ACOUSTICAL SIMULATIONS OF THE HUMAN VOCAL TRACT USING THE 1D AND 2D DIGITAL WAVEGUIDE SOFTWARE MODEL

Jack Mullen, David Howard and Damian Murphy

Department of Electronics
University of York, Heslington, York, YO10 5DD, UK
{jm220|dh|dtm3}@ohm.york.ac.uk

ABSTRACT

This paper details software under development that uses the digital waveguide physical model to represent the sound creation mechanism and environment associated with the production of speech, specifically the human vocal tract. Focus is directed towards a comparison between the existing 1D waveguide method, on which several studies have already been conducted, and the developing 2D waveguide mesh method. The construction of the two models and the application of the tract geometry is examined, in addition, the inclusion of dynamic articulatory variations to increase the ability of such systems to create natural sounding speech is discussed. Results obtained from each suggest that the 2D model is capable of producing similarly accurate vowel spectra to that already accomplished with the 1D version, although speech-like sounds created with the 2D mesh appear to exhibit greater realism.

1. INTRODUCTION

The artificial reproduction of human speech sounds is accomplished to within acceptable levels of naturalness using established and widely used techniques such as linear prediction [1] or formant based synthesis [2]. Although of appropriate quality, these methods do not exploit the benefits inherent in the use of a physics based model in the quest for synthesised sounds of an organic nature.

Physical modelling synthesis is focused on the discretisation of continuous real world mechanics into manageable lumped element systems exhibiting behaviour that approximates the target structure. It has greatly improved the naturalness of sound created in the simulation of musical instruments over spectral reconstruction methods. The underlying mathematical laws governing sound and vibration are used to accommodate a virtual source-environment coupling in such models to allow for a greater scope of representation, and hence produce synthesised sounds of greater realism. Digital waveguide physical modelling can be used to produce accurate representations of vibrating structures, such as the column of air within a clarinet, or a plucked guitar string. Digital waveguides constructed in multidimensional mesh formation can be used to model resonating bodies of air and have been applied in room acoustics simulations [3].

Research focused on the development of a physical model of the vocal tract employing the digital waveguide method in one dimension [4] has produced synthesised speech sounds of an organic nature. Currently little is known about the possible improvements in naturalness of synthesis that may be achievable in a similar model extended towards a multidimensional waveguide system.

2. THE DIGITAL WAVEGUIDE MESH

Based on the discretisation of the acoustic wave equation, the digital waveguide comprises a bi-directional unit delay which forms the left going and right going components of a simulated pressure wave. The sum of such components represents the pressure value at each element. Multiple waveguides connected in a line as in Figure 1, with some applied termination constitutes the basic digital waveguide model of, for example, a resonating tube or string.

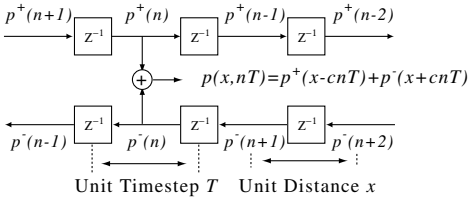

Figure 1: *A 1D Chain of Waveguides*

The length of the modelled structure is apparent in the length represented by each waveguide multiplied by the number of waveguides within the model. Scattering equations are performed, resulting in lossless propagation of an applied excitation along the chain of waveguides with boundary reflections observed at either end. Non-uniform scattering within the model can be set as an impedance discontinuity between waveguides.

This modelling method can be extended to higher dimensions with the construction of a digital waveguide mesh (DWM). The width of the target structure can be set within the model in the same manner as the length, resulting in the 2D DWM, modelling, for example, vibrations on the surface of a drum skin or the propagation of sound in a 2D plane through a room. A 3D DWM can be constructed in a similar manner.

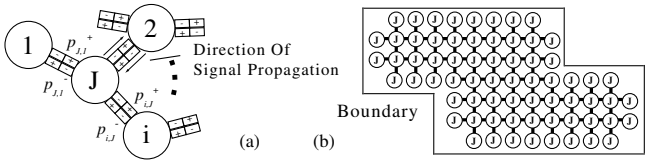

Figure 2: *(a) A Unit Junction, and (b) A Rectilinear Mesh*

Figures 2(a) and (b) detail the scattering junction with i arbitrary connections and the formation of the rectilinear (junctions with 4 neighbours at 90° from each other in a cartesian coordinate system) mesh, respectively. In Figure 2(b), air pressure values labelled $p^+_{J,i}$ indicate an incoming pressure at node J from node i (at i a unit time step before), and those labelled $p^-_{J,i}$ show the outgoing pressure at node J, to node i (reaching node i a time step later). 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 with N equal impedance waveguide connections is:

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

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

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

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

$$p^+_{J,i} = z^{-1}p^-_{i,J} \tag{3}$$

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 is:

$$p_J = (1+r)p^+_{J,1} \tag{4}$$

Equations (1)-(3) can also be derived as an equivalent finite difference scattering algorithm:

$$p_J(n) = \frac{2}{N}\sum_{i=1}^{N} p_i(n-1) - p_J(n-2) \tag{5}$$

This mathematical simplification results in a mesh scattering methodology which is easier to implement and more efficient in terms of memory requirements and speed of computation.

3. THE WAVEGUIDE VOCAL TRACT MODELS

Both 1D and 2D waveguide models representing the vibrating air cavity in the vocal tract between the glottis (vocal folds) and the lips have been constructed. Windows dialog-based software (Figure 3) has been developed to allow the comparison of 1D and 2D performance at equal levels of sophistication so as to highlight the potential benefits available from the increased dimensionality.

Quantitive comparison between the two models are made on the proximity of simulated formant frequencies and their bandwidths to that predicted by natural recorded speech. Although not an exact measure of accuracy owing to the varied nature of speech, this method allows for model parameters to be initialised.

The speech sound producing capability of each of the models depends greatly on accurate data describing the individual geometrical structure of the tract in the creation of each of the vowels. Such information in the form of functions describing the cross sectional area along the tract from x-rays taken of Russian speakers have been used in these simulations [5].

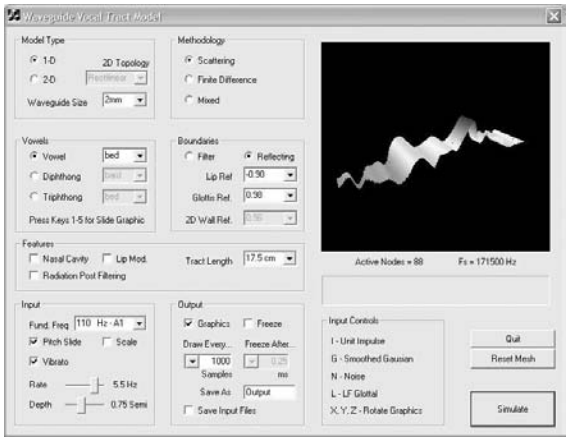

Figure 3: *Waveguide Vocal Tract Modelling Software Screenshot*

3.1. Construction

The 1D model is constructed using a chain of 88 waveguides, the area A of each tube section i determines the waveguide impedance Z_i according to Equation (6), where ρ is the air density and c is the speed of sound.

$$Z_i = \frac{\rho c}{A_i} \tag{6}$$

Figure 4 illustrates the manner in which the discretisation of the tract into many tube sections of differing area is represented as a string of varying impedance waveguides.

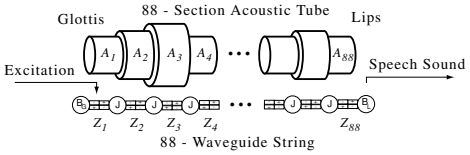

Figure 4: *The 1D Waveguide Vocal Tract Model*

Excitation is applied to the model on the waveguide at the glottis end in the form of either continuous noise, resulting in system spectral characteristics, or the periodic LF glottal flow derivative waveform [6], resulting in speech-like sound. Output is taken from the boundary node at the lip end of the model. Simple reflecting boundaries are applied at either end to simulate the internal reflection at the glottis ($r_{glottis} \approx +1$) and the inverting reflection at the (open) lips end ($r_{lips} \approx -1$).

The 2D model incorporates the width of the physical structure represented by the waveguides as a mesh 88 nodes along from glottis to lips, and multiple waveguides across dependant on the relative diametre value. The diameter values used assume a circular cross sectional area as implied in the 1D tube model. The width apparent in the second dimension represents the distance across the tract as seen in the mid-sagittal plane. The intention is that a 2D model will allow propagation of higher order modal reflections across the tract as well as along. Figure 5(a) illustrates the use of a DWM to model a straight tube open at one end or *quarter wave resonator* (QWR), and Figure 5(b) shows the inclusion of an arbitrary area function along the length of the vocal tract.

Boundary reflection coefficient values are set in three places;

the open lip end r_{lips}, the closed glottis end $r_{glottis}$, and the internal fleshy wall reflection r_{wall}. Excitation is applied in the form of either noise or the LF glottal waveform, injected on to the mesh along the line of nodes closest to the glottis. Output is then measured as an average of all nodes along the lip boundary.

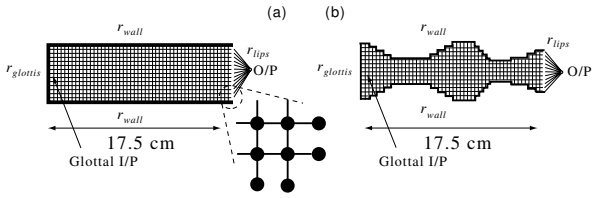

Figure 5: *(a) QWR and (b) The Vocal Tract 2D DWM Models*

Visual output from the two models is presented in the form of an OpenGL 3D graphics window on the main dialog panel (Figure 3) illustrating a planar view of the air pressure within the tract as it varies over time. Data output is saved as a *.wav* file. This allows for continual monitoring of the chain/mesh behaviour at any spatial or temporal point in a simulation.

3.2. Features

Both models can be constructed using the scattering or finite difference methods to allow the study of the various benefits offered by each approach. Mixed modelling, as supported by the use of the KW interface [7], is facilitated in the 1D model; the potential mobility of the 2D model boundaries leads to dismiss the application to the higher dimensional system.

In order to increase the quality of the model such that it is closer to natural speech or singing, many additional factors are included. Vibrato, tremolo and noise can be introduced into the input waveform to include a varied, more human element to the excitation. The simulation of constrictions to the air flow and plosive speech like sounds is achieved with both the modulation of the reflection coefficient at the lips end of the model, and a decrease in tract width modelling the tongue.

3.3. Diphthong Simulation

An important factor in the synthetic production of speech is the ability to create diphthongs and triphthongs, sounds formed from a slide between two or three vowels. In a physics based model this can be accomplished by linear interpolation between area functions set within the model. The 1D model allows for a simple update of the impedance values in an allotted time frame (typically about 250ms for a triphthong), whereas implementing this for the 2D system proves less trivial.

The application of the area function to the 2D model as the amount of waveguides across each section of the tract adds much complexity to the idea of a dynamic model. A mesh with moving boundaries requires dynamic node reallocation to accommodate the addition and removal of extra nodes into the mesh and the adjustment of those surrounding each change. Balancing of waveguide pressures around each of the changing nodes will also be necessary to accommodate this alteration to the model. This feature of the 2D model is currently under construction, but it is believed that once complete it will enhance the potential of the multi-dimensional model to create speech sounds of an organic nature.

4. RESULTS

The output from the simulations produced by the software can be used to gain an understanding of the importance of the use of physical modelling in speech synthesis, specifically any benefits that might be gained by moving to a 2D or 3D formulation.

4.1. Vowel Simulation

The simulation of any speech-like sounds relies heavily on the creation of many of the different vowel types obtainable in the real world, each characterised by its unique formant frequency pattern. The 1D vocal tract model has been seen to exhibit accurate formant simulation and hence vowels that approximate the target with some realism [4]. Similarly, more recent research into the 2D mesh model has shown that it produces speech-like output with accurate formant values for certain vowels [8]. Figure 6 shows the spectrum synthesised with the 1D model when creating the vowel in the word *'boot'*. The frequencies of the simulated formants give a good agreement with those predicted by recordings of natural speech [9]. Although there is up to 30% variability in some of the formant frequencies, the results are considered accurate enough due to the variability of the formant frequencies in measured natural speech.

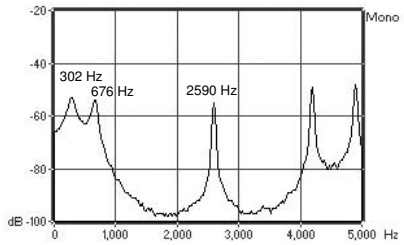

Figure 6: *Frequency Response of the 'boot' Vowel 1D Model*

Figure 7 illustrates the formant pattern produced by the 2D mesh model for the shape of the tract held when creating the vowel in the word *'boot'*. The frequencies of the simulated formants give a good agreement with those of natural speech, with only a 6.4% maximum variability between them.

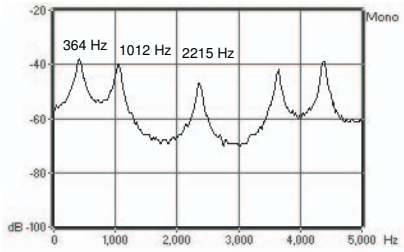

Figure 7: *Frequency Response of the 'boot' Vowel 2D Model*

Initial results show the frequency peaks for both models to be of a reasonable accuracy and as such both methods produce spectra of an expected shape. With application of the LF glottal waveform to both models vowel sounds are achieved that bear a close resemblance to their targets. The 2D model presents a slight increase in

quality of naturalness over the 1D model, which exhibits a small amount of metallic ringing.

4.2. Formant Bandwidths

The main difference in the two spectra in Figures 6 and 7 is observed as a reduction in bandwidth produced by the 1D model. The ability to control the bandwidths of the formants created will have an important implication on the system's overall potential to create realistic sounding speech. The bandwidths of the formants produced by both waveguide models are directly influenced by the reflection values set at the boundaries. The extended control offered by the additional boundaries in the 2D model may therefore present greater flexibility. Current research is involved in the optimisation of reflection coefficient values for both the two (1D) and four (2D) boundary systems. Figure 8 shows the variations in bandwidth of the first formant in the spectrum of the vowel in *'beet'*, achieved by changing $r_{glottis}$, keeping $r_{Lips} = 0.6$. For comparison, 2D formant bandwidth variations achieved by varying r_{wall}, with higher, more realistic reflection coefficient values $r_{lips} = 0.9$ and $r_{glottis} = 0.98$, are also shown in Figure 9. Preliminary results highlight more clearly defined peaks generated by the 2D model when compared to the 1D equivalent, approaching target bandwidths of between $80 - 100Hz$. Similarly, greater sensitivity towards smaller changes in r_{wall} are apparent in the larger range of 2D bandwidths achieved.

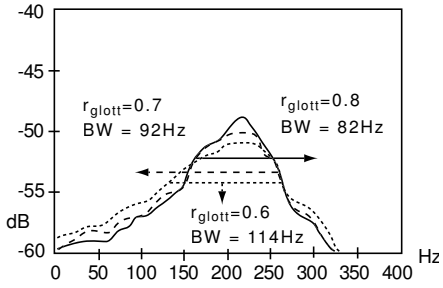

Figure 8: *1D 'beet' Vowel Formant Bandwidth Variations*

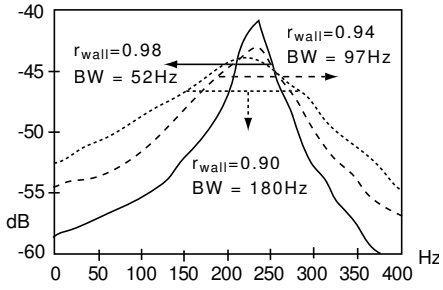

Figure 9: *2D 'beet' Vowel Formant Bandwidth Variations*

5. CONCLUSIONS

The software presented in this paper is designed to allow a thorough comparison of the 1D and 2D waveguide vocal tract models. Current results suggest that the 2D model can accurately simulate vowel formant frequencies, verifying the methods ability to generate the building blocks of speech. The 2D model also allows for greater control of formant bandwidths, giving enhanced sensitivity for smaller changes in mesh reflection coefficients. The ease with which the 1D model employs a vowel-to-vowel slide, when compared with the complex junction reassignment necessary in the 2D case, may present difficulties in justifying the eventual argument of higher dimensional superiority, but it is hoped that diphthongs of greater realism may be achieved with 2D modelling given its superiority for static vowel sounds.

Further features such as mesh boundaries simulating a more realistic frequency dependent reflection, a lip radiation filter accounting for the open end, and a nasal tract, will increase the potential of the 2D model further. With more comprehensive tract geometry data, an additional consideration might be the inclusion of a third dimension to the model, resulting in a ful physically modelled vocal tract. Future testing will involve the use of area functions generated from natural recorded vowels, which can then be re-synthesised with the waveguide models for a direct comparison. The eventual goal is to extend both systems to the sophistication and realism achieved in more developed models, such as the SPASM system, with a view to examining the possible increase in naturalness that may be achieved with the higher dimension model.

6. REFERENCES

[1] J. Makhoul, *Linear Predictive Coding in Electronic Speech Synthesis*, R. R. Donnely and Sons Co., 1984.

[2] Dennis H. Klatt, "Software for a cascade/parallel formant synthesiser," in *Journal of the Acoustical Society of America*, 1980, vol. 67(3), pp. 971–995.

[3] S. A. Van Duyne and J. O. Smith, "Physical modeling with the 2d digital waveguide mesh," in *Proc. International Computer Music Conference*, Tokyo, Japan, 1993, pp. 40–47.

[4] Perry R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model with Applications to the Synthesis of Singing*, Ph.D. thesis, Stanford University, USA, 1991.

[5] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[6] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," in *Quarterly Progress Report*, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweeden, 1986.

[7] Matti Karjalainen, "Mixed physical modelling: Dwg + fdtd + wdf," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2003, pp. 225–228.

[8] Jack Mullen, David M. Howard, and Damian T. Murphy, "Digital waveguide mesh modelling of the vocal tract acoustics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2003, pp. 119–122.

[9] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley and Sons Inc., 2000.