ARTICULATORY VOCAL TRACT SYNTHESIS IN SUPERCOLLIDER

Damian T. Murphy,
AudioLab - Department of Electronics
University of York
York, UK
damian.murphy@york.ac.uk

Mátyás Jani,
Faculty of Information Technology and Bionics
Pázmány Péter Catholic University
Budapest, Hungary
jani.matyas@itk.ppke.hu

Sten Ternström,
Dept. of Speech, Music and Hearing,
KTH
Stockholm, Sweden
stern@kth.se

ABSTRACT

The APEX system [1] enables vocal tract articulation using a reduced set of user controllable parameters by means of Principal Component Analysis of X-ray tract data. From these articulatory profiles it is then possible to calculate cross-sectional area function data that can be used as input to a number of articulatory based speech synthesis algorithms. In this paper the Kelly-Lochbaum 1-D digital waveguide vocal tract is used, and both APEX control and synthesis engine have been implemented and tested in SuperCollider. Accurate formant synthesis and real-time control are demonstrated, although for multi-parameter speech-like articulation a more direct mapping from tract-to-synthesizer tube sections is needed. SuperCollider provides an excellent framework for the further exploration of this work.

1. INTRODUCTION

In recent years it has been possible to model highly detailed 3-D models of the vocal tract based on the capture of Magnetic Resonance Imaging data from human subjects (e.g. [2] [3] [4] [5]). This has further enabled the simulation of acoustic wave propagation within these models and the synthesis of speech, typically limited to sets of vowel sounds. The level of physiological and geometric detail in these 3-D models, and the requirements in terms of accurate spatial sampling for the resulting simulations, implies that real-time synthesis is not usually possible. However, although the vocal sounds produced can sound very natural, the requirement for high-level user control of the 3-D tract shape also causes problems, meaning that articulatory voice synthesis based on these methods is still a non-trivial problem. Hence there is still interest in real-time vocal tract models based on 1-D and 2-D simulation methods using more established techniques such as the Kelly-Lochbaum transmission line [6], or digital waveguide [7], approach, particularly as their reduced complexity offers additional scope for high-level control. In particular, recent work has explored how 2-D models of the vocal tract might be optimised to give results closer to that of more complex, and computationally expensive 3-D simulations [8].

In [9] real-time dynamic articulation of a 2-D profile of the vocal tract was enabled through the use of a dynamically varying digital waveguide mesh based on impedance contour maps. Articulatory control of this model was later offered via the APEX system, a tool that can be used to synthesize sound and generate articulatory voice related parameters, based on the positioning of lips, tongue tip, tongue body, jaw opening and larynx height, all mapped from X-ray data [1]. In this case, APEX generates vocal tract cross-sectional area function information that provides control input to the 2-D dynamic waveguide mesh model [10].

This paper revisits the potential offered by the APEX system for enabling real-time articulatory control of a physical model of the vocal tract for sound synthesis. Whereas APEX was originally written for the legacy Windows operating system, it has now been realised using the SuperCollider environment, a flexible dynamic programming language for real-time audio synthesis [11]. Section 2 of this paper outlines the APEX system and how it can be used to generate vocal tract cross-sectional area information suitable for supplying input parameters to a physical model. Section 3 revisits the 1-D Kelly-Lochbaum digital waveguide model of the vocal tract. Section 4 introduces the new APEX SuperCollider implementation with Section 5 verifying the results produced by the vocal tract model and demonstrating the potential offered by the wider system. Section 6 concludes the paper looking at future directions for this work.

2. THE APEX SYSTEM

The original APEX project was designed to map from tract articulation to formant values with audio synthesis of the results obtained. The APEX model takes sampled fixed position data for certain vocal tract articulations and then derives intermediate values from physiologically motivated interpolation rules [1]. The vocal tract geometry is derived from a large number of X-ray images from a single subject, resulting in sampled two-dimensional contour points for the shape of the mandible (or lower jaw), the hard palate, the posterior pharyngeal (back) wall and the larynx, combined with the articulators - the tongue body, tongue tip, and lips. From this an articulatory profile is constructed and a semi-polar coordinate system defined with respect to specific points along the vocal tract - this enables an artificial mid-line along the length of the tract to be defined, equidistant between the front and back walls. The mid-sagittal distances along this vocal tract mid-line are then calculated between glottis and lips and from these values cross-sectional area functions can be calculated. It is these cross-sectional area values that then enable sound synthesis, and in this work the synthesis engine is based on a 1-D digital waveguide model.

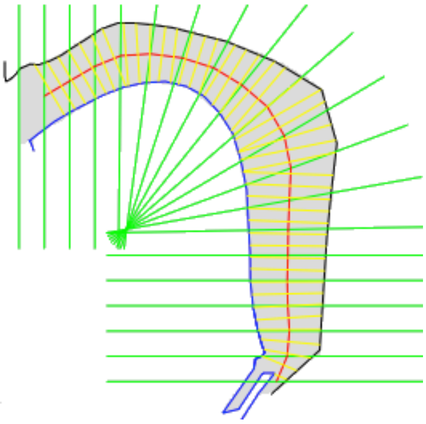


Figure 1: *The APEX vocal tract profile. The back wall of the tract is fixed with the front line depending on the configuration of the vocal tract articulators. The mid-line is first constructed using a semi-polar coordinate system and from this the cross-sectional area functions can be calculated.*

2.1. Voval Tract Cross-Sectional Area Definiton

Figure 1 illustrates this process of defining the cross-sectional area functions. The semi-polar coordinate system is represented by the long lines intersecting the vocal tract profile from an origin external to it. This is used to define the representative mid-line of the tract. New lines are constructed perpendicular to this mid-line so that the distance between the upper (fixed back wall) and lower contours of the vocal tract can be measured. These distances are mapped to equivalent cross-sectional area functions as follows:

$$A(x) = a{\cdot}d^b \tag{1}$$

Where $A(x)$ is the cross-sectional area function associated with a point along the mid-line x, d is the measured distance between the upper and lower contours of the vocal tract, and a and b are constants varying with vocal tract location and individual speaker characteristics.

2.2. Vocal Tract Articulation

In terms of controlling the vocal tract, some parts can be considered fixed: the posterior pharyngeal wall, the hard palate and upper jaw, as represented by the dark, black line in Figure 1. This data has been extracted from APEX X-ray data from a single individual. The moveable parts of the vocal tract considered in this model are detailed as follows.

The *larynx* shape has been extracted from APEX data and is fixed, although the vertical position can be changed. The larynx height (vertical position) varies only a little during speech, with females typically having a higher larynx position, and children even higher. The shorter vocal tract that results causes the formant frequencies to be scaled up relative to the male voice.

The *tongue body* is the most deformable part of the vocal tract and control parameters have been determined by principal component analysis (PCA). In [12] 394 tongue contours were obtained from X-ray images and each sampled at 25 approximately equidistant points. PCA results in a linear, weighted combination of a smaller number of basis functions as follows:

$$S_{target}(x) = S_{neutral}(x) + \sum_{i=1}^{N} c_i(v) PC_i(x) \tag{2}$$

Where x is the index of the point on the sampled tongue contour, $S_{target}(x)$ is the calculated tongue shape, $S_{neutral}(x)$ is a 'neutral' tongue shape, being the average of the measured shapes, and $PC_i(x)$ is the i^{th} basis function. The coefficient c_i is the weighting for the i^{th} basis function, being a 2-dimensional vector value which depends on the vowel, used as a parameter for calculating the tongue shape. From [12], setting $N = 1$ in (2) is found to account for 85.7% of the variance, while setting $N = 2$ achieves. 96.3%. In this work $N = 3$, giving three 2-D control parameters for tongue shape.

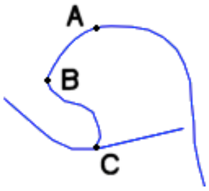


Figure 2: *The tongue tip geometry with the three required points labelled. A is defined as the end of the tongue body, B is the tongue tip and C is the point where the underside of the tongue is fixed to the lower jaw.*

The *tongue tip* is defined according to three points as shown in Figure 2. The tongue body ends at point A on the upper surface of the tongue, point B is the tongue tip and point C is where the underside of the tongue blade is connected to the mouth floor in the lower jaw coordinate system. Hermite interpolation is used between points A and B, with measured data used to form the curve between B and C based on the matched values at these points.

Jaw movement is the translation and then the rotation of the lower jaw coordinate system. The result is that some part of the frontal vocal tract is moved, including both the tongue body and blade, and the mouth floor and teeth (if they were included in this model). The angle for the rotation is determined by:

$$\theta = \frac{J}{2} + 7 \tag{3}$$

Where θ is the angle in degrees, and J is the jaw opening or distance between the upper and lower teeth (in mm).

3. THE 1-D VOCAL TRACT MODEL

The acoustic properties of the vocal tract can be modelled by considering it to be, at its simplest level, a straight tube from glottis to the lips where acoustic wave propagation is predominantly determined by the 1-D lossless wave equation:

$$\frac{\partial^2 p(x,t)}{\partial t^2} = c^2 \frac{\partial^2 p(x,t)}{\partial x^2} \tag{4}$$

where p is acoustic pressure, t is time, c is the speed of sound and x describes the (1-D) coordinate system. The Kelly-Lochbaum 1-D model, equivalent to a 1-D digital waveguide, provides a numerical solution to (4) and is well documented in the literature (e.g. [3] [6] [7] [9]).

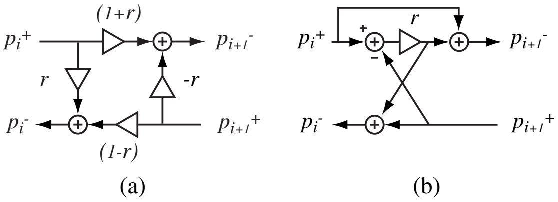


Figure 3: *Kelly-Lochbaum scattering of pressure signals in (a) the standard two-port scattering junction and (b) the one-multiply equivalent*

Summarising, the air column within the tube under consideration is discretized and represented as a series of connected bidirectional 1-D digital waveguide elements. The varied shape of the vocal tract along its length, as quantified by the cross-sectional area function, gives rise to the tract being defined as a series of concatenated acoustic tube elements, varying from glottis to lips. The change in cross-sectional area between tube elements is in turn modelled as a change in acoustic impedance for each waveguide element relative to the corresponding cylindrical tube section. These conditions give rise to the well known Kelly-Lochbaum scattering junction as shown in Figure 3(a) and expressed in the following equations:

$$r = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} \tag{5}$$

$$\begin{aligned} p_i^- &= rp_i^+ + (1-r)p_{i+1}^+ \\ p_{i+1}^- &= (1+r)p_i^+ - rp_{i+1}^+ \end{aligned} \tag{6}$$

Where r is the reflection coefficient between tube sections i and $i+1$ of cross-sectional area A_i and A_{i+1} respectively, p_i^+ is the incoming pressure value to the scattering junction from tube section i, p_{i+1}^+ is the outgoing pressure travelling in the same direction to tube section $i+1$. p_{i+1}^- is the incoming pressure value to the scattering junction from tube section $i+1$, with p_i^- the outgoing pressure value travelling in the same direction to tube section i. As shown in Figure 3(b) this can be expressed more efficiently for computational implementation as:

$$\begin{aligned} p_i^- &= p_{i+1}^+ + w \\ p_{i+1}^- &= p_i^+ + w \end{aligned} \tag{7}$$

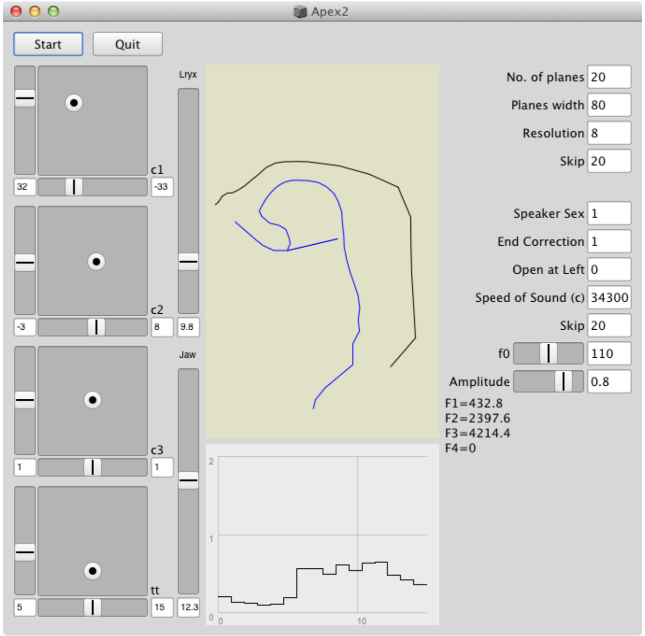
Where:

$$w = r(p_i^+ - p_{i+1}^+) \tag{8}$$

The model is completed with terminations accounting for lip and glottal reflection at each end. The lip end may also be terminated with an appropriate filter to approximate lip radiation effects. With the introduction of appropriate glottal excitation at the glottal end, signal output at the lip end results in speech sounds being produced.

4. APEX SUPERCOLLIDER IMPLEMENTATION

APEX vocal tract articulation control and a synthesis engine based on the 1-D Kelly-Lochbaum digital waveguide have been implemented in the real-time SuperCollider audio processing environment (v3.6). The server-client architecture of SuperCollider results in the APEX graphical user interface (GUI) running on the client with the data-processing and sound synthesis being handled separately on the server. The APEX GUI is shown in Figure 4 and is divided into three panels: the left hand side vocal tract articulation controls, the centre panel vocal tract graphical representation, and the right hand panel system parameter and synthesis information.



32 | -33
-3 | 8 | 9.8
1 | 1
5 | 15 | 12.3

No. of planes 20
Planes width 80
Resolution 8
Skip 20

Speaker Sex 1
End Correction 1
Open at Left 0
Speed of Sound (c) 34300
Skip 20
f0 110
Amplitude 0.8
F1=432.8
F2=2397.6
F3=4214.4
F4=0

Figure 4: *The APEX SuperCollider implementation graphical user interface. The left hand side are the vocal tract articulation controls, the centre panels give a real-time graphical representation of the articulation and resulting cross-sectional area values. The right hand panel allows the user to set parameters relevant to the spatial sampling of the tract shape and the synthesis engine.*

4.1. APEX User Control

The left hand side of Figure 4 shows six vocal tract articulation controls. The initial state of the tract in terms of the tongue body neutral shape and articulator states are loaded from data files at startup. The top three 2-D panel controllers are mapped to the three Principal Component variables derived from (2). The bottom 2-D panel is mapped to the Tongue Tip control (in mm), with the two vertical sliders used to control larynx height and jaw opening, both in mm. Upon user interaction, the parameters from these six articulation controls are passed via the client to a server based unit generator, or UGen, that is used to calculate the cross-sectional area function data based on the shape of the current articulation. The updated articulation data is then returned to the client to enable the centre panel graphical representation of the vocal tract to be updated, providing valuable feedback in terms of user control. It is also possible to view the relative changes in cross-sectional area along the vocal tract itself in the lower of the two centre panels.

The top of the right hand panel displays parameters relevant to the system configuration. The 'No. of Planes' parameter is used to set the number of planes in the semi-polar coordinate system that determines the vocal tract mid-line, with 'Planes Width' setting their width (or length) to ensure intersection with the back wall of the tract. The key parameter in terms of the resulting synthesis is 'Resolution', as this sets the spatial sampling interval (in mm) along the mid-line and hence will directly impact the number of cross-sectional area functions calculated.

One of the overheads with this SuperCollider implementation is the communication between server and client and ensuring that synchronisation is maintained between the two. The processing functions of the UGens are called at fixed time intervals (for the control rate UGens used in APEX the default is the audio sample rate divided by 64). The 'Skip' parameter tells the area function calculator UGen to skip a certain number of these calls between calculations. The minimum number for this parameter therefore depends on the speed of the computer being used and any additional load due to the synthesis engine.

The lower parameters relate to a prototype synthesis engine that calculates formant values based on the cross-sectional area values, passing them to a simple oscillator based formant synthesizer. Formants are calculated based on volume velocity transfer through the vocal tract [13], and although this synthesis method is superseded by the 1-D digital waveguide model discussed here, the calculated formant values displayed act as a useful additional check on the synthesized output. Finally the bottom two horizontal sliders allow run-time control of fundamental frequency and amplitude of the output.

4.2. APEX Vocal Tract Synthesis

The 1-D digital waveguide vocal tract model is also implemented as a server-side UGen, based on a specific implementation of the more general SuperCollider *NTube* synthesizer from Nick Collins' SLUGens plugins distribution [14]. The *KLVocalTract* UGen takes three parameters, the input excitation to the system, an array of $N-1$ reflection coefficients based on N cross-sectional area values (5) and the total delay line length in seconds. This is dependent on the sample rate of the system and the spatial sampling 'Resolution' of the vocal tract. Each cross-sectional area tube length, of which there are N, must have a delay of greater than two samples. Losses are set at the lip and glottis end based on reflection values from [7], such that $r_{glottis} = 0.969$ and $r_{lip} = -0.9$. The one-multiply equivalent scattering junction, as given in (7) and (8), is used for efficiency in implementation.

The excitation function, as shown in Figure 3, is implemented as a SuperCollider client-side, time-reversed glottal flow model, in effect the typical Liljencrants-Fant (LF) glottal flow model without a return phase. For output as sound pressure a +6dB/oct radiation characteristic is added. A single cycle of the excitation, with a fundamental frequency of 110Hz is shown in Figure 3(a) with the resulting spectrum in Figure 3(b).

5. RESULTS

The goal at this stage in the development of the SuperCollider APEX tool is to demonstrate the correct implementation of the 1-D digital waveguide model as part of the overal APEX system, and to highlight its capability for real-time dynamic articulation and voice synthesis.

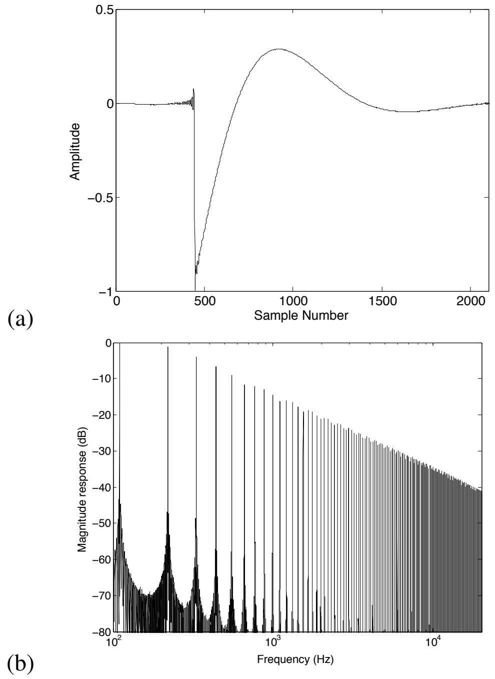


Figure 5: *The 1-D digital waveguide vocal tract excitation model as implemented in SuperCollider, in this case with a fundamental frequency of 110Hz. (a) The result is a time-reversed approximation to the LF glottal flow model; (b) the resulting spectrum.*

5.1. 1-D Vocal Tract Synthesis

Initially the *KLVocalTract* UGen is tested externally to the APEX front end in order to demonstrate its capability for synthesizing voice-like sound. With end termination reflection coefficient values defined as in Section 4.2 and a uniform cross-sectional area function set along the length of the tract, such that there is no internal scattering, should result in quarter-wave resonator like behaviour. In order to test this, *KLVocalTract* is instantiated with 44 equal area function sections, the tract length $L = 0.176$m, speed of sound $c = 343\text{ms}^{-1}$ and the sample rate is set to 192kHz. A white noise source is used as excitation at the glottis end and a 5s output signal is recorded at the lip end. The modal peaks for quarter-wave resonator behaviour can be predicted using:

$$f_{QWR} = \frac{(2m+1)c}{4L} : m \in \mathbb{N}^0 \tag{9}$$

These analytical values are overlaid on the Welch power spectral density of the output in Figure 6 and can be seen to give good agreement to the measurement obtained.

The next stage is to examine the output from *KLVocalTract* for a static cross-sectional area profile that relates more directly to speech-like sound. In this case *KLVocalTract* is instantiated with 44 area function sections under the same conditions as before, but this time based on data from [15] for the /a/ (*Bard*) vowel. As well as white noise excitation, glottal source excitation is also used and 5s output signals are recorded from the lip end. The results

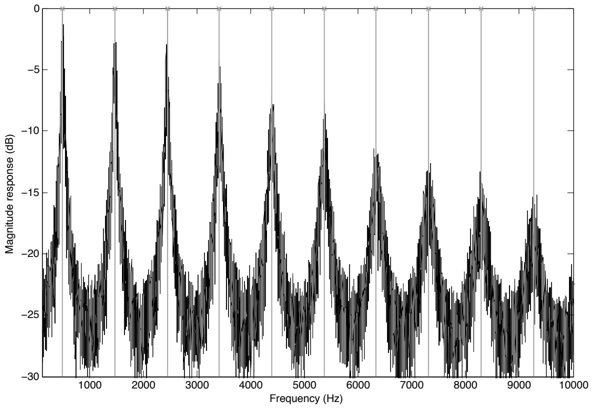


Figure 6: *Welch power spectral density estimate for 5s output from KLVocalTract functioning as a quarter wave resonator with white noise excitation. The grey vertical lines represent the analytical modal peaks for a quarter wave resonator of the same dimensions.*

obtained are compared with average measured formant values for the /a/ vowel based on data from [16] and the results are presented in Figure 7.

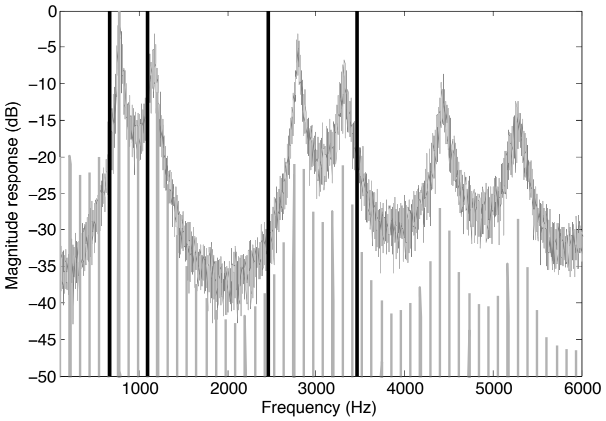


Figure 7: *KLVocalTract synthesis of 5s sound output for the /a/ vowel: Welch power spectral density estimate based on noise (solid dark grey line) and glottal source excitation (light grey lines) compared with average measured values of first four formants (black vertical lines).*

The results give good general agreement, and are comparable with those presented in e.g. [7], [9] which include vowel synthesis/formant values based on a 1-D digital waveguide model (amongst others).

5.2. APEX Real-time Articulatory Synthesis

In this example, the APEX SuperCollider GUI is used as a means to control the *KLVocalTract* UGen, with the dimension and number of the cross-sectional area vocal tract tube sections passed as the input to the synthesizer, together with fundamental frequency values to enable the glottal source excitation to vary pitch accordingly. To test the capabilities of the complete system, pre-calculated APEX data is passed into the GUI, giving the (x, y) values for the three PCA parameters plus the tongue tip, together with fundamental frequency, jaw opening and larynx height positions to enable the scale octave for 'DoReMe' to be articulated. Each note has a start and end set of values over a period of 0.9s, with a transition of 0.1s to the next note. Values are interpolated between states by the APEX system, and the 8s sound output is recorded at the lip end. Two different APEX states are used to obtain the results. In both cases 35 planes are used to calculate the vocal-tract mid-line (the maximum currently allowed by the system), but in the first example a spatial sampling resolution of 4mm is used, and in the second this is set to 5mm, for a sampling rate of 96kHz. For the 4mm case this results in 33-35 vocal tract tube sections, and for the 5mm case 25-27 sections. The number of sections varies slightly as the articulations can actually change the length of the vocal tract model, and in this current implementation, every articulation results in a complete recalculation of the mid-line and cross-sectional area functions, and hence the potential for the total delay-line length to change accordingly in *KLVocalTract*.

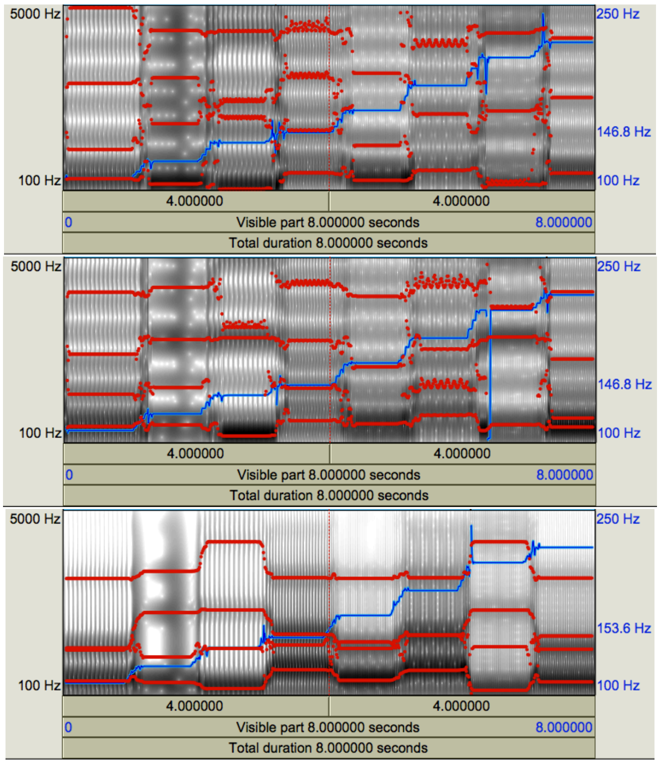

5000 Hz
250 Hz
146.8 Hz
100 Hz
100 Hz
4.000000
4.000000
0
Visible part 8.000000 seconds
8.000000
Total duration 8.000000 seconds
5000 Hz
250 Hz
146.8 Hz
100 Hz
100 Hz
4.000000
4.000000
0
Visible part 8.000000 seconds
8.000000
Total duration 8.000000 seconds
5000 Hz
250 Hz
153.6 Hz
100 Hz
100 Hz
4.000000
4.000000
0
Visible part 8.000000 seconds
8.000000
Total duration 8.000000 seconds

Figure 8: *Praat spectrograms with candidate fundamental frequency and formant values identified and tracked over the 8s of output for the octave scale of 'DoReMe'. The fundamental frequency line corresponds to the right hand scale and is overlaid on the broader formant value calculation spectrogram plot. The top two examples use KLVocalTract with a resolution of 4mm (top) and 5mm (middle). The bottom plot is based on the method presented in [13] and formant synthesis.*

In addition to using *KLVocalTract* and to enable some comparison, the same phrase is synthesized based on the method presented in [13] using a band limited saw-tooth oscillator (SuperCollider *Saw* UGen), and up to four resonant low-pass filters centred on the calculated formant frequencies.

The initial fundamental frequency value in each case is 110Hz. These three sets of results are analysed using Praat, and spectrograms produced with automatic tracking of fundamental frequency and formant frequency values enabled (based on default Praat settings). The results are shown in Figure 8 and corresponding audio examples, resampled to 44.1kHz at 16bit resolution are available for audition at `http://www-users.york.ac.uk/~dtm3/vocaltract.html`.

Note that the first and last notes of the scale have the same articulation values, and so should give the same results apart from an octave shift in fundamental frequency. This is clearly the case in the format synthesis example, although it seems that an additional candidate value has also been found. Interestingly, based on the same default analysis values, the 5mm resolution example gives a closer match. Similarly, the lower resolution case in most cases results in a more consistent tracking of identified formats although both demonstrate similarities and differences with the formant synthesis example. Across transitions it is evident that all examples demonstrate discontinuities due to the interpolations applied between articulation states, especially for fundamental frequency. It is notable that the formant synthesis example demonstrates clearer and smoother transitions although this is expected due to the relative simplicity of the synthesis algorithm used and the fact that formants are synthesized directly (from specific filter settings), rather than indirectly as a consequence of the physical modelling algorithm used.

However, the higher resolution *KLVocalTract* does, for some notes, exhibit better, more consistent formant values (Note 3, Note 5, Note 6, Note 7, when compared with the lower plot in Figure 8)) and ultimately results in a better quality sound output, even if formant tracking transition states are a lot less clear than the lower resolution example. This poor formant tracking at high resolution is due to the nature of the cross-sectional area functions calculated by the APEX controller and how they are applied in *KLVocalTract*. Lower resolution implies fewer tube sections and so fewer (or smaller) changes (and resulting interpolations) to be applied to *KLVocalTract* between articulation states. This is particularly the case when the tract length itself changes due to an applied articulation (e.g. larynx height) as the number of tube sections will also change, resulting in discontinuities in the output of *KLVocalTract* as the delay line length will also have to be updated. This is further compounded if, at the lip end, the jaw opening parameter is not mapped directly to the last tube section that also corresponds to where the output is tapped off. This can result in restricted sound output, and mistuning of expected formant positions, as is evidenced in some of the cases in Figure 8. Finally, it should be noted that this example does apply significant, multi-parameter changes in articulation states with the associated required interpolation of these values. Direct user control of the GUI, where fewer parameters are changed in real-time with smaller degrees of variation, results in much smoother and continuous examples of articulatory synthesis.

6. CONCLUSIONS

This paper has presented a SuperCollider implementation of the APEX articulatory vocal tract algorithm, coupled with the *KLVocalTract* 1-D digital waveguide synthesis engine. The latter has demonstrated its potential for accurate, real-time formant synthesis, based as it is on established methods. The APEX GUI reduces the number of variables required for articulatory speech synthesis to a much more manageable number, and SuperCollider provides an excellent real-time framework for linking these two components while providing additional access to standard sound synthesis unit generator processes and user control options. This first investigation into the potential for such articulatory vocal tract synthesis in SuperCollider has revealed that for realistic speech-like sound output the calculation of cross-sectional tract profiles must be done more efficiently, with direct mapping to the tube sections implemented in the synthesis engine, such that the underlying delay-line length does not have to change. This would enable smoother control of *KLVocalTract* while also enabling its eventual replacement with a modelling engine of higher dimension.

In addition, the articulatory controls require some further refinement. The PC variables need to be constrained such that physically impossible tract profiles are similarly not allowed - for instance the tongue tip should not be able to move through the back wall of the tract. The lips also need to be more completely represented, the tongue body should maintain a constant volume, and in this current implementation, branches on the tract, such as the subapical space and velar port/nasal tract are not represented in either the controller or the underlying model.

A version of the source code for this work is also available at `http://www-users.york.ac.uk/~dtm3/vocaltract.html` such that it might be adopted by the wider SuperCollider community.

7. ACKNOWLEDGEMENTS

We are indebted to Björn Lindblom and Christine Ericsdotter, both at Stockholm University, and to Johan Sundberg at KTH, for generously sharing their expertise and data, painstakingly acquired for the original APEX Windows 3.1 implementation, which was written by Johan Stark. The groundwork for the SuperCollider implementation of APEX was done in 2011 in the M.Sc. thesis of author Jani. The extended version presented here was developed in 2014-2015 by authors Murphy and Ternström. We are grateful also to Gerhard Eckel and Ludvig Elblaus for advice and assistance with SuperCollider. The work was supported in part by the Swedish Research Council (VR), contract D-2013-0378.

8. REFERENCES

[1] J. Stark, C. Ericsdotter, P. Branderud, J. Sundberg, H-J. Lundberg, and J. Lander, "The APEX model as a tool in the specification of speaker specific articulatory behaviour," in *Proc. XIVth ICPhS*, San Francisco, USA, 1999.

[2] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *J. Acoust. Soc. Am*, vol. 128, no. 6, pp. 3724–3738, Dec. 2010.

[3] M. Speed, D. T. Murphy, and D. M. Howard, "Modeling the Vocal Tract Transfer Function Using a 3D Digital Waveguide Mesh," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 2, pp. 453–464, Feb. 2014.

[4] B. Delvaux and D. M. Howard, "A New Method to Explore the Spectral Impact of the Piriform Fossae on the Singing Voice: Benchmarking Using MRI-Based 3D-Printed Vocal Tracts," *PLoS ONE*, vol. 9, no. 7, pp. e102680–15, July 2014.

[5] T. Vampola, J. Horáček, and J. G. Švec, "Modeling the Influence of Piriform Sinuses and Valleculae on the Vocal Tract Resonances and Antiresonances," *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 594–602, May 2015.

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

[7] J. Mullen, D. M. Howard, and D. T. Murphy, "Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 964–971, May 2006.

[8] M. Arnela and O. Guasch, "Two-dimensional vocal tracts with three-dimensional behavior in the numerical generation of vowels.," *J. Acoust. Soc. Am*, vol. 135, no. 1, pp. 369–379, Dec. 2013.

[9] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 577–585, 2007.

[10] D. T. Murphy, S. Shelley, and S. Ternström, "The dynamically varying digital waveguide mesh," in *Proc. 19th Int. Congr. Acoustics*, Madrid, Spain, Sept. 2007.

[11] J. McCartney et al., "SuperCollider," Available at http://supercollider.github.io/, accessed May 15, 2015.

[12] B. Lindblom, "A numerical model of coarticulation based on a principal components analysis of tongue shapes," in *Proc. 15th Int. Congr. Phonetic Sciences*, Barlcelona, Spain, 2003.

[13] J. Liljencrants and G. Fant, "Computer program for vt-resonance frequency calculations," in *Dept. for Speech, Music and Hearing, Quarterly Progress and Status Report STL-QPSR 16(4)*, KTH, Stockholm, 1975, pp. 15–02.

[14] N. Collins, "SC3 plugin contributions," Available at http://composerprogrammer.com/code.html, accessed May 15, 2015.

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

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