Article:
Mullen, Jack, Howard, David M. and Murphy, Damian T. (2007) Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model. IEEE Transactions On Audio Speech And Language Processing. pp. 577-585. ISSN: 1558-7916

https://doi.org/10.1109/TASL.2006.876751

Real-Time Dynamic Articulations in the 2-D Waveguide Mesh Vocal Tract Model

Jack Mullen, David M. Howard, and Damian T. Murphy

***Abstract*—Time domain articulatory vocal tract modeling in one-dimensional (1-D) is well established. Previous studies into two-dimensional (2-D) simulation of wave propagation in the vocal tract have shown it to present accurate static vowel synthesis. However, little has been done to demonstrate how such a model might accommodate the dynamic tract shape changes necessary in modeling speech. Two methods of applying the area function to the 2-D digital waveguide mesh vocal tract model are presented here. First, a method based on mapping the cross-sectional area onto the number of waveguides across the mesh, termed a *widthwise mapping* approach is detailed. Discontinuity problems associated with the dynamic manipulation of the model are highlighted. Second, a new method is examined that uses a static-shaped rectangular mesh with the area function translated into an impedance map which is then applied to each waveguide. Two approaches for constructing such a map are demonstrated; one using a linear impedance increase to model a constriction to the tract and another using a raised cosine function. Recommendations are made towards the use of the cosine method as it allows for a wider central propagational channel. It is also shown that this *impedance mapping* approach allows for stable dynamic shape changes and also permits a reduction in sampling frequency leading to real-time interaction with the model.**

***Index Terms*—Acoustic impedance, acoustic resonators, acoustic waveguides, speech synthesis, vocal system.**

I. INTRODUCTION

CURRENT technologies in artificial speech generation are at a stage of near perceived realism. Many state of the art text-to-speech (TTS) systems use sample-based concatenative synthesis [1], [2]. Phonemes taken from recorded natural speech are spliced together to create new words and sentences not present in the original utterance. The recordings are of a spoken voice, typically that of an actor, reading aloud for several hours. A large database is constructed as each possible diphone (the transition between the middle of one phoneme and another) is extracted and categorized to be recalled for concatenation at a later time. Such a scheme is, however, restricted to regeneration only of sounds present in the original recordings. Even limited processing of the signal may prove detrimental to the resulting naturalness. Furthermore, the user may only communicate with the vocal identity provided with the system.

Manuscript received September 26, 2005; revised February 17, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

The authors are with the Audio Laboratory, Intelligent Systems Group, Department of Electronics, York University, Heslington, York, YO10 5DD, U.K. (e-mail: jm220@ohm.york.ac.uk; dmh@ohm.york.ac.uk; dtm3@ohm.york.ac.uk).

Digital Object Identifier 10.1109/TASL.2006.876751

Articulatory vocal tract modeling attempts to recreate the behavior of the human speech apparatus to simulate the process of speaking, rather than simply the sounds it generates. The tract is an acoustic resonator with various mouth features that constantly alter, constrict and stop the way vibrations travel through it. These articulations and tract shape changes, combined with the glottal source produce the sounds we perceive as speech. An effective vocal tract model will display the ability to dynamically adapt, accommodating the changes in the tract area function. Such a dynamic model can be used to simulate a diphthong—a slide between two vowels, for example /aU/ in the word house. Area function changes can also be made to represent constrictions to the air flow in the model, giving lateral articulation such as the /l/ in the word *lip*. Plosive articulations can be modeled in a similar way, forcing a momentary stop and then release of the air flow, such as the /p/ in the word *put*.

Frequency-domain articulatory modeling makes use of one-dimensional (1-D) area function data to parameterize the tract into a cascaded series of filters, each representing the different sections of the tract [3]. The transfer functions of the filters are adjusted according to the associated tract movements to simulate the manipulations to the airflow. More recent synthesis models have been developed that use three-dimensional (3-D) shape data of the complex system to generate the transfer functions [4], [5]. However, numerical simulation of the actual vibrations within the system is reduced to a 1-D representation for simplicity and real-time response.

Equivalent time-domain articulatory vocal tract models have been developed, where the wave motion itself is directly synthesized. Many studies have been conducted into the use of a 1-D chain of waveguides to simulate the resonances of the tract [6]–[8]. More recent research has been focused on improving the underlying propagational model to give better simulation of the resonant cavity [9], [10]. Research into higher dimensionality of the wave propagation model has shown that equivalent formant patterns to the existing 1-D models can be generated using a two-dimensional (2-D) waveguide mesh [11]. It has also been detailed that formants produced by the mesh model follow an approximately linear bandwidth variation in response to changes in boundary reflection parameters [12]. As such, the 2-D waveguide mesh tract has been presented as an alternative development to the original 1-D piecewise acoustic tube model, parallel to advances based on enhanced order area function approximation and improving the planar wave propagation mechanism. In addition, the extra dimensionality allows for simulation of cross-tract modes, and the modeling of the split in the air channel used in creating lateral sounds, such as /l/.

Current articulatory models have been used to synthesize simple words with near realism, but the method is not

comparable to concatenative synthesis at its current state of development. The complicated and variable nature of the movements of mouth features such as the tongue and lips need to be parameterized and applied to an accurate, dynamically adaptable model of the vocal tract cavity. In order to confirm the application of multidimensional signal processing techniques to time-domain articulatory tract modeling as a potentially advantageous research direction, dynamic ability must first be validated.

Work presented in this paper demonstrates further developments in 2-D waveguide vocal tract modeling. A new method for area function application is demonstrated which allows for stable dynamic changes to be made to the tract shape. Two methods of introducing the effects of the area function onto the tract model are examined. Furthermore, this adaptation also allows for a reduction in system sampling frequency, which leads to real-time performance. This paper is organized as follows, Section II introduces the topic of synthesis using physics based representation. The notion of the acoustic waveguide is introduced in 1-D and then expanded into the 2-D mesh. This theory is used to outline the structure of the 1-D time-domain articulatory tract model in Section III. Section IV recounts how the 2-D mesh is used to synthesize the acoustics of the vocal tract, with reference to the problems associated with making on-the-fly changes to the area function. New techniques for dynamic articulatory synthesis using the 2-D model are given in Section IV-B. Section IV-C then shows how those techniques also allow for real-time performance to be achieved in the model.

II. Physical Modeling Synthesis

Numerical simulation of any real-world process requires a valid discretisation of the problem domain, and a definition of physics derived laws governing behavior within the system. In the context of acoustical physical modeling much focus has been placed on the use of the 1-D digital waveguide for real-time synthesis of bores and pipes. The extension of this technique towards a 2-D and 3-D digital waveguide mesh can be used in a multidimensional simulation of acoustic wave propagation within a membrane, rigid structure or air cavity, although the increased processing load can result in non realtime performance.

A. One-Dimensional Digital Waveguide

The digital waveguide physical model defines the unit element within a 1-D system to be a bi-directional digital delay line [13]. The units are connected together in a chain or *ladder* configuration as demonstrated in Fig. 1. This discrete form assumes the system to be of a linear time invariant nature.

Based on a discrete version of the d'Alembert solution to the one-dimensional wave equation, the total pressure $p(x, nT)$ at the waveguide at distance x (or a number of waveguide unit lengths nd) along the chain and at time interval T is the sum of left-going (p^-) and right-going (p^+) components at each time step as in (1), where c is the wave speed

$$p(x, nT) = p^-(x + cnT) + p^+(x - cnT). \tag{1}$$

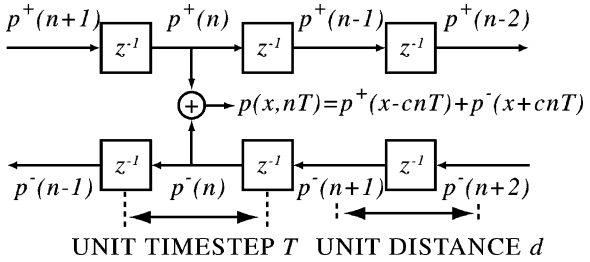

Fig. 1. One-dimensional chain of waveguides

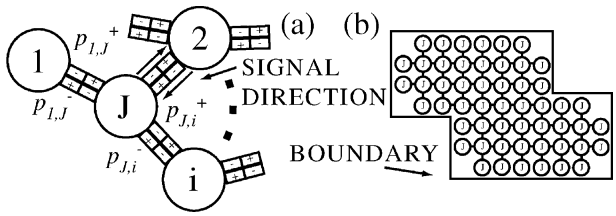

Fig. 2. (a) Unit junction and (b) a rectilinear mesh.

Application of an input to the 1-D system and then continuous iteration of scattering and timestep equations to each element constitutes propagation of a traveling wave through the modeled medium.

B. Two-dimensional Digital Waveguide Mesh

This 1-D case can be extended to create a *lattice* of waveguides, or a digital waveguide mesh (DWM) resulting in a 2-D representation of the propagating medium [14]. In the mesh, scattering junctions are created where multiple waveguides meet. The basic DWM configuration is the rectilinear mesh, where junctions are placed at regular intervals on a cartesian-coordinate grid, such that each has 4 neighboring junctions at 90° from one another. Fig. 2(a) and (b) detail the scattering junction with i arbitrary connections and the formation of the rectilinear mesh, respectively.

In Fig. 2(a), air pressure values labeled $p^+_{J,i}$ indicate an incoming pressure at node J from node i (at i a unit time-step before), and those labeled $p^-_{J,i}$ show the outgoing pressure at node J, to node i (reaching node i a time-step later). As in (1), the pressure $p_{J,i}$ on each waveguide is then the sum of its two components

$$p_{J,i} = p^+_{J,i} + p^-_{J,i}. \tag{2}$$

The pressure p at each junction J with N intersecting waveguides, each of impedance Z_i is [14]

$$p_J = 2\frac{\sum_{i=1}^{N} \frac{p^+_{J,i}}{Z_i}}{\sum_{i=1}^{N} \frac{1}{Z_i}}. \tag{3}$$

The time-step, n, is then incremented to distribute all junction output pressures along waveguides to become neighboring junction input pressures

$$p^+_{J,i}(n) = p^-_{i,J}(n-1). \tag{4}$$

Scattering (3) can also be derived as an equivalent finite difference algorithm [15]. This mathematical simplification removes the terms involving incoming and outgoing waveguide pressures, reducing junction parameters to just pressure values and time indices. This results in a mesh implementation that shows significant improvements in terms of memory requirements and computation time [16].

Mesh boundaries are typically simulated using scattering equations derived from impedance matching techniques. A proportional amount of signal that is incident upon the boundary impedance Z_2 is reflected back into the mesh of impedance Z_1. This leads to a reflection coefficient $r = (Z_2 - Z_1/Z_2 + Z_1)$. The pressure on a single connection boundary node, as in node 1 in Fig. 2(a), is

$$p_1 = (1 + r)p^+_{1,J}. \tag{5}$$

Such one-port boundaries provide accurate reflection only to wavefront components that are parallel to the connecting waveguide and therefore perpendicular to the boundary. The remaining content of the incident wave receives an approximation to the reflection. However, these effects are only particularly evident at low reflections ($r \approx 0$). Recent developments in boundary modeling have seen the use of impedance layers and spatial filters to reduce the directional dependency [17].

The sampling frequency of the N-dimensional mesh is determined by the distance represented by each waveguide element, d and the wavespeed c

$$f_s = \frac{1}{Ts} = \frac{c\sqrt{N}}{d}. \tag{6}$$

The ability of the rectilinear mesh to perform uniform scattering, however, deteriorates as a function of direction and frequency due to dispersion error [14]. Furthermore, its square construction limits the valid frequency output content to $f_s/4$ [14]. Alternative methods of mesh construction have resulted in the development of triangular [18] and bilinearly deinterpolated [19] topologies, both of which reduce the problem to within acceptable levels. Frequency dependent dispersion error can be compensated for by the inclusion of frequency warping, where additional processing of the input and output signals is used to adjust for unwanted frequency shifts in the spectrum [19].

A further extension of the waveguide modeling technique can be used to implement a 3-D model of a resonating cavity. Waveguide structures of various topology can be used to create models of small cavities or large acoustical spaces such as a room or concert hall [16], [20], [21]. Accurate simulation of a source within the space can be achieved with either direct injection or convolution with the room impulse response (RIR) measured from the mesh.

III. One-Dimensional Waveguide Vocal Tract Model

A thorough treatment of the theory and construction of the 1-D piecewise vocal tract model can be found in the relevant literature [7], [8], [22], [23], [24], and in that presented previously

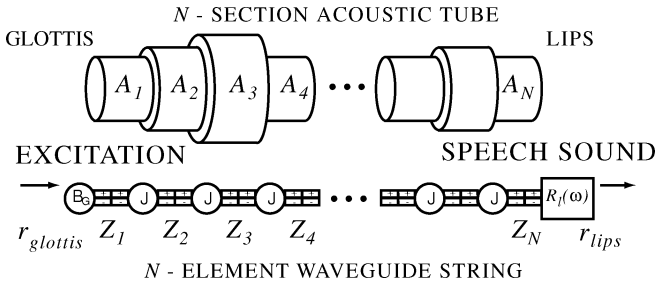

Fig. 3. One-dimensional waveguide vocal tract model.

form this work [12]. This section, therefore, is intended purely as an overview of the fundamental model, giving also a brief introduction to more recent improvements.

The acoustical properties of the human vocal tract can be modeled by considering it to be, at its simplest level, a straight tube from the glottis to the lips. The air column within is discretized and represented as a series of connected 1-D waveguides, as demonstrated in Fig. 3. The varied shape of the tract along its length, quantified as an *area function*, is modeled in the changing acoustic impedance Z_i of each waveguide relating to cylindrical tube section i. Frequency dependent reflections accounting for lip radiation and glottal reflection are placed at each end. With the introduction of glottal excitation, such as the three-mass model [25], or the LF waveform [26], speech sound emanates from the lip end.

The reciprocal relationship between cross-sectional area A_i and impedance Z_i and how they relate to both wavespeed c, air density ρ, and pressure p and volume velocity v for each waveguide is described in (7). Simply put, tract changes such as a constriction (decreased area) are modeled with an increased impedance around that point. Kelly-Lochbaum scattering junctions are used at the intersection between each waveguide [6]. Waveguide impedances to either side are used to define signal reflection and transmission in each direction

$$Z_i = \frac{p}{v} = \frac{\rho c}{A_i}. \tag{7}$$

The 1-D model in this form amounts to a connected series of cylindrical tube elements. This could be considered a spatially sampled system at *zeroth* order. The use of waveguides representing propagation in conical tube elements has been presented as an improvement to this model [9]. Benefits offered by the scheme are a first-order area function approximation and signal scattering derived from spherical, rather than planar wave propagation. However, it has been noted that the conical tract model introduces additional processing demands on a system, equal to those would result in a doubling of spatial resolution in the cylindrical version, offering no further improvements in accuracy [27]. Furthermore, the filters needed to facilitate conical wave propagation are unstable in certain conditions. This can be solved with a further enhancement to the propagation mechanism where the use of a discretized form of Webster's horn equation may be used to redefine the underlying algorithms to better approximate the modeled space [10].

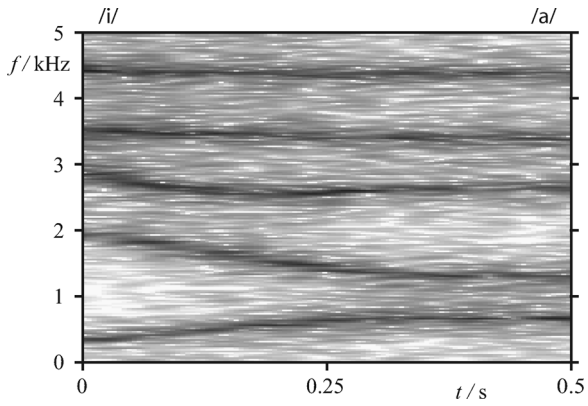

Fig. 4. One-dimensional /i/ to /a/ vowel slide spectrum

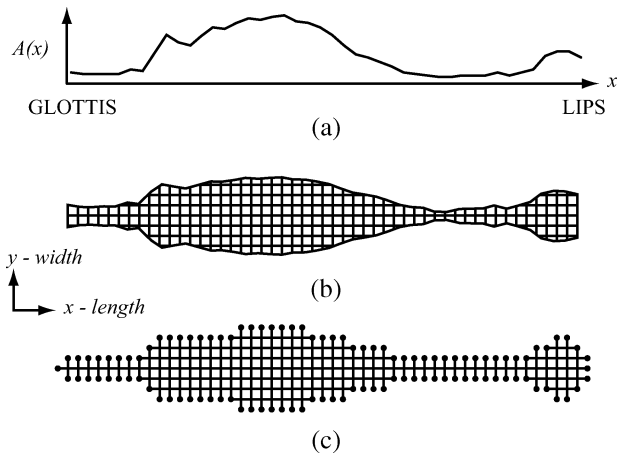

Fig. 5. Forming the widthwise /i/ vowel waveguide mesh. (a) Cross-sectional area function, (b) diameter map, (c) spatial discretization.

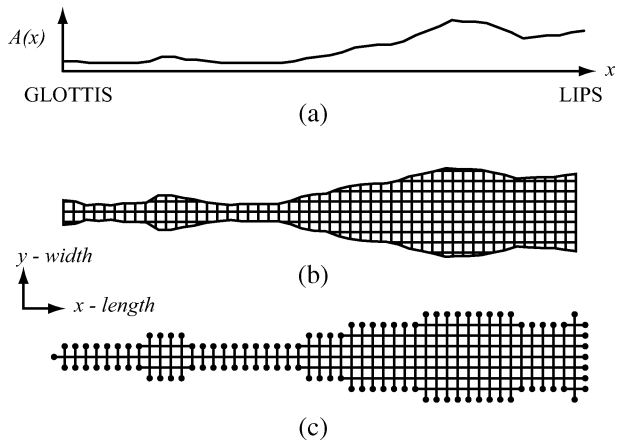

Fig. 6. Forming the widthwise /a/ vowel waveguide mesh. (a) Cross-sectional area function, (b) diameter map, and (c) spatial discretization.

A. Dynamic Articulations

Area function updates are applied directly to the model as changes in waveguide impedance. Linear interpolation between start and target values is used to avoid discontinuities in the resulting waveform. The spectrogram in Fig. 4 illustrates the change in formant pattern generated when noise is applied to a 1-D waveguide model undergoing a dynamic slide between /i/ and /a/ vowel area functions [28]. Glottal input to the same simulation results in a a vocal-slide between the two modeled vowels.

IV. Two-Dimensional Waveguide Vocal Tract Model

Acoustic wave propagation in the vocal tract can also be modeled using the 2-D DWM. The shape of the air cavity between the glottis and the lips is used to generate a waveguide mesh with resonant behavior approximating that of the modeled vowel. Resulting spectra should also include higher order cross-tract propagational modes inherent in the higher dimensional representation, and therefore present synthesis of increased accuracy. Two methods of mapping the area function onto the mesh model are presented as follows.

A. Widthwise Mapped Dynamic Articulations

The 1-D area function is generated using 3-D MRI scans of the tract held in various vowel positions [28]. The complex shapes generated are condensed into a series of equivalent circular cross-sectional areas along the length of the tract. In order to map the shape data onto a 2-D mesh the cross-sectional area value A at each length-wise spatial sampling instant is converted into a tube radius r. A mesh representing a 2-D plane through the tract from the glottis to the lips is constructed such that its diameter, defined as a number of waveguides across the tube width, follows this relationship. As such, the mesh uses 1-D area function data extended into a symmetrical 2-D form. This process is illustrated in Figs. 5 and 6. In each, the area function is shown in (a), followed by the diameter map (b), and then how this is discretized into a 2-D mesh (c).

The spectrogram in Fig. 7 illustrates the change in formants observed during the noise-excited dynamic slide between the /i/ [Fig. 5(c)] and /a/ [Fig. 6(c)] vowel mesh shapes. The overlayed dotted lines taken from the formant peaks in Fig. 4 verify the 2-D system resonances to be similar to, but not exactly the same as those generated by the 1-D model. This small discrepancy in frequency may be considered negligible as the variable nature of speech and specifically formant frequencies from person to person allows for some variation in this matter. More importantly, it is considered that the perceived likeness of the resulting sound output of the 2-D mesh to the vowels which were modeled remains a good match.

Although it produces accurate formant synthesis, the *widthwise mapping* approach to area function application does not fully accommodate dynamic changes in tract shape. The vocal tract configuration used to generate the /i/ to /a/ vowel slide in Fig. 7 involves a widening in mesh width at the mouth, and a narrowing towards the middle. Referring to the mesh layout around these areas in Figs. 5(c) and 6(c) it is clear that this transition will require additional waveguides to be added around the mouth, and removed from the middle region. These changes force surrounding junctions to alter their behavior. This dynamic restructuring of waveguides at run-time can be problematic in maintaining the continuity laws governing the mesh scattering equations. For example, an increase in width might see a one-connection boundary junction changing to a four-port scattering junction, resulting in the averaging of the single incoming pressure

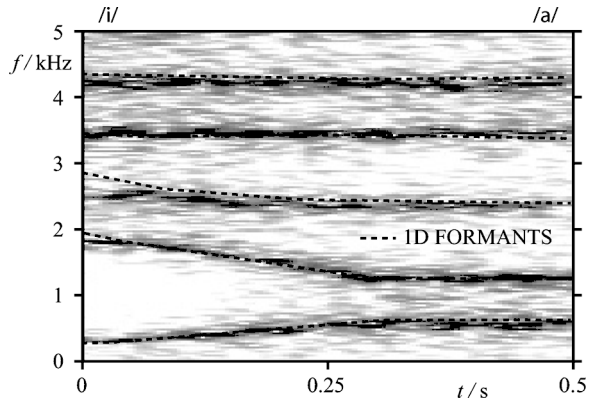

Fig. 7. Two-dimensional mesh width mapped /i/ to /a/ vowel slide spectrum.

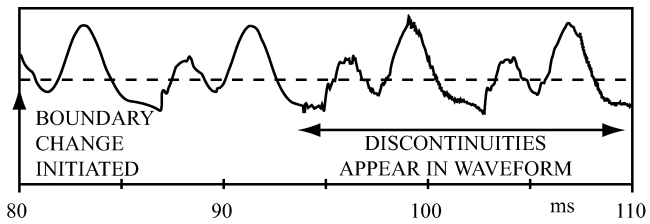

Fig. 8. Waveform discontinuities introduced by changing mesh boundary properties.

value across four outputs and hence a sharp step in pressure gradient in the new mesh area. Some attempts were made to develop a junction with the capability to accommodate such changes. Scattering (3) was reconfigured such that errors introduced by additional or removed pressure inputs were spread evenly across existing connections as to minimize their effects. It was found, however, that the manipulation of the equations was often in contradiction with the underlying acoustic theory, and hence introduced more instabilities rather than less. Minimum disruption to the pressure balance at each junction was achieved simply by defining new pressure components to be set to zero and that lost pressure components are disregarded.

The distance involved in moving a boundary between minimal and maximal area function values is about 20 mm. The changes are infrequent, as the number of junction manipulations is negligible when compared to the number of samples in the given duration for the transition. For example, using (6) the waveguide size in a high resolution mesh sampled at 120 kHz is about 4 mm. In the approximately 500 ms required for the transition in a diphthong, this would result in five junction changes at each moving boundary point over 60 000 samples. However, the small discontinuities propagate across the mesh and are still audible in the output waveform. Fig. 8 shows the output generated for 30 ms after a boundary change is initiated in the 2-D model during a vowel slide. The LF glottal flow derivative model was used as excitation [26]. Approximately four cycles of the output waveform are shown after a boundary alteration instant. Discontinuities are clearly visible about 14 ms after a step in boundary movement, which are audible as a high frequency click in the output.

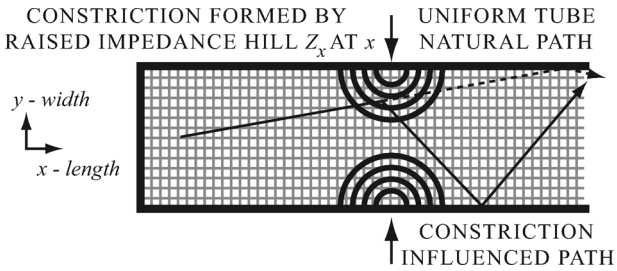

Fig. 9. Raised impedance hills causing constriction in DWM tube.

B. Impedance Mapped Dynamic Articulations

In order to achieve full dynamic tube-area manipulation without the introduction of waveform discontinuities a technique has been developed that uses a different method to map the area function onto the 2-D mesh model. Considering that the use of impedance to represent area in the 1-D model allows for stable alteration of tract shape, a 2-D vocal tract mesh model using increased dimensional impedance representation has been used to simulate realistic sounding diphthongs. In Section IV-A a constriction to the flow in the tract was implemented with a reduction in mesh width. This new method introduces raised impedance areas, or hills on to a rectangular straight-tube mesh.

According to (7), a narrowing in the 1-D tract implies a decrease in cross-sectional area, equivalent to an increase in impedance of the single waveguide at that point. It follows that a band of increased impedance waveguides across the width of a rectangular 2-D mesh at a point of constriction will introduce similar forwards-backwards reflection-transmission as the 1-D KL junctions mentioned in Section III. However, if the band is of constant impedance across, then no cross-tract reflections will occur, thus making the use of multidimensionality superfluous.

A minimum impedance channel is defined through the middle of the mesh to act as a direct propagation path, and raised impedance sectors are introduced to either side to act as the resistance to propagation in areas of constriction. A wavefront approaching the constriction experiences some transmission through the center channel and reflection due to the higher impedances, both across and along the mesh. Fig. 9 demonstrates how the application of raised impedance hills towards the edges of the mesh alters its resonant behavior.

We define $Z_{\min}$ as the lowest impedance value (or largest area $A_{\max}$) across the range of vowels required and construct a rectangular mesh of length $l = 17.6$ cm and width corresponding to that maximal opening $w = 2\sqrt{A_{\max}/\pi}$. With equal impedance across the rectangular mesh, the model exhibits the same lengthwise resonant behavior as a uniform tube of cross sectional area $A_{\max}$. An impedance map is constructed whereby each impedance value Z_x along the length of the 1-D area function is translated into multiple values across the width of the 2-D mesh. The minimum value $Z_{\min}$ remains at the center of the tube and any constrictions greater than this appear as impedance hills either side, approaching Z_x at the tract inner walls. Two different functions have been applied to the shape of the impedance hills to demonstrate this scheme. First, a linear increase from $Z_{\min}$ in

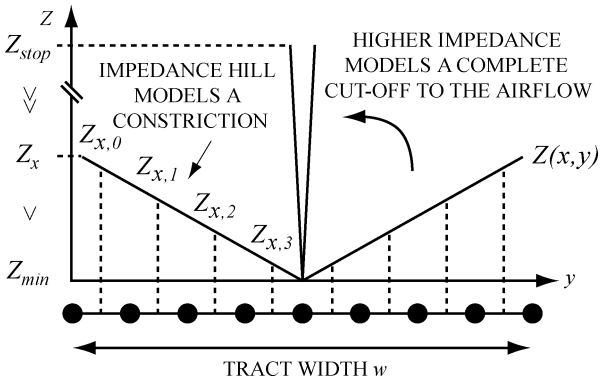

Fig. 10. Constriction modeled by a linear increase of waveguide impedance values from the center to the edges of the mesh at Z_x.

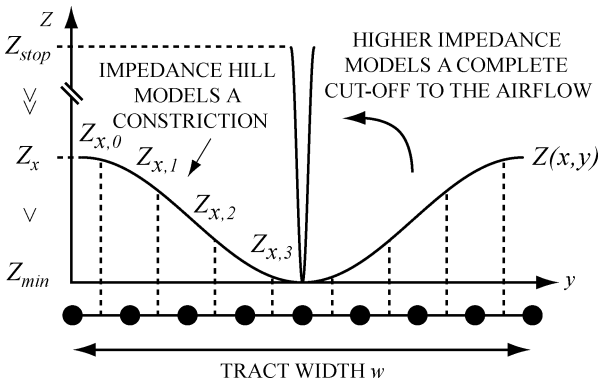

Fig. 11. Constriction modeled by a raised-cosine increase of waveguide impedance values from the center to the edges of the mesh at Z_x.

the center, to Z_x at the mesh edge, is illustrated in Fig. 10. Junctions shown are those across the width of the mesh at a point x along its length. The grey dotted lines indicate the waveguide impedance value used for each point across the width of the mesh. An example of the manner in which a plosive articulation is modeled using this method is included in the diagram. The high impedance $Z_{\text{stop}} \gg Z_{\text{min}}$ represents a virtual cut-off to the flow, where the linear impedance hills are large enough to restrict flow beyond the constriction.

The second demonstration of this method uses an inverted raised cosine window, scaled according to the equivalent 1-D impedance value Z_x and central channel minimum Z_{min}, to realize the impedance hills. The impedance variation across the y-axis of the 2-D mesh of total width w is described in (8). The equation is evaluated at each point along the x-axis (length) in relation to the corresponding impedance value Z_x taken from the 1-D area function

$$Z(x,y) = Z_x - \frac{(Z_x - Z_{\text{min}})}{2}\left[1 + cos\left(2\pi\left(\frac{y}{w} - \frac{1}{2}\right)\right)\right]. \tag{8}$$

Fig. 11 shows the impedance shape formed across the width of the rectangular mesh using the raised cosine function (8).

With this system applied to the mesh, the *impedance map* for the /i/ and /a/ vowels would appear as in Figs. 12 and 13, respectively. In the diagrams the 1-D area function is detailed in (a) and the straight tube mesh with overlayed raised cosine impedance

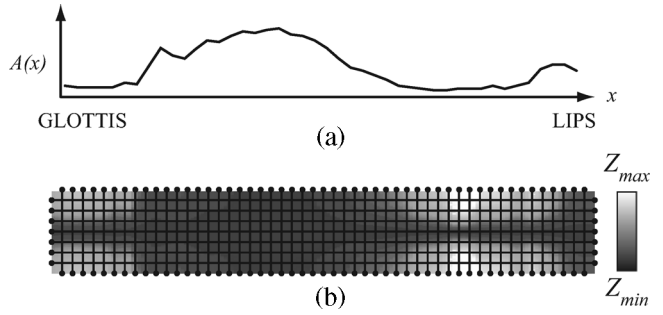

Fig. 12. Forming the impedance mapped /i/ vowel waveguide mesh. (a) Cross-sectional area function and (b) rectangular mesh with raised cosine impedance map.

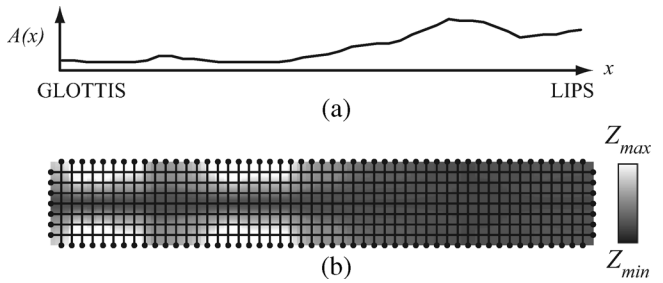

Fig. 13. Forming the impedance mapped /a/ vowel waveguide mesh. (a) Cross-sectional area function and (b) rectangular mesh with raised cosine impedance map

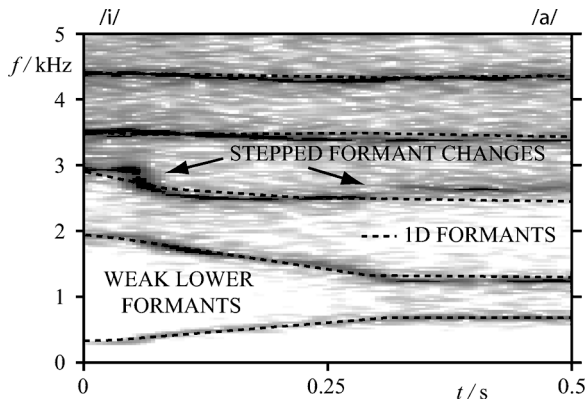

Fig. 14. Two-dimensional mesh linear impedance mapped /i/ to /a/ vowel diphthong.

map is shown in (b), where a higher impedance is represented with a lighter shade of grey. The minimum impedance channel can be observed as the darker area along the center, and the lighter, higher impedance constrictions are clear towards the edges of the mesh. The equivalent linear impedance maps are not included as little visible difference is apparent between the two functions at the scale chosen for the diagram.

The two methods can be verified with an examination of the formant patterns generated when each is used to generate the vowel slide /i/ to /a/. A linear interpolation is used to define the transition between the two area functions in both cases. The dynamic behavior of the linear impedance mapped model is shown in the noise excited spectrogram in Fig. 14, to be very similar to the widthwise mapped equivalent in Fig. 7.

The overlayed dotted line highlights that the changes in formants are also very close to those generated using a 1-D model

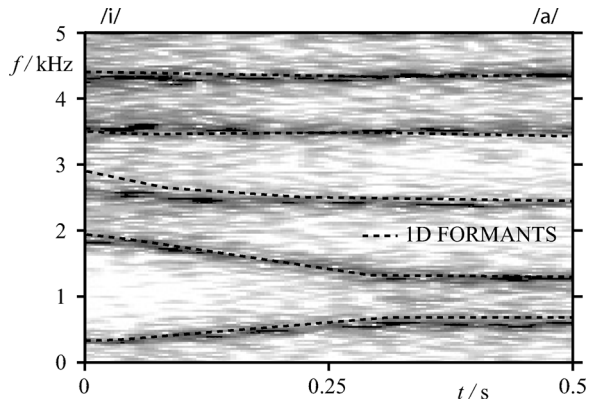

Fig. 15. Two-dimensional mesh raised cosine impedance mapped /i/ to /a/ vowel diphthong.

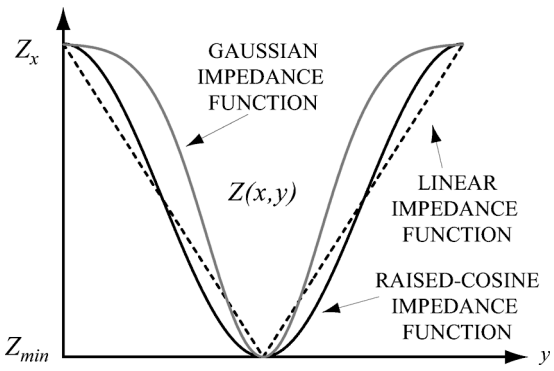

Fig. 16. Comparison of possible cross-mesh impedance functions: linear, gaussian and raised cosine.

in Fig. 4. The lower two formants, however, are attenuated by approximately 12 dB compared to the third and fourth. Furthermore, the third formant can be seen to follow a slightly erratic path, giving stepped changes rather than the smooth transition required. Speech-like output generated using this model contains a buzzing quality where the glottal waveform is emerging with little of the important lower resonances imparted onto it. This is considered an artifact of the very narrow channel created with the linear impedance hills. The central $Z_{\min}$ lower impedance region has zero physical width (see Fig. 10), and so has little of the desired effect as a direct propagational path.

The raised cosine impedance mapped model vowel slide formants are shown in Fig. 15. Again, the frequencies appear at similar values to those generated with both the 1-D and 2-D widthwise models. Issues identified with the linear mapping function are no longer present. The relative strength and smoothness of transition of the formants is more in line with those observed in the 1-D model.

The wider central channel provided by the raised cosine method (Fig. 11) provides an increased acoustic throughput, and therefore does not restrict the signal propagation as with the linear map. Fig. 16 highlights this difference. Comparison is drawn between the raised cosine impedance function, the linear version (dotted line), and another possible consideration, a gaussian curve (grey line) with a standard normal distribution. The raised cosine function offers a wider central channel whilst still providing a smooth transition in the increased impedance effects towards the tract inner walls. Furthermore, vowel sounds

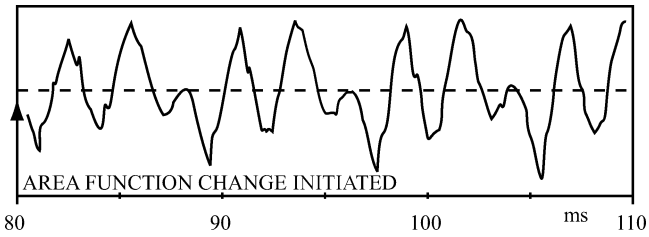

Fig. 17. Impedance mapped model waveform after area function change.

generated with the raised cosine function are considered to be the most natural of the three simulations. As such, it is selected for use in further simulations.

There are now no discontinuities audible in the resulting diphthong created using the application of the LF glottal waveform to the model during the slide between the two area functions. This is demonstrated in Fig. 17 where the output waveform after an area function change contains none of the high frequency discontinuities highlighted in the *widthwise* equivalent in Fig. 8.

C. Real Time Performance

Previous work has shown that the 2-D mesh vocal tract model based on a *widthwise* area function implementation offers speech sound generation on a non real-time basis [11], [12]. Parameters are set within the model and the system is left to generate the output. As indicated in (6), sampling frequency and hence simulation times vary largely with mesh resolution; clearly more waveguide sections, closer together within the spatially sampled area will result in a longer run time. Owing to its construction, the *widthwise* 2-D vocal model requires a high resolution mesh to incorporate the small distances appearing in some states of the tract. For example, the distance between the lips in the /u/ vowel area function [28] can be as little as 8 mm. For minimum mesh resolution two waveguides (two boundary junctions and one standard scattering junction) are needed to suitably model this narrow tube section. Given (6), a 2-D system synthesizing wave propagation at $c = 343$ ms^{-1} using a waveguide size $d = 4$ mm results in a sampling frequency of $f_s = 120$ kHz. With such a mesh requiring of the order of 200–300 junctions, real-time performance is not currently an option. Questions also arise when considering the sort of mesh resolution needed to accurately model a near or complete stop to the air flow, for example as seen in plosive articulation.

The *impedance mapping* techniques place no limitations on minimum width, as the mesh retains its rectangular shape throughout. Therefore a 2-D system may be constructed with a waveguide size $d = 11$ mm, which gives a sampling frequency of $f_s = 44.1$ kHz, and given the $f_s/4$ restraints for the rectilinear mesh, a valid bandwidth of approximately 11 kHz. Such an arrangement employing a rectilinear mesh topology comprises 60 waveguide junctions. Exploiting this reduction in sampling frequency, software has been developed which demonstrates real-time 2-D DWM vocal tract model vowel shape manipulation. The model allows for real-time user interaction using a mouse to bring about smooth vowel slides resulting in diphthongs and sharper area function changes effecting momentary constrictions to the airflow for simulation of plosive articulation.

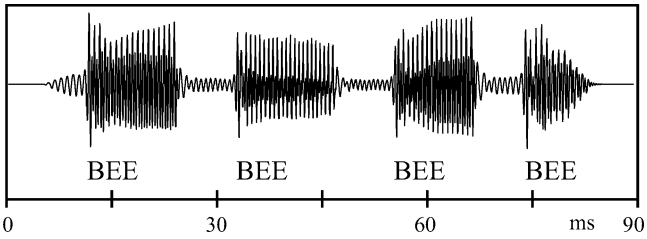

Fig. 18. Simulated plosive articulation with the 2-D impedance mapped mesh.

The software constructed to test the real-time dynamic 2-D model is available for download and use at http://www-users.york.ac.uk/~dtm3/vocaltract. Fig. 18 shows example output from the software. In the simulation, voiced plosive articulation is modeled in the 2-D impedance mapped /i/ vowel mesh using a mouse controlled slider which represents the area function at the lip end. The plosive is generated with a constriction to the tract as indicated by the high impedance Z_{stop} in Fig. 11. The waveform has four distinctive points where a stop and release of pressure is generated as a direct result of the real-time constrictions made with the user interface. Combined with LF glottal excitation, each plosive part of the waveform is a synthesized version of the word *bee*. Other vowels and area function manipulations at different places along the tract may be used to simulate further voiced and non-voiced plosive articulation.

D. Discussion

Articulatory vocal tract modeling is currently still in its infancy. A sophisticated artificial speech system based on a simulation of the complex biological and physical processes observed in the vocal tract may eventually exhibit naturalness close to that already shown in sample based methods. Once such a system is defined it also may prove interesting, as is possible in other physical models, to examine its behavior beyond that observed in the real world. Adapting the input parameters and bounding limits could give scope for experimentation, such as the changing of material properties within the tract, or even the application of non-human tract area data.

Vocal tract modeling methods based on defining wave propagation in 1-D are widely known and accepted due to their simplicity and low processing requirements, making them ideal for real-time simulation. The addition of extra dimensionality to the model will increase its ability to accurately simulate the tract resonances. However, current levels of advancement in computer processing power indicate that high resolution multidimensional acoustical modeling in real-time is not yet possible. Furthermore, the techniques used to model such structures, specifically those used in 3-D DWM room acoustic simulation, are currently under development and as yet little has been done to include dynamic considerations.

A technique presented here as an alternative method of area function application to the 2-D DWM vocal tract model has been shown to allow stable dynamic changes to be made. Results from Section IV-B show the impedance based tract shape application to generate nearly identical formant patterns to those created using the same area function as applied to a highly spatially sampled, and therefore accurate 1-D model. The process of translating the area function data onto a mesh impedance map was analyzed with two approaches; one using a linear impedance increase and one based on a raised-cosine function. The steep gradient of the linear function proved to restrict the width of the central channel which was inherent in the model design. It was therefore concluded that the raised-cosine-based constrictions provide more accurate formant simulation, and more realistic sounding vowel synthesis. This configuration is shown to facilitate a slide between two vowels without the waveform discontinuities introduced using width based shape mapping discussed in Section IV-A. The alternative manner in which constrictions are applied to the tract using the impedance based method also allows for a reduction in sampling frequency and, as discussed in Section IV-C, realtime interaction with the tract model. It is considered that output results bear an audible improvement to those generated with a highly spatially sampled 1-D model using the same area functions and glottal input.

E. Future Considerations

Further research is needed into the use of multidimensional signal processing techniques in dynamic time-domain acoustic modeling of the vocal tract. In order to verify it as a potentially advantageous method of artificial speech production focus may be needed towards the completion of the 2-D model using boundary filters to model the tract energy losses due to frictional, thermal and yielding wall effects. The use of different topology waveguide structures, such as the triangular [18] or interpolated [19] mesh, should also be investigated in order to exploit the increased valid frequency output range and improved dispersion characteristics. A move towards a 3-D model using a complete 3-D scan of the tract shape may bring about highly accurate vocal tract acoustic simulations. The use of full MRI scan data with complex-shape cross-sectional area data rather than the circular form used for 1-D simulation should increase the naturalness of synthesis. It is considered that the impedance mapping of area function presented in this paper could be easily adapted for use in a 3-D model. A 17.6-cm-long cuboid rectangular waveguide structure could have the area function set within the impedance of waveguides through each 2-D slice across the tract. Such a system would also lend itself well to the non-circular tract shapes found in the complex 3-D scans. Considerations could also be made towards the inclusion of lengthwise shape changes to the mesh, such as those observed in the protrusion of the lips for the /u/ vowel and weather this can be achieved with some form of the impedance mapping technique presented in this paper. Much further work would be involved in drawing up a set of rules as to how the model might move during simulation of actual speech, such that it might be used in natural sounding speech synthesis. Furthermore, alongside technological advances in tract scanning, these methods could eventually be used to create a personalized model for the speech impaired.

V. Conclusion

Recent developments in 2-D waveguide mesh modeling of the vocal tract have been discussed. Problems with discontinuities appearing in the output waveform introduced by moving the mesh boundaries have been addressed. Simulation of the tract changes taking place in speech have been considered using

a new method of area function application which defines waveguide impedance as the variable quantity rather than the mesh shape itself. This technique allows for stable manipulation of the area function for the synthesis of diphthong and plosive speech sounds. In addition, the possibility of a reduction in system sampling frequency also emerges, allowing for real-time performance to be achieved.

REFERENCES

[1] P. A. Taylor, A. Black, and R. Caley, "The architecture of the festivalspeech synthesis system," in *Proc. 3rd ESCA Workshop in Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 147–151.

[2] T. Dutoit, "High-quality text-to-speech synthesis: an overview," *J. Elect. Electron. Eng. Australia: Special Issue on Speech Recognition and Synthesis*, vol. 17, no. 1, pp. 25–37, 1998.

[3] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 7, pp. 955–967, 1987.

[4] O. Engwall, "Tongue Talking—Studies in Intraoral Speech Synthesis," Ph.D. dissertation, Royal Inst. Technol., Stockholm, Sweden, 2002.

[5] P. Birkholtz and D. Jackel, "A three-dimensional model of the vocal tract for speech synthesis," in *Proc. 15th Int. Congr. Phonetic Sciences (ICPhS)*, 2003, pp. 2597–2600.

[6] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. 4th Int. Congr. Acoustics*, Copenhagen, Denmark, 1962, pp. 1–4.

[7] J. Liljencrants, "Speech Synthesis With a Reflection-Type Line Analogue," Ph.D. dissertation, Royal Inst. Technol., Stockholm, Sweden, 1985.

[8] P. R. Cook, "Identification of Control Parameters in an Articulatory Vocal Tract Model With Applications to the Synthesis of Singing," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1991.

[9] V. Välimäki and M. Karjalainen, "Improving the kelly-lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Yokohama, Japan, 1994, pp. 615–618.

[10] D. P. Berners, "Acoustics and Signal Processing Techniques for Physical Modeling of Brass Instruments," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1999.

[11] J. Mullen, D. M. Howard, and D. T. Murphy, "Digital waveguide mesh modelling of the vocal tract acoustics," in *Proc IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2003, pp. 119–122.

[12] ——, "Waveguide physical modeling of vocal tract acoustics: Flexible formant bandwidth control from increased model dimensionality," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 964–971, May 2006.

[13] J. O. Smith, "Physical modeling using digital waveguides," *Comput. Music J.*, vol. 16, no. 4, pp. 74–91, 1992.

[14] S. A. Van Duyne and J. O. Smith, "Physical modeling with the 2D digital waveguide mesh," in *Proc. Int. Comput. Music Conf.*, Tokyo, Japan, 1993, pp. 40–47.

[15] M. Karjalainen and C. Erkut, "Digital waveguides versus finite difference structures: equivalence and mixed modelling," in *EURASIP J. Appl. Signal Process.*, New Paltz, NY, 2004, vol. 2004, no. 7, pp. 978–989.

[16] 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-04)*, 2004, pp. 268–273.

[17] A. Kelloniemi, "Improved adjustable boundary condition for the 2-d digital waveguide mesh," in *Proc. 8th Int. Conf. Digital Audio Effects (DAFx-05)*, Madrid, Spain, 2005, pp. 237–242.

[18] F. Fontana and D. Rocchesso, "Signal-theoretic characterization of waveguide mesh geometries for models of two dimensional wave propagation in elastic media," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 152–161, Mar. 2001.

[19] L. Savioja and V. Välimäki, "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. 184–193, Mar. 2000.

[20] 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.

[21] L. Savioja and V. Välimäki, "Interpolated rectangular 3D digital waveguide mesh algorithms with frequency warping," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 783–790, 2003.

[22] B. H. Story, "Physiologically-Based Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract," Ph.D. dissertation, Univ. Iowa, Ames, 1995.

[23] J. O. Smith, Physical Audio Signal Processing: Digital Waveguide Modeling of Musical Instruments and Audio Effects Stanford Univ., 2004, Online Book [Online]. Available: http://ccrma.stanford.edu/jos/pasp

[24] V. Välimäki, "Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters," Ph.D. dissertation, Lab. Acoust. Audio Signal Process., Faculty of Elect. Eng., Helsinki Univ.Technol., Helsinki, Finland, 1995.

[25] B. H. Story and I. R. Titze, "Voice simulation with a bodycover model of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1249–1260, 1995.

[26] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," in *Quarterly Progress Report*. Stockholm, Sweden: Speech Transmission Lab., Royal Inst. Technol., 1986.

[27] H. W. Strube, "Are conical segments useful for vocal-tract simulation," *J. Acoust. Soc. Amer.*, vol. 114, no. 6, pp. 3028–3031, 2003.

[28] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 100, no. 1, pp. 537–554, 1996.

Jack Mullen received the M.S. degree in electronic engineering with music technology from the University of York, York, U.K., in 2002. The final year M.Eng. project was research into digital waveguide mesh boundary implementation for room acoustics modeling. He is currently pursuing the Ph.D. degree from York University in multidimensional waveguide vocal-tract modeling.

His research interests include speech and acoustical modelling.

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

He became a Lecturer in speech and hearing sciences at UCL in 1979 and he moved to York in 1990. He gained a 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, and a Fellow of the Institute of Acoustics and a Member of the Audio Engineering Society.

Damian T. 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 Audio Consultant since 2002 and 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. 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.

Dr. Murphy is a member of the Audio Engineering Society.