Waveguide Physical Modeling of Vocal Tract Acoustics: Flexible Formant Bandwidth Control From Increased Model Dimensionality

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

***Abstract*—Digital waveguide physical modeling is often used as an efficient representation of acoustical resonators such as the human vocal tract. Building on the basic one-dimensional (1-D) Kelly–Lochbaum tract model, various speech synthesis techniques demonstrate improvements to the wave scattering mechanisms in order to better approximate wave propagation in the complex vocal system. Some of these techniques are discussed in this paper, with particular reference to an alternative approach in the form of a two-dimensional waveguide mesh model. Emphasis is placed on its ability to produce vowel spectra similar to that which would be present in natural speech, and how it improves upon the 1-D model. Tract area function is accommodated as model width, rather than translated into acoustic impedance, and as such offers extra control as an additional bounding limit to the model. Results show that the two-dimensional (2-D) model introduces approximately linear control over formant bandwidths leading to attainable realistic values across a range of vowels. Similarly, the 2-D model allows for application of theoretical reflection values within the tract, which when applied to the 1-D model result in small formant bandwidths, and, hence, unnatural sounding synthesized vowels.**

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

I. Introduction

SPEECH synthesis plays an important role in modern communications. Well-established techniques, such as linear predictive coding [1] and formant-based methods [2] are widely used in the analysis/synthesis of transmitted or generated speech and produce speech-like sound of a reasonably organic nature; an important factor of the quality of such a scheme. In such methods, the frequency patterns present in natural speech are reconstructed using bandpass filters to introduce the desired resonances into a model of the human vocal tract. While both accurate and efficient, these methods do not allow for controlling parameters that are directly related to physical properties of the speech system. Such intuitive semantic control over many aspects of the produced sound is inherent in a physics-based model.

The field of virtual acoustics modeling draws many parallels with both speech synthesis and musical instrument modeling. Much speech synthesis theory is derived from the notion that the vocal tract is an acoustic resonator [3], and similarly, the modeling of sound propagation within vibrating structures corresponds to both room acoustics and instrument bodies. Geometrical acoustic techniques such as ray-tracing [4] are typically used to represent the many propagation paths a sound wave follows between source and receiver in an enclosed space. This information can then be used to construct an approximation to a room impulse response (RIR), often combining the energy-time response with head-related transfer function (HRTF)-based directivity information for auralization purposes. The digital waveguide mesh (DWM) physical model provides an alternative method of modeling sound propagation in an enclosed space, exhibiting accurate low frequency characteristics [5] and diffraction effects [6] not inherent in ray-tracing.

Similarly, physical modeling as applied to speech synthesis offers an alternative approach over those based on spectral reconstruction. Physical models of the vocal tract [7]–[9], have shown that sounds approaching natural speech can be synthesized using a chain of one-dimensional (1-D) waveguides representing the length of the tract from the glottis to the lips. A two-dimensional (2-D) model extending the 1-D case by incorporating variable tract width along the length of the model has been introduced [10], and the software used to analyze the model has been discussed [11]. The model is examined here in greater detail, with particular reference to the control available over formant bandwidths of synthesized vowels.

The contribution of this paper is a novel method for time-domain simulation of the acoustics of the human vocal tract. Increased dimensionality has been presented as an alternative improvement to the basic Kelly-Lochbaum (KL) 1-D waveguide vocal tract model over methods simulating enhanced order area function approximation. In particular, the model has been used to show how vowels can be synthesized using the additional reflecting boundary parameter as an effective control over formant bandwidth. The model indicates that there should be clear advantages in moving toward a full three-dimensional (3-D) model, giving extensive control over many physical parameters of the human vocal system. This paper is arranged as follows. Section II introduces the concept of physical modeling with reference to both musical instrument and room acoustics modeling. Digital waveguide theory is then introduced for both a 1-D and 2-D physical model. Section III details articulatory vocal tract modeling in terms of the development of the 1-D waveguide model and the concept and construction of the proposed 2-D waveguide vocal tract model. Vowel simulation results in the form of formant frequency and bandwidth values are presented in Section IV and discussed in Section IV-D.

Manuscript received July 26, 2004; revised April 25, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

The authors are with the Media Engineering 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/TSA.2005.858052

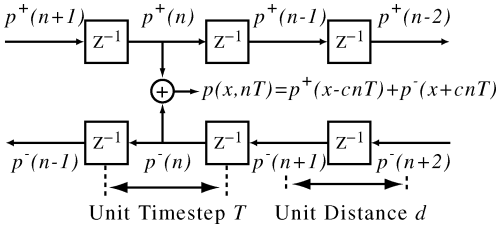

Fig. 1. 1-D chain of waveguides.

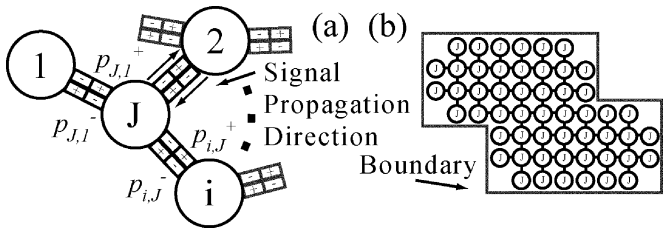

Fig. 2. (a) Unit junction and (b) rectilinear mesh.

II. Physical Modeling Synthesis

The application of physics-based modeling in the artificial reproduction of natural sound allows for a physical process to be directly represented, and, hence, accurately synthesized with output results bearing semantic relationship to input parameters. Physical modeling employs a simplified discretization of real world mechanics in order to sufficiently represent the target physical structure such as a musical instrument or analog circuit, or process such as the collision of two objects. For example, a vibrating body can be represented as a lumped element system, sampled at a suitable resolution, where each element obeys the physics-derived laws governing interaction with its neighbors. With constraints applied to the signal-propagating medium and inputs defined, the virtual model exhibits natural behavior that approximates real world expectations. This method can be thought of as the definition of two aspects; a virtual *resonator* and *exciter*. For example a 1-D chain of interconnected elements with fixed terminations representing a guitar string, coupled with a 3-D resonating cavity representing a guitar body, might comprise a virtual resonator. An input might then be defined as an initial or continuous external variation of a number of system parameters, resulting in excitation to the system. This method has proved beneficial in the quest for realistic sound synthesis of a continuously interactive nature over previously implemented nonphysical methods due to its real-world representation [12]. In addition, the principles behind physical modeling allow for low memory requirements to be placed upon a system as large sample lookup tables are not needed. However, representations of larger structures, such as concert halls, require large numbers of interconnected units, resulting in a significant increase in processing power, and usually nonreal time performance.

A. One-Dimensional Digital Waveguide

The digital waveguide physical model defines the unit element within a 1-D system to be a bidirectional 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 (LTI) nature.

Based on a discrete version of the D'Alembert solution to the 1-D 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 said to be the sum of a left-going (p^-) and a right-going (p^+) components at each time step as in (1), where c is the wave speed, typically 343 ms^{-1} in the context of an acoustic pressure wave

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

Application of an input to the string system and then continuous iteration of scattering and timestep equations to each element constitutes propagation of a travelling 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]. Scattering junctions are formed where multiple waveguides meet, such that in its most basic form—the rectilinear mesh—waveguides form a cartesian-coordinate grid joining junctions placed at 90° from each of its four neighbors. Fig. 2(a) and (b) details 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 can be shown to be

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

The application of the following three equations to each node in the mesh gives rise to accurate lossless scattering of pressure.

- The pressure p at a lossless junction J with N equal impedance waveguide connections is derived from (3) to be

$$p_J = \frac{2}{N}\sum_{i=1}^{N} p^+_{J,i}. \tag{4}$$

- The pressure output $p^-_{J,i}$ on each waveguide connected to a junction is directly related to its input

$$p^-_{J,i} = p_J - p^+_{J,i}. \tag{5}$$

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

$$p^+_{J,i} = p^-_{i,J}(n-1). \tag{6}$$

Mesh boundaries are simulated using scattering equations derived from impedance matching techniques, allowing for a proportional amount of incident energy to be reflected back into the mesh, as defined by the reflection coefficient r, such that the pressure on a single connection boundary node, as in node 1 in Fig. 2(a), is

$$p_1 = (1+r)p^+_{1,J}. \tag{7}$$

Equations (4)–(6) can also be derived as an equivalent finite difference scattering algorithm

$$p_J(n) = \frac{2}{N}\sum_{i=1}^{N} p_i(n-1) - p_J(n-2). \tag{8}$$

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.

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

$$f_s = \frac{1}{T_s} = \frac{c\sqrt{N}}{d}. \tag{9}$$

The ability of the rectilinear mesh to perform uniform scattering, however, deteriorates as a function of direction and frequency due to dispersion error. This results in a reduction in propagation wavespeed in axial-directions for higher frequency components. Alternative methods of mesh construction have resulted in the development of triangular [15] and bilinearly deinterpolated [16] topologies, both of which reduce the problem to within acceptable levels. Frequency-dependant dispersion error can be compensated for by the inclusion of frequency warping, where pre- and postprocessing of the mesh adjusts for unwanted frequency shifts in the spectrum [16].

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 [17]–[19]. Accurate simulation of a source within the space can be simulated with either direct injection or convolution with the RIR measured from the mesh.

III. Vocal Tract Modeling

Each vowel, typically identified by its first three resonances or *formant* frequencies, is formed by creating constrictions in the vocal tract using the tongue, lips, and jaw muscles. The basic form of the tract is often modeled as a 17.5-cm straight tube of uniform cross-sectional area, open at one end, giving resonances at approximately 500, 1500, and 2500 Hz; a reasonable match to the formants of the neutral vowel—/e/. Further vowel sounds are created as the formant frequencies change through the application of constrictions to the tube, quantified by an *area function*. Such measurements are taken as cross-sectional area data from X-rays of the human vocal tract during utterances of Russian vowels [3], or from magnetic resonance imaging [20] and then applied to the model such that it behaves as an acoustic resonator of the same shape. With an applied area function, the tract model takes the form of a connected series of tube elements with varying cross-sectional area and, hence, varying impedance function. Bounding limits to the tract amount to a reflecting closed end at the glottis and a negative reflection at the open lip end. It has been indicated [21], [22] that a small amount of energy is radiated out through the fleshy inner walls of the tract, that the glottis reflects almost all energy incident upon it—about 97%, and that approximately 10% of energy in the tract is radiated out through the mouth, although this varies with frequency and opening area. Application of a periodic glottal input and frequency dependent lip reflection completes the model. With wave propagation assumed to be planar and linear, time and/or frequency domain solutions to the acoustic wave equation are used to model the pressure variations in the human vocal tract, producing speech like sound. Dynamic changes made within the parameters of the model represent the movement of tract features, such as the lips and tongue. Such articulatory modeling provides direct synthesis of the speech production mechanisms and exploits physically related control parameters that vary slowly enough to provide low bit rate speech data.

In a frequency domain acoustic tube model, physical parameters are mapped onto matrices representing the transfer function of each section. Tract resonances follow a reasonably linear nature and can, therefore, be suitably modeled using a number of 2×2 (ABCD) chain matrices representing the characteristic impedance of each tube section [23]. Each matrix includes parameters relating to the tube area, wall vibrational behavior, and viscous losses. The cascaded matrices then combine to form the complete transfer function of the vocal tract. A time-domain glottal model can be introduced, as either a periodic signal such as the Liljencrants–Fant waveform [24], or a mass and damper system using two [25], three [26], or up to eight [27] masses to model each vocal fold.

A. One-Dimensional Waveguide Vocal Tract Model

The traditional time-domain method of simulating the acoustical properties of the human vocal tract as a series of different impedance cylindrical tubes uses the KL scattering junction [7] at each intersection between segments. The impedance discontinuity between two sections of tube arises from the change in cross-sectional area A between two points, and acts to cause scattering of pressure signals through both the transmission and reflection of some amount of the propagating wave signal across the section discontinuity. The pressure p and velocity v components in each section i can be directly related to the corresponding impedance Z_i and cross-sectional area A_i using air density ρ and the wavespeed c

$$Z_i = \frac{p}{v} = \frac{\rho c}{A_i}. \tag{10}$$

The KL scattering equation is a special case of the more general N-connection waveguide junction (3) where $N = 2$. It describes, in discrete form, the signal flow out of a junction as

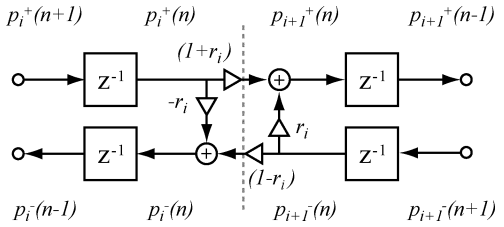

Fig. 3. Kelly-Lochbaum scattering junction.

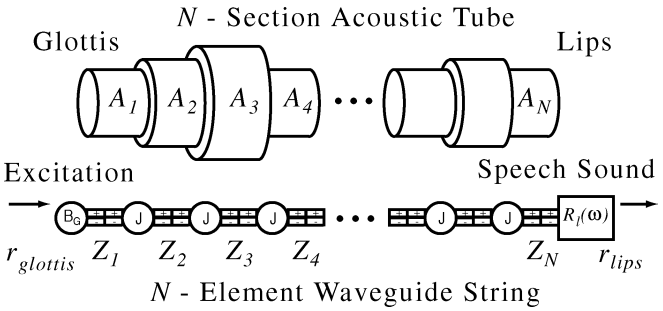

Fig. 4. One-Dimensional waveguide vocal tract model.

illustrated in Fig. 3 in terms of pressure input to section $i+1$ as right-going $p_{i+1}^{+}(n)$ and section i as left going $p_i^{-}(n)$

$$p_{i+1}^{+}(n) = (1+r_i)p_i^{+}(n) + r_i p_{i+1}^{-}(n) \tag{11}$$

$$p_i^{-}(n) = -r_i p_i^{+}(n) + (1-r_i)p_{i+1}^{-}(n) \tag{12}$$

where the reflection coefficient between sections i and $i+1$ can be related to the impedances $Z_{i,i+1}$ and cross-sectional area $A_{i,i+1}$

$$r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} = \frac{Z_i - Z_{i+1}}{Z_i + Z_{i+1}}. \tag{13}$$

Fig. 4 illustrates the construction of the 1-D model and how the area function is set according to the waveguide impedances. Boundary reflection coefficient values are set such that a positive reflection exists at the closed glottis end ($0 < r_{\text{glottis}} < 1$), and a negative reflection is present at the open lips ($0 > r_{\text{lips}} > -1$). A frequency-dependant lip reflection $R_l(\omega)$ can also be applied as a one-pole lowpass digital filter [8], where r_{lips} is the gain value in the pass band.

Improvements on the basic KL 1-D model have been suggested. Increased accuracy in the model can be achieved with substitution of the low-resolution spatially sampled cylindrical tube sections for conical elements that better follow the changes in tract area function [9]. This amounts to a first-order approximation of the area function, whereas the cylindrical waveguide method can be considered zeroth order. Spherical pressure wave propagation is used to define the 1-D wave equation in a conical tube as follows:

$$\frac{\partial^2 p(r,t)}{\partial r^2} + \frac{2}{r}\frac{\partial p(r,t)}{\partial r} = \frac{1}{c^2}\frac{\partial^2 p(r,t)}{\partial t^2} \tag{14}$$

where c is the wave speed, r is the distance from the tip of the cone, and $p(r,t)$ is acoustic pressure. Partial differential (14) has a discretized solution such that the total pressure $p(n)$ in a waveguide representing a conical tube element, where r is the distance to the effective cone-tip, is

$$p(r, nT) = \frac{1}{r}\left[p^{+}(r - cnT) + p^{-}(r + cnT)\right]. \tag{15}$$

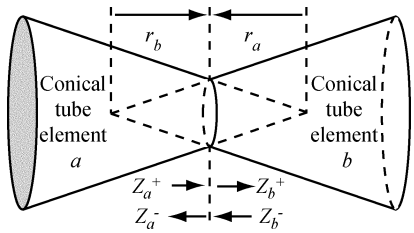

Fig. 5. Junction between two conical tubes.

The acoustic impedance Z of conical tubes varies as a function of frequency. This results in the introduction of a reflection filter $R(\omega)$ into the scattering junction at the impedance discontinuity present between the two conical tube sections a and b of equal connecting area, but different taper. This is illustrated in Fig. 5, where Z_a^+ and Z_a^- are the impedances out of and into cone a, respectively (and similarly for cone b), which act to cause frequency-dependant scattering at the junction of the form

$$R^{+}(\omega) = \frac{\frac{1}{Z_a^{+}(\omega)} - \frac{1}{Z_b^{+}(\omega)}}{\frac{1}{Z_a^{-}(\omega)} + \frac{1}{Z_b^{-}(\omega)}}. \tag{16}$$

The closer area function approximation offered by conical tube segmentation increases the accuracy of the model in terms of agreement with target formant frequencies. However, when compared to the cylindrical equivalent with double spatial resolution, the conical model presents no further improvements in performance at a similar computational complexity [28]. As such, the conical waveguide model introduces similar complexity and benefit as a cylindrical model with half the waveguide size.

The conical waveguide junction introduces stability issues. At the impedance discontinuity between two conical tube sections, scattering derived from spherical wave propagation and simplified to plane wave propagation does not include the gradual increase in wavefront curvature. A spherical wavefront passing between two conical sections experiencing an increase in cone taper (widening of tube) naturally displays a bulging of the wavefront to accommodate the $1/r$ relationship (effective distance from the tip of the cone) in the new section. In the waveguide model, this change in wavefront shape is sudden rather than gradual, and equivalent to a missing section of tube volume. Similarly, propagation through decreasing taper junctions experience a sudden reduction in wavefront curvature, amounting to a doubly defined volume [29]. Providing these erroneous volume sections are small compared to tract volume changes, then reflection filters can be used to simulate stable wave propagation in the conical junction.

An alternative to conical tube modeling exists in the form of a discretization of Webster's horn equation [29]. Developed as a model of the bell of a brass instrument, it can also be used to describe propagation in the human tract. The effect of tract inner

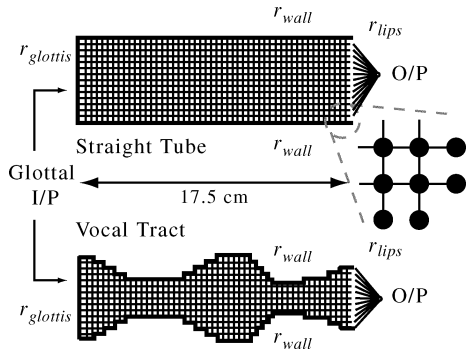

Fig. 6. Straight tube and vocal tract 2-D DWM models.

wall curvature can be better approximated with the application of a wave motion in a flared horn-type geometry. By modeling a gradual change in propagating wavefront radius rather than the abrupt change offered by the conical element method, dispersion is introduced into the scattering between sections leading to wave reflectance of greater accuracy.

B. Two-Dimensional Waveguide Vocal Tract Model

The 2-D model implements the width of the tract in the same manner as length is included in the 1-D KL model. Removing the plane-wave motion assumptions gives propagation across as well as along the tract allowing for simulation of higher order modes. The area function data [3] is translated into width data, assuming a circular cross-sectional area, and then used to determine the number of waveguides across each length segment. The constructed mesh is then analogous to a 2-D plan of the air cavity through the tract, from the glottis to the lips in the mid-sagittal plane.

Fig. 6 illustrates the mesh constructed to represent a straight tube and an arbitrary vowel shape with the inclusion of the width data. The nature of the 2-D model increases the controlling parameters available when compared to the 1-D case. The introduction of the extra boundary into the system along the inside length of the tract wall, labeled as r_{wall} in Fig. 6, allows for flexible control of formant bandwidths as will be examined in Section IV. Glottal excitation can be introduced into the system by either a direct injection of a periodic signal such as the LF model, or using a constant flow controlled by the changing impedances of the waveguides, modulated by a glottal opening area-function. Both input methods can be applied along the width of the glottal end of the mesh, taking advantage of the second dimension by using the curvature of the vocal cords to focus the injected pressure toward the middle of the tract. Similar advantages may also be gained from the increased dimensionality of the model in terms of the application of articulatory features. The effects of speech modifiers such as the lips, teeth, and tongue may also be included in the cross-tract plane to accurately model their influence and again offering improved semantic control. The intention is that similar control benefits to those offered in using the chain matrix approach to articulatory speech synthesis can be exploited in a wave scattering-based method.

Fig. 7 illustrates the graphical output used in simulations for the /u/ vowel 2-D waveguide-mesh model. The length of the

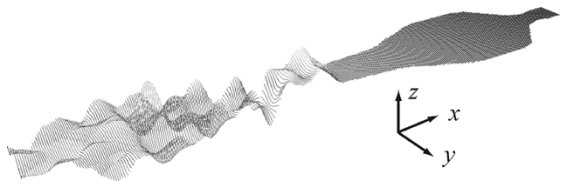

Fig. 7. 2-D /u/ vowel model 0.25 ms after applied Gaussian impulse at the glottal end.

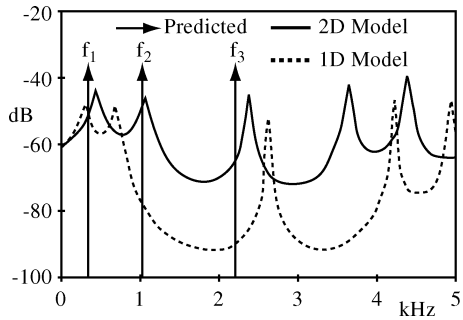

Fig. 8. Noise excited 1-D-2-D model spectra: /u/ vowel.

tract is modeled by the waveguides along the x-axis, the diameter is modeled by waveguides along the y-axis, and pressure signal magnitudes are represented on the z-axis. The 2-D arrangement of junctions results in signal scattering in both the x and y axes, and, hence, the propagation of higher order resonant modes not inherent in the 1-D model.

One current limitation of the 2-D model is the restriction on sampling frequency. The constrictions within the vocal tract for particular shapes can result in diameters as small as 8 mm, in for example, the distance between the lips during production of the vowel /u/. For adequate mesh resolution a minimum of two waveguides (one scattering and two boundary junctions) are required across the narrowest sections to ensure boundary junctions are not connected together. A choice of waveguide size of 4 mm results in a system sampled at $f_s = 121.3$ kHz, constituting a total number of between 200 and 300 junctions, depending on vowel shape. This sample rate upper-bound currently suggests that real time system performance may not yet be achieved. Clearly, a speech synthesis system based on a 2-D model will introduce additional complexities and computational requirements. At its current stage in development, it is intended simply as a research tool and not as a successor to existing real-time speech methods.

IV. Results

A. Vowel Synthesis

Measurements, in the form of noise excited spectral responses, taken from the 2-D mesh model of the vocal tract show its potential to simulate accurate formant peaks [10], in terms of frequency and bandwidth, both important factors in the realistic simulation of vowels.

Fig. 8 shows an analysis of the 2-D waveguide mesh model in the /u/ vowel configuration, with preliminary reflection values of $r_{\text{lips}} = -0.9$, $r_{\text{glottis}} = 0.97$, and $r_{\text{wall}} = 0.97$. Constructed in this way, a 2-D model with a waveguide size of 4 mm gives 220 nodes. Also included in Fig. 8 are formant patterns generated

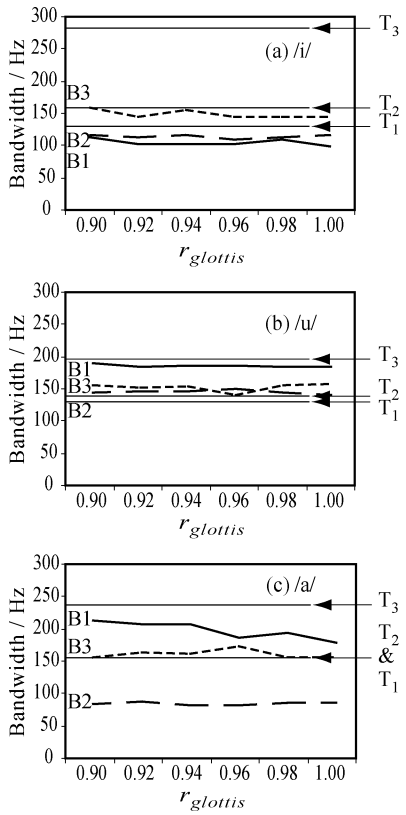

Fig. 9. Two-dimensional waveguide vocal tract model—formant bandwidths varying with r_{glottis} for the (a) /i/, (b) /u/, and (c) /a/ vowels.

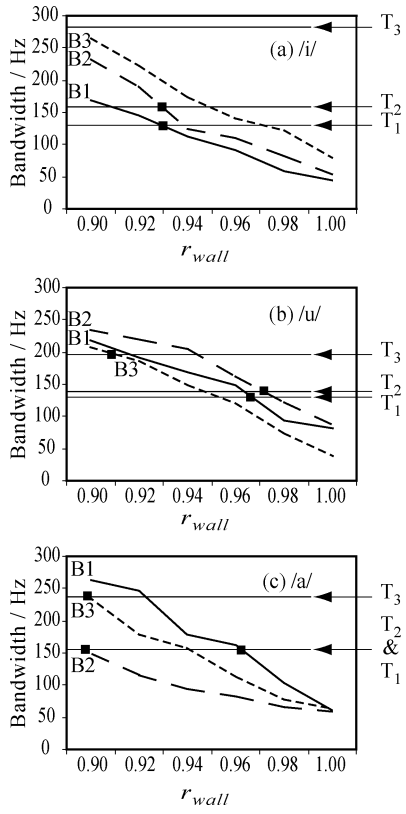

Fig. 10. Two-dimensional waveguide vocal tract model—formant bandwidths varying with r_{wall} for the (a) /i/, (b) /u/, and (c) /a/ vowels.

using a 1-D KL vocal tract model for comparison. Results presented in [28] indicate that a higher resolution spatially sampled 1-D model gives increased accuracy in area function approximation. As such, a high resolution 1-D model was used with a waveguide size of 0.5 mm (350 nodes), in order to give sufficient area function accuracy and, hence, reasonable comparison. The KL model is the basis for many vocal tract physical models, and due to its simplicity and ease of implementation, acts as a good benchmark for comparison purposes. The 2-D model can be seen to produce formants which follow a better match to given predicted frequency values [30] than those produced by the KL model. In simulations of different vowels, the 2-D model performs better in most, equal in some, and slightly worse in others, but the small differences between the frequency values of the 1-D and 2-D peaks are minor discrepancies considered negligible owing to the large variability of speech. Therefore, it is considered that a reasonably high-resolution 2-D model is considered as accurate as a highly spatially sampled 1-D model.

Vowel spectra shapes are often described in terms of the first three formant frequencies f_1, f_2, and f_3, and their respective bandwidths B_1, B_2, and B_3. Values taken from measured speech predict the first three bandwidth values to average around $B_1 = 140$, $B_2 = 150$, and $B_3 = 210$ Hz [30], although this differs from vowel to vowel. Spectra in Fig. 8 give 2-D bandwidth values at around 80–140 Hz. This is a reasonable match to target values, although optimization of formant bandwidths using the increased control offered by the extra boundary should demonstrate improvements in the 2-D model for accurately simulating vowel sounds.

B. Two-Dimensional Model Formant Bandwidths

Fig. 9 shows the variation in formant bandwidth attainable using r_{glottis} as a control parameter, with r_{wall} set to a high reflection value of 0.97 and $r_{\mathrm{lips}} = -0.9$. This is similar to the manner in which bandwidth adjustment is achieved in the simple KL 1-D vocal tract model with tract wall losses modeled as a small attenuation in each junction. Target bandwidths for the Nth formant for each of the three vowels in Fig. 9(a) "*bead*," (b) "*booed*," and (c) "*bard*" are indicated by the label T_N. Clearly, none of the target bandwidths are achieved with the 2-D model using r_{glottis} as a control parameter. It can also be seen that very little variation is present in the bandwidths when the dominating reflecting boundary of r_{wall} is kept constant and r_{glottis} is varied. As a parallel to the manner in which reflection coefficients, and, hence, formant bandwidths, are altered in the simple KL model, this result highlights how ineffective such boundary adjustment is when attempting to control tract losses in such models.

The effect of using r_{wall}, the additional parameter introduced with the use of the 2-D model, as a controlling parameter for bandwidth can be seen in Fig. 10. The formant bandwidths generated by the 2-D model have been examined through a range of r_{wall} with the remaining boundaries fixed at theoretical values, $r_{\mathrm{lips}} = -0.9$ and $r_{\mathrm{glottis}} = 0.97$. The bandwidth values are more in line with expectations as indicated by the presence of two [Fig. 10(a)], three [Fig. 10(b)], and three [Fig. 10(c)] success-intersection points. The sole missing success-intersection point in the 2-D simulations exists for B_3 in the /i/ vowel simulation [Fig. 10(a)], although the measured error from a successful intersection was marginal at less than 50 Hz for $r_{\mathrm{wall}} =$

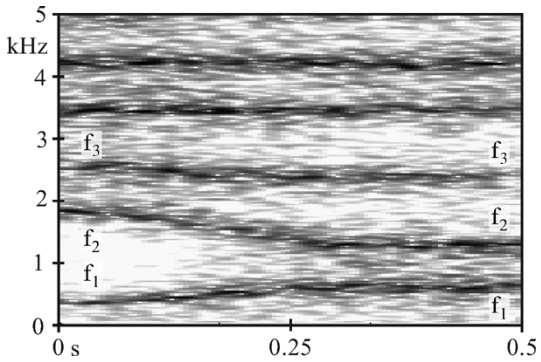

Fig. 11. /i/ to /a/ vowel diphthong.

0.9. Note that all three bandwidth values for each formant are equally responsive to changes in r_{wall}, presenting an approximately linear response. As such, r_{wall} can be selected to give a best-fit match to desired bandwidth values for each vowel.

The simulated bandwidths do, however, remain interrelated and currently cannot be individually tuned. Further research remains in the field of separating control of each of the three formant bandwidths, allowing fully optimized bandwidths for each formant of each vowel. This may be achieved by identifying which tract sections have a dominating influence on individual bandwidths and allowing for variable r_{wall} values along the length of the tract. Alternatively, the introduction of frequency dependent boundaries may give extended control over such issues.

C. Two-Dimensional Dynamic Ability

The ability of the model to reproduce the dynamic changes in the tract during speech further validates it as a potential synthesis tool. A linear interpolation between tract area functions with an applied glottal input generates a dynamic slide between two vowels. The spectrogram in Fig. 11 illustrates the changing formants as the 2-D mesh model boundaries move to vary the constrictions in the tract. The continuity laws applied to the waveguide mesh do not allow for dynamic changes to be made to its structure. However, the changes in boundary configuration are small as changing boundaries have pressure values set to zero, and are infrequent with respect to the sample frequency. As such, only a small discontinuity is introduced into the model, apparent as a small audible high-frequency click in the output.

With a change in the application of the area function using loaded junctions or fractional delay filters, it is proposed that such dynamic changes, and, hence, articulatory capabilities are attainable without the undesirable discontinuities.

D. Discussion

Technological advances of the 1-D tract model have amounted to an increase in accuracy of area function approximation [9], [28], [29]. Success in this area has either been in terms of increased spatial sampling resolution, or a more thorough treatment of the wave propagation mechanism itself. Both methods result in accurate simulation of formants frequencies and, hence, a realistic vowel synthesis.

Formant bandwidths also contribute to the naturalness of vowel sounds. The ability to adjust synthesized bandwidth values toward those observed in natural speech will also increase the power of a vocal tract model. In the 1-D model, energy reflected back into the tract is largely governed by coefficients r_{lips} and r_{glottis}. Results from the 2-D model shown in Section IV-B give little variation in bandwidth when r_{glottis} is used as a controlling parameter with wall losses set, equivalent to bandwidth variation in the traditional 1-D KL model. In contrast, when the additional reflection coefficient r_{wall} is used as a controlling parameter bandwidth values follow an approximately linear pattern of adjustment, and, hence, optimum values can be achieved. The theoretically predicted values set in r_{lips} and r_{glottis} are valid and can, therefore, remain fixed. From Fig. 10 it can be concluded that for a reasonable match to target formant bandwidths a value of $r_{\mathrm{wall}} = 0.92$ should be used as an appropriate minimum-error point between success points in the 2-D graph simulations. These three values conform to logical expectations in the human tract. The majority of losses should exist at the lips, where sound is actually radiated, with some vibrational and heat conduction losses present in the fleshy inner walls of the tract, and a high reflection at the glottis, which is closed for a majority of its cyclic motion.

The small variations in bandwidth between each vowel should then be accommodated by the model itself in terms of the number of boundary junctions across the lip opening, and their effective combined internal reflection and, hence, their contribution toward the resultant bandwidth. Vowels with greater-area lip opening such as /i/ and /a/ will have a greater number of r_{lips} junctions, giving rise to more sound output from the model. Conversely, smaller lip-openings as in /u/ will contain fewer boundary junctions, and so reflections back into the tract will be greater, and bandwidths will be smaller, as required by the lower target values in Fig. 10(b) when compared to those in Fig. 10(a) and (c).

The resulting output generated using the 2-D model are considered a good likeness to the respective target vowel sounds. The manner in which the constrictions are applied to the tract model is an important factor in the quality of resulting vowel sounds and as such future work will include optimization of the techniques used to apply the area function to the 2-D tract model. Research into fully optimized independent bandwidth control may introduce a variable r_{wall} along the tract length or frequency dependent reflections. It is also considered that with further development of the manner in which the area function is applied to the model, it will be possible to reduce the high resolution required to model small tube sections.

V. Conclusion

A 2-D waveguide mesh model of the human vocal tract has been presented as an alternative development to the standard Kelly–Lochbaum 1-D vocal tract model. Results from simulations of the tract using area functions based on specific vowels show accurate synthesis of formant frequencies. Formant bandwidths measured from simulations highlight the increased control available from the 2-D model. The nature of its construction introduces an extra mesh boundary, incorporating the losses along the length of the tract. This distribution of loss more closely follows real-life vocal tract acoustics. As such, it allows for application of reflection coefficients to the lip

and glottis ends of the 2-D model that are close to those recommended by biological analysis, while maintaining desired formant bandwidth values.

The ability of the 2-D model to create synthesized sounds of greater naturalness is yet to be fully established, but current speech sounds generated give clear audible improvement upon those generated by the basic 1-D model. Future work involves the inclusion of improved dynamic constriction changes, further mouth features and a full 3-D system. It is believed a powerful comprehensive-representation physical model of the vocal tract can be used to create synthesized speech with a high level of naturalness and allowing intuitive semantic control over model functionality.

REFERENCES

[1] J. Makhoul, *Linear Predictive Coding in Electronic Speech Synthesis*. Chicago, IL: R. R. Donnelley, 1984.

[2] D. H. Klatt, "Software for a cascade/parallel formant synthesiser," *J. Acoust. Soc. Amer.*, vol. 67, no. 3, pp. 971–995, 1980.

[3] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[4] A. Krokstad, S. Strøm, and S. Srøsdal, "Calculating the acoustical room response by use of a ray tracing technique," *J. Sound Vibr.*, vol. 8, no. 1, pp. 118–125, 1968.

[5] L. Savioja, J. Backman, A. Järvinen, and T. Takala, "Waveguide mesh method for low-frequency simulation of room acoustics," in *Proc. 15th Int. Congr. Acoustics (ICA)*, vol. 2, Trondheim, Norway, 1995, pp. 637–640.

[6] D. T. Murphy and M. J. Beeson, "Modeling spatial sound occlusion and diffraction effects with the digital waveguide mesh," in *Proc. AES Int. Conf.*, 2003, pp. 207–216.

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

[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. Spoken Language Processing (ICSLP)*, Yokohama, Japan, 1994, pp. 615–618.

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

[11] ——, "Acoustical simulations of the human vocal tract using the 1-D and 2-D digital waveguide software model," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFx-04)*, Naples, Italy, 2004, pp. 311–314.

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

[13] ——, "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 2-D digital waveguide mesh," in *Proc. Int. Computer Music Conf.*, Tokyo, Japan, 1993, pp. 40–47.

[15] 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.

[16] 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.

[17] S. A. Van Duyne and J. O. Smith, "The 3-D tetrahedral digital waveguide mesh with musical applications," in *Proc. Int. Computer Music Conf.*, Hong Kong, China, 1996, pp. 9–16.

[18] L. Savioja and V. Välimäki, "Interpolated rectangular 3-D digital waveguide mesh algorithms with frequency warping," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 783–790, Nov. 2003.

[19] 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.

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

[21] J. Liljencrants, "Speech synthesis with a reflection-type line analogue," Ph.D. dissertation, Royal Inst. Technol., Stockholm, Sweden, 1985.

[22] M. Kob, "Physical modeling of the singing voice," Ph.D. dissertation, Aachen University of Technology, Aachen, Germany, 2002.

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

[24] "Quarterly Progress Report," Speech Transmission Laboratory, Royal Inst. Technology, Stockholm, Sweden, 1986.

[25] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal folds," *Bell Syst. Tech. J.* , vol. 51, pp. 1233–1268, 1972.

[26] 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.

[27] I. R. Titze, "The human vocal cords: a mathematical model, part I," *Phonetica*, vol. 28, pp. 129–170, 1973.

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

[29] D. P. Berners, "Acoustics and signal processing techniques for physical modeling of brass instruments," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1999.

[30] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. New York: Wiley, 2000.

Jack Mullen received the M.Eng. degree in electronic engineering with music technology from the York University, 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 working towards the Ph.D. degree at York University in multidimensional waveguide vocal-tract modeling with an expected completion date of 2006.

His research interests include acoustical and music instrument modeling.

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

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, a Fellow of the Institute of Acoustics, and a member of the Audio Engineering Society.

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

In 1999, he was a 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 Audio Consultant since 2002 and has been a Visiting Lecturer in the Department of Speech, Music, and Hearing, at 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.