MODELING VOICED STOP CONSONANTS USING THE 3D DYNAMIC DIGITAL WAVEGUIDE MESH VOCAL TRACT MODEL

Amelia J. Gully[1], Benjamin V. Tucker[2]

[1]Department of Language and Linguistic Science, University of York, UK, [2]Department of Linguistics, University of Alberta, Canada
amelia.gully@york.ac.uk, bvtucker@ualberta.ca

ABSTRACT

Three-dimensional (3D) acoustic simulations of the vocal tract are showing significant promise for the study of speech acoustics. Recent models have demonstrated dynamic behaviour, but currently only vowels and sonorant consonants have been reproduced, limiting the applications of such models. We present a method for producing voiced stop consonants in an intervocalic context, using a 3D digital waveguide mesh (DWM) simulation based on magnetic resonance imaging data of the vocal tract. The synthetic output demonstrates appropriate formant transitions leading to several intelligible stop consonants. This method represents a step towards a complete phoneme inventory for 3D vocal tract models, and demonstrates the suitability of the 3D DWM vocal tract model for the simulation of dynamic speech elements. The proposed method also offers considerable opportunity for controlled perceptual study of the acoustics of voiced stop consonants.

Keywords: speech synthesis, stop consonants, vocal tract modeling, digital waveguide mesh.

1. INTRODUCTION

Three-dimensional (3D) acoustic simulations of the vocal tract, usually based on magnetic resonance imaging (MRI) data, have appeared in the literature over approximately the last decade (e.g. [12, 16, 1]). These detailed models offer unparalleled opportunities for research into the physics of speech production and accurate simulation of speech. However, many 3D vocal tract models are static, capable of producing held vowels only. Although recent models have been shown to be capable of synthesizing diphthongs and sonorants [6, 4], obstruents have not yet been simulated, limiting the use of 3D simulations to the study of isolated phonemes. This paper presents a method for synthesizing voiced stop consonants based on the 3D dynamic digital waveguide mesh (DWM) vocal tract model presented in [6].

A stop consonant occurs when there is a complete occlusion of the vocal tract. Stop consonants have three phases based on the movement of the articulators: closure, hold, and release. The movement of the articulators results in formant transitions, which are important cues for identifying the place of the articulation of the stop consonant [3]. Stop consonants show a wide range of variability in natural speech, often omitting, or only partially realising, one of the above phases [17]. A 3D vocal tract simulation capable of modelling all aspects of stop consonant variability would be of considerable value for controlled studies of their perception.

Stop consonants have been well studied using one-dimensional (1D) simulations of the vocal tract. For example [13, 14, 15] describe a 1D vocal tract model based on vocal tract area functions derived from MRI data. Stops in an intervocalic context are implemented as a closing gesture local to the occlusion location, superimposed on an underlying vowel-vowel transition across the rest of the tract. This approach has been found to appropriately reproduce the acoustics of stop consonant production. An alternative system [2] uses a 3D model of the vocal tract articulators, based on MRI data, to control a 1D vocal tract simulation. However, a true 3D simulation method offers opportunities for more detailed study, particularly regarding acoustic phenomena not reproduced under the plane wave assumption inherent in 1D simulations.

An early study using a 2D dynamic DWM vocal tract model [10] found that momentary occlusions of the modeled tract produced stop-like sounds, in particular /b/, but no systematic investigation into the acoustics of the simulated stops was performed. By combining the methods introduced in the above studies with 3D acoustic simulations based on the detailed vocal tract geometry, it is anticipated that intelligible simulations of stop consonants can be achieved. Existing 3D vocal tract airway models lack a nasal tract or turbulence modelling, so this study considers only the voiced English stops /b/, /d/ and /g/, which feature a closed velum and are identifiable without a turbulent burst [8]. Stops are sim-

ulated in an intervocalic context using a 3D DWM technique. A longer-term aim of this research is to study the effect of varying model parameters on stop consonant perception.

2. METHOD

2.1. MRI data collection

Volumetric MRI data was collected for one phonetically-trained subject (female British English speaker, age 28) in a GE 3T Signa Excite MRI scanner, with the following parameters: 3D GRE sequence, TR 4.736ms, TE 1.68ms, FA 5°, 80 contiguous 2mm sagittal slices with no gap. This resulted in a 16 second scan time. All static British English phonemes were captured and this study uses data for the vowels /i/ and /a/ and the hold phases of stop consonants /b/, /d/, /g/. Due to scan time limitations, it was not possible to obtain MRI data for stops in vowel-specific contexts, so the scan subject was instructed to produce them in a 'neutral' (schwa) context. It was shown in [15] that consonant constrictions can be implemented as a localized constriction on a vowel-vowel articulatory transition in a 1D vocal tract model. This study tests this approach with 3D volumetric data, superimposing the closures for /b/, /d/ and /g/ upon the vocal tract shapes for the vowels, as illustrated for /i/ in Figure 1.

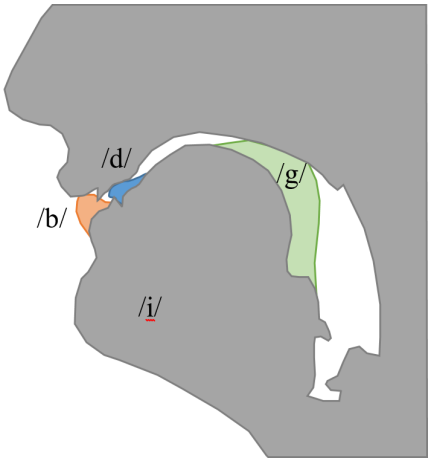
The MRI data were segmented using the automatic region growing algorithm in itk-SNAP [18] and hand-corrected to remove leakage into surrounding tissues. An additional scan capturing the teeth was used to remove teeth from the segmentations using a method similar to that of [19]. The segmentations were converted to 3D Cartesian grids, following [6], to permit DWM modelling.

2.2. The DWM model

The digital waveguide mesh (DWM) is a multidimensional numerical acoustic modeling technique, previously demonstrated for vocal tract modelling in [10, 12, 6], with [10] and [6] introducing a dynamic DWM vocal tract model in 2D and 3D respectively.

The 3D dynamic DWM approach models the vocal tract as part of a cuboid domain comprising a Cartesian mesh of short interconnected waveguides. Acoustic pressures are propagated throughout this domain during a simulation, and acoustic scattering takes place where acoustic impedance (Z) changes. Each waveguide is assigned a Z value corresponding to air ($Z_{air} = 399\ \mathrm{Pa\,s\,m^{-3}}$) or surrounding tissue ($Z_{tissue} = 83666\ \mathrm{Pa\,s\,m^{-3}}$), based on its equivalent location in the MRI data [6]. By interpolating between different impedance 'maps' over the course of a simulation, dynamic movement of the vocal tract is simulated. A source signal representing the glottal flow waveform is added sample-by-sample at a mesh location corresponding to the physical location of the glottis, and the output sound pressures are recorded at a mesh location corresponding to a microphone position close to the mouth, to obtain a synthetic speech waveform.

Figure 1: Midsagittal profile from MRI data for vowel /i/, showing the locations of superimposed consonantal closures for /b, d, g/.

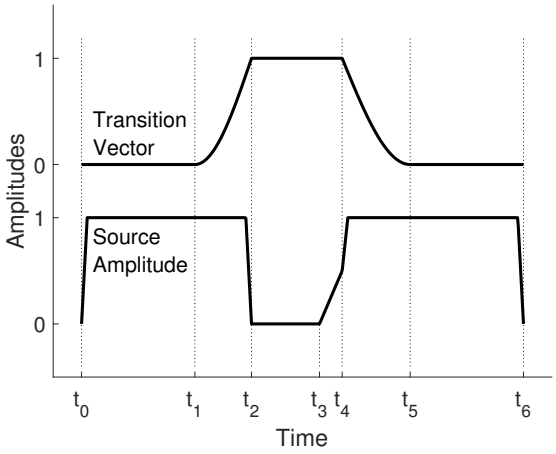


2.3. Articulation

Articulation in the DWM simulation is modeled by interpolating between impedance maps. Simulating /ibi/, for instance, would require two impedance maps—one each for /i/ and /b/—and at any time sample, the impedance values throughout the domain will be some weighted combination of these. As a result, rather than true articulator movement during a transition, any location where the articulator has been or will be takes on an intermediate impedance value. This non-physical behaviour has been shown to produce perceptually-acceptable diphthongs in [6], but its applicability for obstruents is not yet known. The time-dependence of the weights is described by a transition vector varying between 0 (vowel) and 1 (stop; i.e. a complete occlusion of the vocal tract).

In [2], the transition vector between one set of vocal tract articulator positions and another is given as a sigmoid function. However, because the articulator approaches a 'virtual target' beyond vocal tract boundaries, the *airway* closure is not released until almost halfway through the gesture. Since the proposed model considers transitions between airway

Figure 2: Transition (top) and source amplitude (bottom) vectors for 3D dynamic DWM simulation of VCV utterances.



shapes not articulator positions, the transition vector uses a quarter wavelength sine function to approximate the resulting airway shape transition, as illustrated in Figure 2. This results in abrupt transitions at $T = t_2$ and $T = t_4$ where the occlusion is formed or released by the action of the moving articulators.

2.4. Voice Source

The dynamic 3D DWM vocal tract model makes use of a source signal approximating the glottal flow (in this case the Rosenberg pulse [11]), which is input at a location in the model equivalent to the physical position of the larynx [6]. The pitch contour of the source signal is obtained from recordings of the MRI subject's speech averaged across VCV utterances. Jitter and shimmer are also added to the source signal to improve naturalness.

Since the simulation method does not currently feature interaction between the vocal tract and the voice source, the amplitude of the source signal is adjusted in conjunction with the transition vector, to mimic the natural onset and offset of phonation caused by the tract occlusion. As a result, it is possible to precisely control voice onset time (VOT)—the duration between the release of the closure and the onset of phonation—which is another important acoustic cue for stop consonants [9]. An example source amplitude envelope, demonstrating a negative VOT, can be seen in Figure 2.

The combination of articulator positions, transition timings and VOTs are expected to result in intelligible simulated VCV utterances. The next section presents the results of these simulations.

3. RESULTS

Spectrograms for the simulated /b, d, g/ in each context /i, a/ are shown in Figure 3 (pre-emphasis applied). The stops have an onset duration ($t_2 - t_1$) of 50ms, a closure duration ($t_4 - t_2$) of 80ms, a VOT ($t_3 - t_4$) of -20ms and an offset duration ($t_5 - t_4$) of 80ms. These parameters are within published ranges for stop consonants and were found to provide acceptable results across all vowel/stop combinations, although different parameter combinations were more suitable for certain VCVs. Audio examples are available at [7]. Impressions from informal listening are given below, but full perceptual testing is planned for the near future.

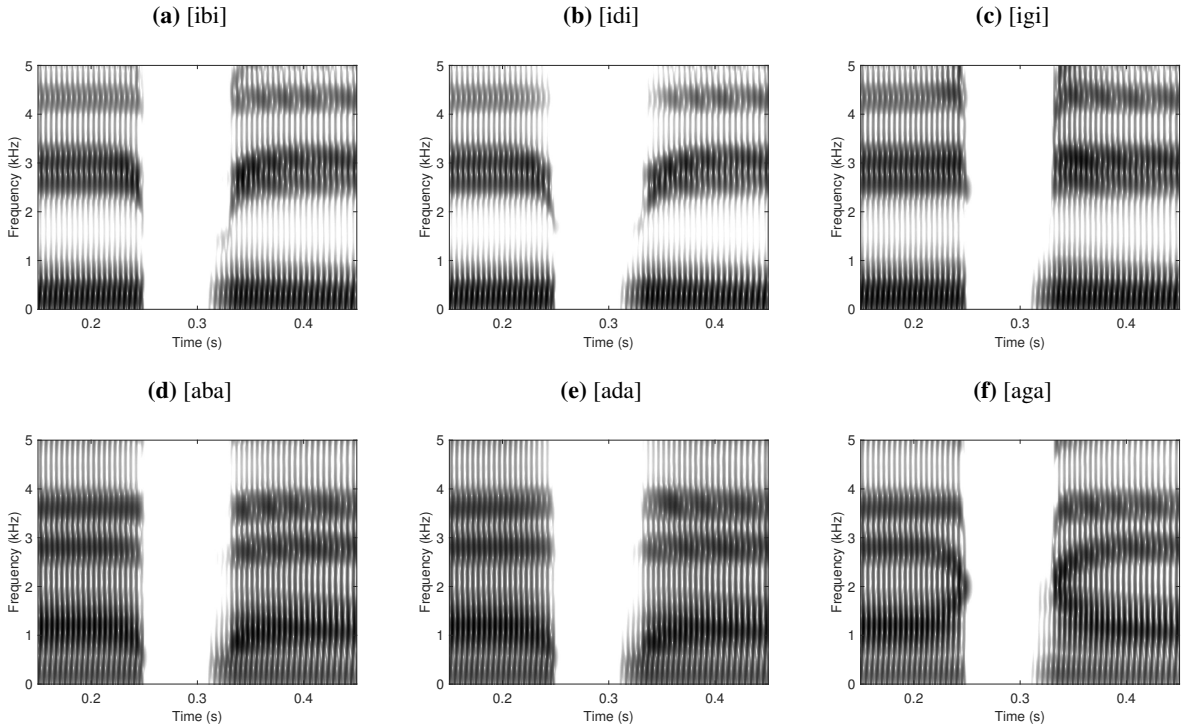
Simulations for /b/ (left column of Figure 3) show the decrease in F2 characteristic of a bilabial stop in both vowel contexts, with the decrease particularly significant for in the /i/ context given the high vowel F2. The audio data supports this, with /b/ intelligible in both vowel contexts. Simulations for /g/ are illustrated in the right column of Figure 3. An appropriate F2 transition, and a clear F3 transition, are demonstrated for /aga/, leading to an audible /g/ sound. Almost no transition is visible for /igi/, perhaps owing to the already high F2 for /i/, and a /g/ is not reliably perceived. Finally, simulations for /d/ are shown in the centre column of Figure 3. The F2 transition for /idi/ is similar to that for /ibi/ albeit slightly shorter, and for /ada/ the second formant decreases rather than increasing as expected. As a result both of these stops sound like /b/ rather than /d/.

4. DISCUSSION

The successful simulations of /ibi/, /aba/ and /aga/ demonstrate that the 3D dynamic DWM vocal tract model is capable of synthesizing acceptable stop consonants. This is particularly encouraging given the non-physical nature of phoneme transitions using the current method (as described in Section 2.3). Furthermore, this suggests that superimposing a consonant closure upon the vocal tract shape for a vowel is valid even in 3D simulations, providing further support to the hypothesis that articulators not directly involved in the occlusion have limited impact upon the acoustics of the resulting consonant, as previously demonstrated in 1D [14].

In both vowel contexts, the simulations of alveolar stops have formant transitions more appropriate for bilabials, and indeed were perceived as such. This indicates that the timing parameters were appropriate for a stop, but the occlusion was sufficiently advanced to appear bilabial. This may be due to the underlying MRI data, as overcompensation for

Figure 3: Spectrograms for simulated VCV utterances with $t_2 - t_1 = 50$ms, $t_4 - t_3 = 20$ms, $t_5 - t_4 = 80$ms. Vowels extend for 200ms either side of consonant closure but are cropped for clarity.



the effect of gravity while supine in an MRI scanner has previously been observed in e.g. [5], resulting in an advanced tongue position. Hyperarticulation is a natural consequence of holding articulations for an MRI scan [5], but a 16-second hold phase is particularly unnatural and may inherently result in a vocal tract shape that differs from that of normal speech. It is also probable that limitations of the simulation method, described in [6], such as having a single high value for the vocal tract wall impedance, or the non-physical phoneme transitions described above, contribute to erroneous formant transitions.

The absence of a burst at the release of the stop does not appear to have been necessary for the perception of /ibi/, /aba/ and /aga/, but may be another factor affecting the identification of the alveolar consonants. Work on the simulation method is ongoing, to address the flow issue among other improvements. Furthermore, the optimum transition and source amplitude vectors will have subtly different shapes for each place of articulation, which may also provide cues to place of articulation.

The major strength of the 3D vocal tract modeling approach is that the user has fine-grained control over the geometry of the vocal tract airway once the model has been created. In addition to distance from the glottis, degree, length, and skewness of a constriction, as used in 1D models [14], it is possible to exert precise control over all aspects of the closure. Furthermore, transition vectors can be defined for different parts of the vocal anatomy independently. Upcoming work will make use of this flexibility to explore the impact of different closure shapes upon the perception of stop consonants.

5. CONCLUSION

This paper has presented evidence that 3D acoustic simulations of the vocal tract are capable of producing acceptable voiced stop consonants, representing a step towards a full phoneme inventory for such models. Future work will involve detailed study of the impact of changes in timing and articulation parameters on stop consonant perception.

ACKNOWLEDGMENTS

This work was partially supported by a Worldwide Universities Network Research Mobility Programme award.

6. REFERENCES

[1] Arnela, M., Guasch, O., Dabbaghchian, S., Engwall, O. 2016. Finite element generation of vowel sounds using dynamic complex three-dimensional vocal tracts. *Proc. 23rd Int. Congr. Sound Vib.* Athens, Greece.

[2] Birkholz, P. 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS One* 8(4). e60603.

[3] Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., Gerstman, L. J. 1952. Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America* 24(6), 597–606.

[4] Dabbaghchian, S., Arnela, M., Engwall, O., Guasch, O. 2018. Reconstruction of vocal tract geometries from biomechanical simulations. *International Journal for Numerical Methods in Biomedical Engineering* 0(0).

[5] Engwall, O. 2000. Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA. *Proc. INTERSPEECH* Beijing, China. 17–20.

[6] Gully, A. J., Daffern, H., Murphy, D. T. 2018. Diphthong synthesis using the dynamic 3D digital waveguide mesh. *IEEE/ACM Trans. Audio Speech and Language Process.* 26(2), 243–255.

[7] Gully, A. J., Tucker, B. V. Aug. 2019. Audio from: 'Modeling voiced stop consonants using the 3D dynamic digital waveguide mesh vocal tract model'. *http://doi.org/10.5281/zenodo.2616613*.

[8] Halle, M., Hughes, G. W., Radley, J. A. 1957. Acoustic properties of stop consonants. *J. Acoust. Soc. Am.* 29(1), 107–116.

[9] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3), 384–422.

[10] Mullen, J., Howard, D. M., Murphy, D. T. 2007. Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model. *IEEE Trans. Audio Speech and Language Process.* 15(2), 577–585.

[11] Rosenberg, A. E. 1971. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.* 49(2), 583–590.

[12] Speed, M., Murphy, D. T., Howard, D. M. 2014. Modeling the vocal tract transfer function using a 3D digital waveguide mesh. *IEEE/ACM Trans. Audio Speech and Language Process.* 22(2), 453–464.

[13] Story, B. H. 2005. A parametric model of the vocal tract area function for vowel and consonant simulation. *J. Acoust. Soc. Am.* 117(5), 3231–3254.

[14] Story, B. H., Bunton, K. 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification. *J. Speech Lang. Hear. Res.* 53(6), 1514–1528.

[15] Story, B. H., Bunton, K. 2017. An acoustically-driven vocal tract model for stop consonant production. *Speech Commun.* 87, 1–17.

[16] Takemoto, H., Mokhtari, P., Kitamura, T. 2010. Acoustic analysis of the vocal tract during vowel production by finite-different time-domain method. *J. Acoust. Soc. Am* 128(6), 3724–3738.

[17] Warner, N., Tucker, B. V. 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *J. Acoust. Soc. Am.* 130(3), 1606–1617.

[18] Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., Gerig, G. 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31(3), 1116–1128.

[19] Zhang, J., Honda, K., Wei, J. 2018. Tooth visualization in vowel production mr images for three-dimensional vocal tract modeling. *Speech Communication* 96, 37 – 48.