Howard is a Chartered Engineer, a Fellow of the Institution of Electrical Engineers, a Fellow of the Institute of Acoustics, and a Member of the Audio Engineering Society.

Andy Tyrrell (SM'96) received the First Class Honors degree in 1982 and the Ph.D. degree in 1985, both in electrical and electronic engineering.

He joined the Electronics Department, York University, York, U.K., in April 1990; he was promoted to the Chair in Digital Electronics in 1998. Previously, he was a Senior Lecturer at Coventry Polytechnic, Coventry, U.K. From August 1987 and August 1988, he was Visiting Research Fellow at Ecole Polytechnic Lausanne, Switzerland, where he was researching into the evaluation and performance of multiprocessor systems. His main research interests are in the design of biologically inspired architectures, artificial immune systems, evolvable hardware, FPGA system design, parallel systems, fault-tolerant design, and real-time systems. In particular, over the last six years his research group at York has concentrated on bio-inspired systems. This work has included the creation of embryonic processing arrays, intrinsic evolvable hardware systems, and the immunotronics hardware architecture. He is Head of the Intelligent Systems Research Group at York and Head of the Department. He has published over 160 papers in these areas.

Dr. Tyrrell was General Programme Chair for ICES 2003 and Programme Chair for IPCAT2005.