A STUDY ON DYNAMIC VOCAL TRACT SHAPING FOR DIPHTHONG SIMULATION USING A 2D DIGITAL WAVEGUIDE MESH

Anocha Rugchatjaroen and David M. Howard
Audio Lab, Department of Electronics, University of York
York, UK
ar718@york.ac.uk and
david.howard@york.ac.uk

ABSTRACT

This paper presents a study of an articulatory-based speech synthesis based on a 2D-Digital Waveguide Mesh (2D-DWM) to model acoustic wave propagation in the oral tract. It is employed to study the effects of changing oral tract area, and in particular, of moving the articulators during the production of diphthongs. The operation of the synthesizer including details of how diphthongs are produced are discussed. The results support earlier findings that the wall reflection coefficient is inversely proportional to the formant bandwidth.

1. INTRODUCTION

Understanding the acoustic nature of speech signals provides the basis for the design of speech synthesis systems. The first attempt to produce vowel-like sound was reported by Ch. G. Kratzenstein in 1779 [1]. Twelve years later von Kempelen simulated a vocal tract mechanically using bellows as a power source, an ivory reed as the sound source and a leather sack as the oral tract (sound modifiers). These two devices are the results of the earliest research into articulatory-based speech synthesis, since they are direct models of the speech production system.

The shape of the vocal tract can be altered using the main articulators (tongue, jaw and lips) and the resulting tube shape modifies the sound source from the glottis during voiced speech. To simulate articulation, Childers (2000) separates the modelling into two parts (the articulatory model and the acoustic model). The articulatory modelling is the part that physically considers shaping of the vocal tract, which could be viewed as a structure of concatenated small ducts. The sizes of those are available from a set of cross-sectional area data from computerised tomography CT or magnetic resonance imaging MRI scans. These data are also used in the acoustic modelling as parameters to represent the volume of the tract [2].

In this article, we consider acoustic modelling and focus on the control parameters available in the 2D-Digital Waveguide Mesh (2D-DWM) to study the effect of dynamic modelling in the time domain. The evaluation will focus on the reflection coefficient of the wall of the tract in moving boundaries (articulator), which can be found in diphthongs.

The diphthong is a production of changing one monophthong to another. In English there are eight diphthongs /eI/, /aI/, /OI/, /@U/, /aU/, /I@/, /E@/ and /U@/ (Speech Assessment Methods Phonetic Alphabet (SAMPA) [3] phonetic symbols) classified into two groups – closing and centering. The classification relates to the direction of the changing positions of the vowels in the vowel quadrilateral [4].

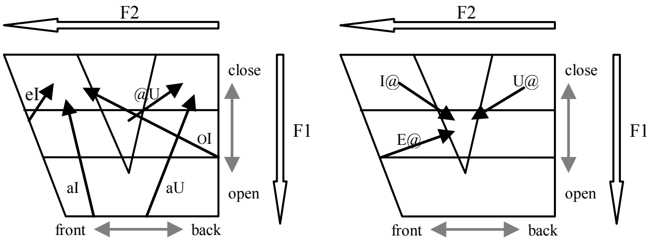


Figure 1 : *The changing position of vowels in English diphthongs on the vowel quadrilateral (after [3]).*

Mullen (2006) describes the link between the reflection coefficient and formant bandwidth as an inverse relationship between the coefficient and the bandwidth [5]. He also shows an example of a synthesized /I@/ diphthong where he analysed its changing formant frequency, but he did not examine the whole range of English diphthongs. This article extends his work by examining the relationship in more detail, particularly during the synthesis of other diphthongs.

2. 2D-DIGITAL WAVEGUIDE MESH

The 2D-Digital Waveguide Mesh (2D-DWM) is one of the algorithms that has been used for simulating acoustic wave propagation. It has firstly been used to simulate a sound propagation in the human vocal tract in [6]. It is an adaptation of the 1D Digital Waveguide. The 1D simulates the wave propagation on a chain of waveguides connected together and passing the wave parameters in both directions. For the 2D version, more ports are attached to connecting points (junctions) to create a higher dimensionality for the travelling wave. At the junction, the mesh system works as its topology. In this article we use four port scattering junctions, which have four connecting ports attached to a junction. The propagation is done by calculating the pressure output of port *k* using:

$$p_k^- = p_J - p_k^+ = \frac{2\sum_{i=1}^{N} Y_i p_{J,i}^+}{\sum_{i=1}^{N} Y_i} - p_k^+ \qquad (1)$$

where: p_J is a junction pressure, p_k is pressure at the k^{th} port, '-' and'+' indicates direction of travelling parameters (incoming and outgoing), N is a number of attached ports (in case of the previous example, N is 4) and Y_i is an admittance of the i^{th} port.

The admittance is a variable that is used as a weight of each waveguide. It is calculated from the area function that the waveguide represents. Here we are considering the vocal tract, and therefore it is based on the cross-sectional areas along the tract. A set of areas along the length of the tract is created based on scanning (e.g. MRI) and these are perpendicularly oriented to the direction of the air flow. The areas are used to calculate the admittance by considering them in terms of their radius.

$$Y_x = \frac{A(x)}{\rho c} = \frac{\pi\, r^2(x)}{\rho c} = \frac{1}{Z_x} \qquad (2)$$

where r(x) is a radius of the area at x, x indicates the location of the cross-sectional area A(x), ρ is density of the air, c is the velocity of the air, Z_x *is an impedance at x.*

The 2D-DWM in [7] improves the meshing algorithm by applying a cosine function to control the impedance across the tube on the Y-axis. Figure 2 shows increased impedance along the edges of the straight tube. The flipped up bell curve represents the impedance hill Z_x at the point of constriction x. $Z_{x,0\ \dots\ n}$ represent impedance values at different y positions and the maximum impedance value of a constriction x is $Z_{x,0}$ and the minimum is $Z_{x,n/2}$ or Z_{min} which equals to the Z_{tube}.

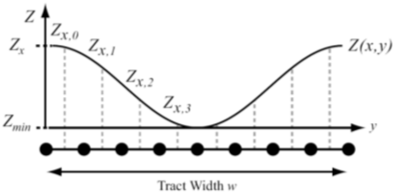


Figure2 : *Raised impedance hills causing a constriction in a straight tube and plotted raised cosine impedance hills on either side of the constriction from [7].*

Equation 3 shows the cosine smoothing across the impedance hill *Z(x,y)* where w represents a tract width (from [7]).

$$Z(x,y) = Z_x - \frac{(Z_x - Z_{min})}{2}\left[1 + \cos\left(2\pi\left(\frac{y}{w} - \frac{1}{2}\right)\right)\right] \qquad (3)$$

Since the simulation in 2D has one more boundary condition than in 1D (wall reflections), the reflection coefficient of the wall is involved in the increasing of the impedance values at the wall when $y = 0$ and $y = n$. The experiment, assessing the flexibility of the wall reflection coefficient is shown in the next section.

3. EXPERIMENT AND RESULTS

As there are four boundaries to the mesh, each is being represented and controlled by the reflection coefficients – lips, glottis and the two side-walls. The wall reflection coefficient affects the formant bandwidth but the lip and glottis reflection coefficients do not [6]. The higher the wall reflection coefficient the lower is the resulting formant bandwidth. A small formant bandwidth leads to a greater potential distinction between different vowels by reducing the overlap between adjacent formants when they are close in frequency.

In our experiment, we are studying the effect of the wall reflection coefficient in diphthongs when the wall or the area data is moving or changing. The system tests the effect of movement by simulating the wave propagation using white noise as the sound source to allow tracking of all of the changing resonant frequencies. The reflection coefficients are varied as follows: 0.90, 0.92, 0.94, 0.96, 0.98 and 1.0. Figure 3 shows formant bandwidth of eight synthesized diphthongs overlaid using a Hamming window with the following parameters: length 0.049 seconds, 0.7 pre-emphasis, 0.01 second of frame interval and LPC order 12. The data presented in the figure are the average of the formant bandwidths when the tract is changing its size.

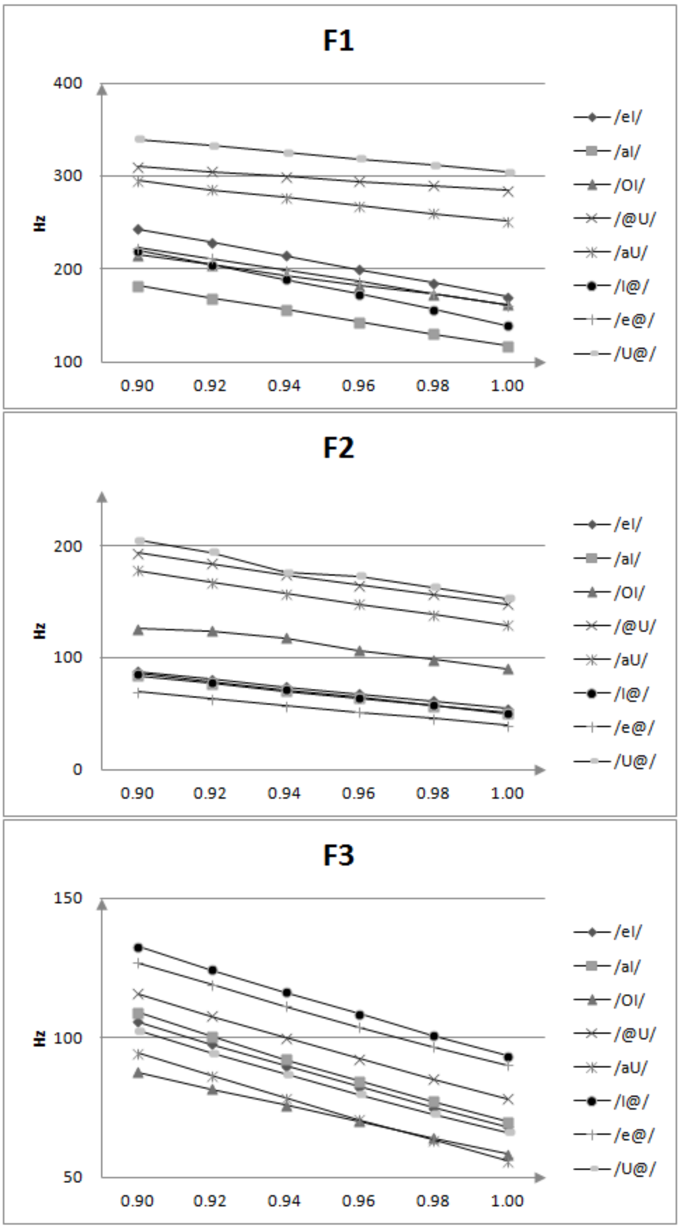


Figure 3 : *The formant bandwidth of eight synthesized English diphthongs using various wall reflection coefficient (0.90, 0.92, 0.94, 0.96, 0.98 and 1.0).*

Figure 3 shows the results of speech analysis for the bandwidths of F1, F2 and F3. They are decreasing when the coefficient is getting slightly higher. This means that the damping in the time domain or the absorption of the sound energy by the moving boundaries could also be controlled by fixing the wall reflection coefficient. Meanwhile, the distinctiveness of each monophthong

end-point component in each diphthong can be reduced by increasing the formant bandwidth or decreasing the coefficient.
Figure 4 shows an example of the bandwidth variations during a diphthong /@U/ averaged over 4 different waveguide sizes. The experiment used the same configuration as in previous one but plotted the averaged bandwidth frame by frame. It still depicts an effect of how decreasing the bandwidth was when we increase the coefficient. On the other hand it also reports that the 2D-DWM resonates F1 in wider bandwidth than F2 which is opposite of what happens in natural speech.

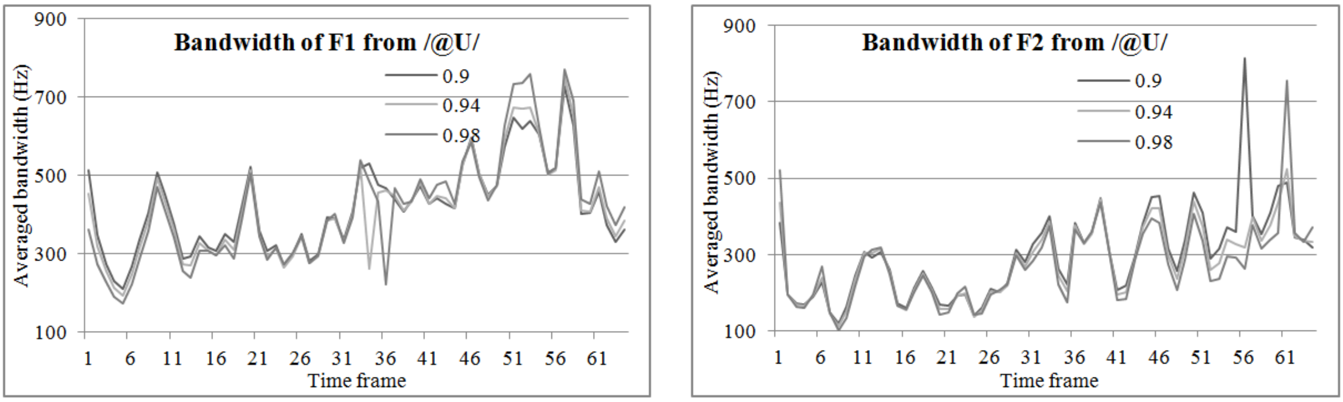
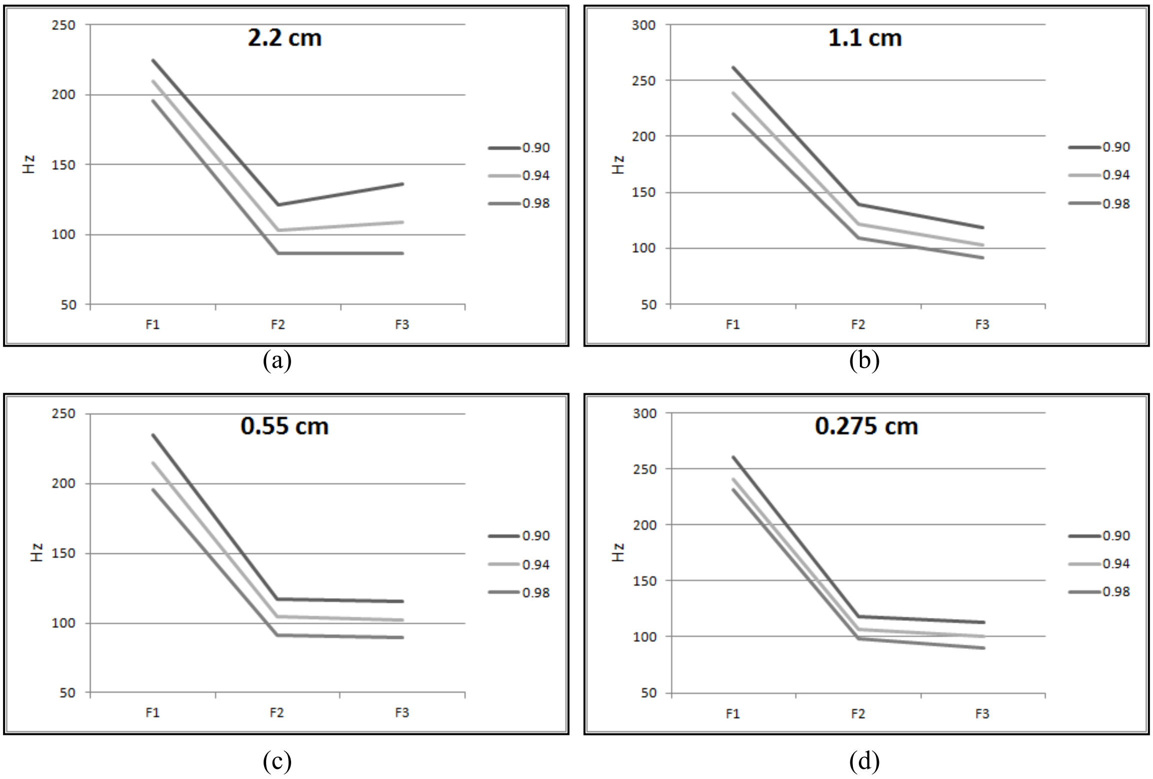


Figure 4 : *Formant bandwidth analysis results of F1 (left) and F2 (right) of each time frame from in synthesized /@U/ averaged from using waveguide size 2.2, 1.1, 0.55 and 0.275 cm.*

Moreover, we also evaluate the effect of changing waveguide size on the bandwidth in order to understand the effect of using a more dense mesh. The results show the same trend of decreasing of the bandwidth when the reflection coefficient is increased in all waveguide sizes tested. Figure 5 shows the analysis results overlaid using 0.049 sec of Hamming window length with 0.7 Pre-emphasis, 0.01 sec of a frame interval and LPC order 12 on the frequency analysis.



(a)

(b)

(c)

(d)

Figure 5 : *Formant bandwidth results from different waveguide size using wall reflection coefficient as 0.90, 0.94 and 0.98. (a) waveguide size = 2.2 cm, (b) waveguide size = 1.1 cm (c) waveguide size = 0.55 cm (d) waveguide size = 0.275 cm.*

These results show that there are some disadvantages to using the 2D-DWM when the formant bandwidth is wider than that of natural speech. However, as we mentioned before, they do give us some idea of how to control them using the reflection coefficient configuration. A set of synthesized sounds from our experiment can be found in formant plots in the appendix. They are synthesized results of using 0.55 cm of the waveguide size. Small movements on F1 are the results of usage of area based 2D-DWM as discussed in section 4.7.1 in [7] but they show reasonable F2 movements.

4. DISCUSSION AND CONCLUSION

In Vocal tract modelling of diphthongs the moving boundaries will change the area data and this will affect formant frequencies. In the present research we looked at an additional effect of this movement, namely, that on formant bandwidth. The results not only support the inverse relationship between the wall reflection coefficient and the bandwidth which was discussed in [6], they also show that this relationship holds true for all English diphthongs when synthesised. In detail, when the first three formants are morphed from one vowel to another the damping is still linearly reverse proportional to the wall reflection coefficient. This suggests that we can control the formant bandwidth even when the boundaries are moving which means there is a possibility in using a flexible wall reflection coefficient to achieve more human-like synthesised speech. Applying flexible wall reflection coefficients with 2D-DWM will be advantageous to the simulation of co-articulation in the vocal tract. In addition, changing of the waveguide size did not affect the above-mentioned relationship between the reflection coefficient and bandwidth.

All in all, these results suggest that when synthesising English diphthongs using a 2D-DWM, the formant bandwidths can be controlled by utilising flexible wall reflection coefficients. In order to gain more support for this finding a perceptual study will be conducted.

5. ACKNOWLEDGMENTS

We would like to thank colleagues Christin Kirchhuebel and Aglaia Foteinou for all their support.

6. REFERENCES

[1] S. Lemmetty, “Review of Speech Synthesis Technology,” Master's Thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, 1999.

[2] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley, NY, 2000.

[3] J. C. Wells (1989), Computer coded phonemic notation of individual languages of the European Community [SAMPA transcription]. *Journal of the International Phonetic Association*, 19, 32-54.

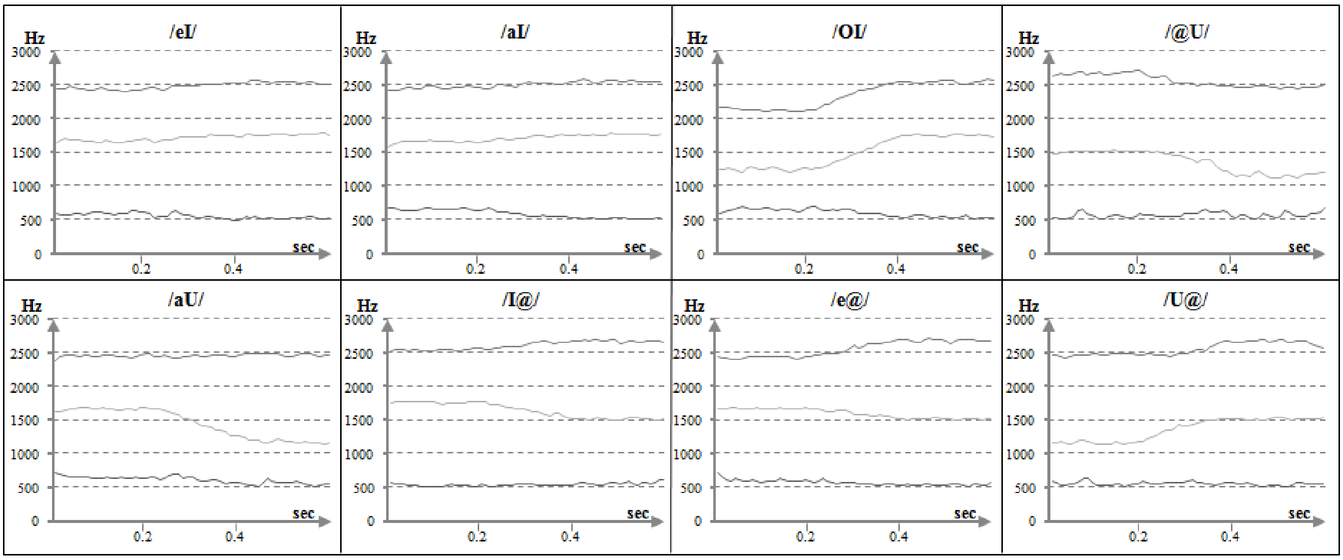
[4] D. M. Howard and J. A. S. Angus, *Acoustics and Psychoacoustics*, Elsevier, Oxford, fourth edition, 2009.

[5] J. Mullen, D.M. Howard, and D.T. Murphy, “Waveguide Physical Modeling of Vocal Tract Acoustics: Flexible Formant Bandwidth Control from Increased Model Dimensionality,” *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 3, pp. 964-971, 2006.

[6] J. Mullen, D.M. Howard, and D.T. Murphy, “Acoustical simulations of the human vocal tract using the 1D and 2D digital waveguide software model,” in *Proc. 7th Int. Conf. On Digital Audio Effects (DAFx-04)*, Italy, 2004, pp. 311-314.

[7] J. Mullen, *Physical Modelling of the Vocal Tract with the 2D Digital Waveguide Mesh*, Ph.D. dissertation, Dept. Elect., University of York, April 2006.

7. APPENDIX: F1, F2, F3 OF THE EIGHT SYNTHESIZED ENGLISH DIPHTHONGS


/eI/
/aI/
/OI/
/@U/
/aU/
/I@/
/e@/
/U@/
Hz
sec