Characteristics of Two-Dimensional Finite Difference Techniques for Vocal Tract Analysis and Voice Synthesis

Matt Speed, Damian Murphy, David M Howard

Audio Lab, Department of Electronics, University of York, Heslington
mdas100@ohm.york.ac.uk, dtm3@ohm.york.ac.uk, dh@ohm.york.ac.uk

Abstract

Both digital waveguide and finite difference techniques are numerical methods that have been demonstrated as appropriate for acoustic modelling applications. Whilst the application of the digital waveguide mesh to vocal tract modelling has been the subject of previous work, the application of comparable finite difference techniques is as yet untested. This study explores the characteristics of such a finite-difference approach to two-dimensional vocal tract modelling. Initial results suggest that finite difference techniques alone are not ideal, due to the limitation of non-dynamic behaviour and poor representation of admittance discontinuities in the approximation of three-dimensional geometries. They do however introduce robust boundary formulations, and have a valid and useful application in modelling non-vital static volumes, particularly the nasal tract.

Index Terms: vocal tract, finite difference time domain, computation, acoustic, simulation

1. Introduction

The Kelly-Lochbaum (KL) model [1] was amongst the earliest vocal tract models, and remains a popular approach to vowel synthesis. At its most fundamental level the model consists of two parallel and opposite-travelling delay lines with commuted losses and delays. This implementation is essentially equivalent to the digital waveguide, despite predating its formal development by many years [2]. The digital waveguide presents a method for modelling sound propagation in a medium by using a discrete form of d’Alembert’s travelling wave solution to the one-dimensional wave equation. This can similarly be implemented as a bi-directional delay line with commuted losses and delays. Waveguides can be connected together in various topologies and dimensionalities, resulting in the Digital Waveguide Mesh (DWM) [3]. This provides a description of wave propagation over a multi-dimensional spatio-temporal sampling grid, not dissimilar to those generated by existing Transmission Line Matrix (TLM) techniques[4]. The DWM has become increasingly popular for acoustic modelling, although it is not the only means of providing a discretised form of the wave equation. Finite-Difference Time-Domain (FDTD) techniques employ difference operators to directly discretise the wave equation. These can be arranged in the same multi-dimensional arrangements as their DWM counterparts. Since FDTD techniques propagate a physical variable (rather than the wave variables of DWM techniques) they are often referred to as ‘Kirchoff’ methods. The equivalence of DWM and FDTD techniques has been well investigated [5]. Each has its own advantages and disadvantages, and the two can be used together for best effect [6]. Mullen implemented a real-time voice synthesiser by taking advantage of dynamic impedance mapping in the DWM [7]. While FDTD can not currently be operated dynamically, it offers a significant computational saving over an equivalent DWM model [8] (by assuming the propagative medium to be of uniform impedance). With this computational saving in mind it is valuable to consider the contribution FDTD modelling can make within the scope of a hybrid-paradigm voice model.

2. The Finite-Difference mesh

As with the DWM, FDTD techniques provide a discretised representation of the wave equation. This is derived from the combination of equations describing the conservation of mass and conservation of momentum [9]. Regardless of the dimensionality of the system it describes, the wave equation remains a second order partial differential equation. The one-dimensional wave equation is given in (1), with (2) being its two-dimensional counterpart.

\[
\frac{\partial^2 p(x, t)}{\partial x^2} = c^2 \frac{\partial^2 p(x, t)}{\partial t^2} - \frac{\partial p(x, t)}{\partial x} \tag{1}
\]

\[
\frac{\partial^2 p(x, y, t)}{\partial x^2} = c^2 \left( \frac{\partial^2 p(x, y, t)}{\partial x^2} + \frac{\partial^2 p(x, y, t)}{\partial y^2} \right) \tag{2}
\]

Either form of the equation can be easily discretised using forward and backward difference operators in turn. Of crucial relevance at this stage is the relationship between the temporal and spatial discretisation intervals (as described by (3)), as these will determine the wave speed of pressure propagation. This factor is determined by the Courant number, \(\lambda\), and the dimensionality of the system.

\[
\lambda = \frac{cT}{\Delta x} \tag{3}
\]

To minimise numerical dispersion and satisfy the Courant stability condition [10] the Courant number is set as \(\lambda = 1\) for a one-dimensional mesh and \(\lambda = \frac{1}{\sqrt{2}}\) for a two-dimensional mesh. The resulting 2D discrete update equation will take the form of (4), where \(x\) and \(y\) are coordinates of a two-dimensional grid, and \(t\) is the time-sample index.

\[
p_{x,y}^{t+1} = \lambda^2 (p_{x+1,y}^{t} + p_{x-1,y}^{t} + p_{x,y+1}^{t} + p_{x,y-1}^{t}) + 2(1 - 2\lambda^2)p_{x,y}^{t} - p_{x,y}^{t-1} \tag{4}
\]

2.1. Boundary Modelling

Recent work in FDTD-based acoustic modelling has reconsidered the problem of implementing frequency dependent boundaries based on the assumption of a locally reacting wall [9]. Pre-
vious techniques, paralleling established DWM approaches using 1-D connectivity at a boundary, introduced ambiguities regarding the correct choice of Courant number. One-dimensional connections are constrained to $\lambda = \frac{1}{4}$ by geometrical correctness, whereas $\lambda \leq \frac{1}{4}$ meets the Courant stability condition for the 2-D case. The new formulation creates a two-dimensional grid at the boundaries of a 2-D FDTD mesh, removing the ambiguity regarding the Courant number. Both frequency independent, and dependant boundaries have been developed by Kowalczyk and van Walstijn in [9], and the frequency independant formulation is used here. At the boundary node, neighbouring nodes within the boundary medium (‘ghost’ points) are removed from the update equation by substituting the discrete form of (5), describing the boundary condition where $\xi_b$ is the normalised wall impedance.

$$\frac{\partial p(x,t)}{\partial t} = -\xi_b \frac{\partial p(x,t)}{\partial x} \tag{5}$$

Equation (6) shows the discretised form of (5).

$$p_x^{t+1} = p_x^{t-1} - \frac{1}{\lambda \xi_b} (p_x^{t+1} - p_x^{t-1}) \tag{6}$$

3. Implementation

The relationship between spatial and temporal sampling intervals is dictated by the Courant number as defined in (3). The distance between spatial sampling points is in turn defined by the sampling rate assumed in the system, according to:

$$f_s = \frac{c}{d} \sqrt{N} \tag{7}$$

where $N$ is the number of dimensions in the system being considered, $c$ is the speed of sound and $d$ is the spatial sampling interval.

Increasing the sampling rate of the system shortens the spatial sampling interval, increasing the density of the sampling grid, and hence improving the accuracy with which it is possible to closely match the bounding geometry of the model. Meshing anything other than planar and perpendicular surfaces which are in length a perfect multiple of the spatial sampling interval will result in a boundary geometry that is only an approximation to the continuous boundary of the modelled system.

The vocal tract models used in this study were developed from Magnetic Resonance Imaging (MRI) of four vocal tract shapes (/a/, /E/, /i/, /u/) [11]. This MRI data describes the cross-section of the vocal tract at a number of mid-sagittal intervals (28 for /a/, 25 for /E/, 29 for /i/ and 31 for /u/). Having calculated the cross-sectional area at each of these intervals, models are produced to represent these changing cross-sectional areas (as described by Stevens[12]). Input was in the form of a bandlimited impulse, injected at the closed end of the cylinders (to simulate the closed-phase of the glottal cycle) and off the centre line of the tract model to encourage adequate excitation of all cross-tract modes. Output was taken at the opposite end of the cylinders, to simulate system output at the lips (although lip radiation is not taken into account). Simulations were run for 0.2s, which is adequate time for the establishment of resonant modes regardless of the propagational path pursued.

4. Validating the 2D FDTD Method

Testing was carried out to ensure the accuracy and validity of the output in terms of the correct prediction of resonant mode positions, by comparing simulated output to mode positions calculated for structures with well defined (and mathematically stated) acoustic behaviour. This validation is crucial if the approach is to be used for identification or synthesis of vocal tract formant patterns. Two scenarios are considered, a basic cylinder, and the concatenation of two cylinders with different cross-sectional areas (as described by Stevens[12]). Input was in the form of a bandlimited impulse, injected at the closed end of the cylinders to simulate the closed-phase of the glottal cycle and off the centre line of the tract model to encourage adequate excitation of all cross-tract modes. Output was taken at the opposite end of the cylinders, to simulate system output at the lips (although lip radiation is not taken into account). Simulations were run for 0.2s, which is adequate time for the establishment of resonant modes regardless of the propagational path pursued.

4.1. Basic Cylinder

Since we are testing a series of 2D representations of 3D cylinders, the natural first test is the simulation of a single cylinder. The cylinder tested was of length 0.17m (assumed average length of male vocal tract), and of width 0.03m. A mesh was generated with the sample rate corresponding to a spatial sampling interval of $d = 0.26 \times 10^{-4}$m. The resulting rectangular mesh was of size 71 by 48 nodes. Reflection coefficients of $r = 1$ were applied across all cylinder boundaries, to represent an entirely closed and reflective condition. The resonant behaviour of such a cylinder is well understood, and early resonant modes of this plane can be predicted using the universal modal equation (8):

$$f_{x,y,z} = \frac{c}{2} \sqrt{\left(\frac{N_x}{l}\right)^2 + \left(\frac{N_y}{w}\right)^2} \tag{8}$$

where $l, w$ are cylinder length and width respectively, $N_x, N_y$ are harmonic numbers in each dimension and $c$ is the speed of sound. Resonant mode positions were calculated up to 15kHz using (8). The resulting frequency response is shown in Figure 1, using a 65536 point FFT.

Figure 1: Frequency response of 0.17m closed cylinder described in section 4.1.
All resonant modes were present and correct up to approximately 15kHz. Under simulation there was a 0.195% error in model length, and a 0.173% error in model width. This error in model length will introduce a shift of 1.0Hz to simulated lengthwise modes (N_e), and a shift of approximately 10Hz to widthwise modes (N_a). This, combined with errors due to potential off-centre peaks within a frequency bin in the FFT explains the very small errors in resonant mode placement.

4.2. Stevens’ Cylinders

After successful testing of the cylindrical response, the next step is a two-stage concatenated model. Though not as straightforward in calculation as the simple cylinder, two concatenated cylinders provide a useful analogue with the vocal tract through perturbation theory. In [12] Stevens mathematically describes the effect of pharyngeal narrowing and oral widening. It is in effect a transition from a neutral open tract condition to vowel /a/. Considering a single cylinder of half the vocal tract length (0.085m) open at one end and closed at the other, it will generate resonant modes at f_0 = 1kHz and n f_0 where n is odd. Concatenating a second cylinder with a discontinuity in area function causes the resonant modes to split in a manner determined by perturbation theory and the relationship between their cross-sectional areas, as defined by:

\[ f'_0 \approx f_0 \left( 1 \pm \frac{2}{\pi} \sqrt{\frac{A_1}{A_2}} \right) \]  

(9)

In the limit, if the two cylinders are of identical cross-sectional area the resonances would separate so far as to match the resonances of a single cylinder of length 0.17m. This can be tested by producing a narrow ‘pharyngeal’ tube (0.009m x 0.085m - 16 x 135 nodes) and connecting it to a wider ‘oral’ tube (0.03m x 0.085m - 48 x 135 nodes). The resulting resonant mode positions should match the resonances calculated from (9) for an equivalent cylinder of length 0.085m. The resulting obtained response is shown in Figure 2, with the expected resonant mode positions plotted as vertical lines. The errors in resonant mode position highlight a fundamental shortcoming in 2D vocal tract modelling. Firstly it should be acknowledged that every 2D mesh is in essence a ‘membrane’. For planar structures, as in Section 4.1 the FDTD mesh can completely describe the resonant behaviour. It cannot however be assumed equivalent to a cylinder, since the resonant behaviour in only a single plane of the structure is represented. With this in mind the relationship between two adjacent cylinders exhibiting a discontinuity in cross-sectional area should be reconsidered. Stevens [12] shows that the modal behaviour caused by concatenation is determined by the difference in cross-sectional area, and therefore the acoustic admittance apparent from each cylinder to its neighbour. The 2D membrane simulation provides only a ‘top-down’ approximation to the cross-sectional area, essentially replacing the area ratio of (9) with the ratio of cylinder diameters. The resonant mode transformation itself is therefore only an approximation to that of the 3D system. Where there are more cylinders involved (as is the case with vocal tract models used here) the modal behaviour will become both unpredictable and not necessarily representative of the vocal tract itself. It should be noted that this error does not occur in 1D, nor 3D FDTD modelling. In 1D models admittance is determined explicitly as a system parameter to facilitate scattering. The equivalent plot for a 1D model is included in Figure 2, this demonstrates the expected mode positions. In 3D models the admittance discontinuity is modelled implicitly since the model geometry is completely described. In the 2D model it is possible to adjust the diameters of each cylinder to provide an approximation to the correct resonant mode separation. Whilst this might provide an accurate formant pattern, this approach will compromise cross-tract modes, thereby voiding the advantage 2D modelling holds over 1D modelling.

5. Discussion

Figure 4 shows the frequency response of the outputs obtained from each of the four vocal tract vowel models (shown in Figure 3), under excitation conditions described in Section 3. The frequency response is based on a 65536 point FFT of the impulse response obtained off centre-axis at the ‘mouth’ of the model. Figure 5 shows the /a/ model during simulation.

![Figure 2: Frequency response of Stevens concatenated cylinders - Simulated with one- and two-dimensional techniques.](image)

![Figure 3: 2D vowel models - From top: a) /I/ b) /u/ c) /d/ d) /a/.](image)

Note the relatively large resonant mode amplitudes at frequencies above 3kHz in Figure 4, which would be unusual if these were spoken vowels. It is helpful to remember that this is a simulation of the resonant behaviour of the vocal tract alone. The effect of combination with the -6dB/Octave decay of the source is not considered, nor are the effects of lip radiation. The similarity between the four frequency response plots is striking, particularly below 4kHz. The models display very different geometrical arrangements (as demonstrated in Figure 3), yet these differences have very little impact upon early formant positions. This is discouraging, since the first 4 formants are crucial to vowel identification, suggesting these 4 vowels would sound very similar indeed. Despite their identical lengths, the differences in tract model profiles should lead to greater changes in formant position (particularly for f_2, f_3 and
as observed in Figure 4). That the higher frequency resonant mode positions (above 4kHz) become more varied between models, suggests that longitudinal resonant mode placement is compromised. This is likely to be a symptom of the poor admittance discontinuity modelling afforded by the 2D mesh, as explored in Section 4.2.

Figure 5: Impulse propagation during simulation of /ad model.

6. Conclusions

Whilst its computational efficiency and ease of implementation makes the 2D FDTD mesh an attractive proposition, there are two main factors which constrain its application in vocal tract modelling. Firstly the structure of the FDTD mesh cannot be adjusted during simulation without compromising the stability of the model. This rules out its implementation in any region of a vocal tract model intended to move, for example the oral cavity (when tongue movement is modelled) or at the lips. Secondly, it has been shown that in two-dimensional models scattering due to admittance discontinuities is not representative of an equivalent three-dimensional geometry. While this may be avoided by adjusting the section diameters, this will move the error into the cross-tract modes, compromising the physical analogue and complicating parameterised control of the tract model. These factors demonstrate that without combination with wave techniques, the usefulness of 2D FDTD techniques is limited. Where resonant behaviour is of lesser influence the 2D FDTD mesh could perhaps be used without similar concern, a possible application being an integrated model of the nasal tract. The 2D FDTD boundary formulation is especially robust, having eliminated ambiguity regarding the Courant number. For this reason the frequency dependent variant particularly is likely to be of benefit to a hybrid Kirchoff/Wave model.

Further work should approach a formal treatment of the resonant equivalence of the 2D FDTD mesh and 3D geometries, and furthered hybridisation of Kirchoff and Wave based techniques.

7. Acknowledgments

The MRI data used in this study was kindly provided by Olov Engwall, at TMH, KTH, Stockholm. Author MS is funded by a University of York EPSRC Doctoral Training Grant.

8. References