Introduction and Background [250]

There are many different ways to synthesise sound, from adding sine tones together (additive synthesis) to using oscillators as ways to modulate the sound (FM synthesis). However, these methods have a few issues, namely that sound synthetic and lack the warm related to acoustic instruments and they’re harder to control [1].

Physical modelling is different from these ways of synthesis, instead of focusing on sine waves, it models the system based on real-world equations. It produces a more natural sound, and allows the user to use more meaningful ways to define it. The method of physical modelling synthesis that will be used is digital waveguide (DWG) synthesis. This works by modelling wave propagations found in tubes or strings using multiple delay lines.

Physical modelling definition

Use of voice synth in real world

Voice synthesis is different from other types of synthesis, the voice is a very rich and expressive sound source, so this adds more challenges. Due to the rich variety of the voice, there are numerous ways to synthesise the voice, most models start by looking at the vocal tract as a 17cm tube with the lips and mouth at one end and the glottis at the other (Figure 1).

The vocal tract can be represented using the source-filter model where be glottis is the source, and the larynx and lips are the filters changing the way the sound is produced. Using the Kelly-Lochbaum vocal tract model we can split it up into multiple sections, each section has a specific area which changes relating to the vowel being produced (Figure 2). We can then model the vocal tract as a series of standard digital waveguides. This idea was used in collaboration with Max Matthews in the first synthesis of singing voice on an IBM computer, it is now possible to use AI to create custom voices using software such as resemble.ai [2]

*Figure 1 Vocal Tract [3]*

*Figure 2 Area of Vocal Tract, from the Glottis end [4]*

Code Deign

The code is based on the digital waveguide vocal model and the work done in Labs. It is based of the work by Kelly Lochbaum, and implements the vocal tract as a series of cylindrical tubes with differing impedances, at the intersections between each tube it uses a Kelly Lochbaum scattering junction [5]. Figure 3 shows the 1D model and details how the area components (A) relate to the impedance (Z) with the scattering junctions (J) being between each section.

The code produces diphthongs, monophthongs, plosives as well as allowing the use of noise, LF and LF with Vibrato as the excitation. Plosives can be inserted at any pint during the vowel production, and diphthongs can be produced between any two monophthongs. This can produce some diphthongs that would not naturally be produced by the human voice. The table below highlights all the monophthongs (vowels) the system can produce, the other table highlights diphthongs the system can produce with the two variables needed to make it sound.

Clicking on the IPA representation will link to the examples further down the page.

User Inputs

| **IPA** | **Code Variable** | **Description** |
| --- | --- | --- |
| **ɑ** | a | bart/father |
| **æ** | ae | bat/lad |
| **ε** | bird | bird |
| **i** | I | beet/see |
| **o** | O | ball/law |
| **ɵ** | Q | bod/not |
| **u** | u | food/soon |
| **ʊ** | U | foot/put |
| **ʌ** | V | but |

Development

When developing the model, the place to start would be looking at the excitement of the system. The system allows Liljencrants-Fant (LF) glottal flow model to synthesise the voice, the system reads in the two LF inputs, one with vibrato and one without, provided in the lab and uses it as excitement for the system. In addition to the LF inputs, the system allows 1 second of noise to be used as an excitement, below are the experiments with different types of noise, unit impulse and LFs. The phoneme being synthesised is <b>a</b>.

After completing the lab and implementing the different excitement types, the next stage was to focus on was creating plosives. Plosives are as simple as stopping the sound midway through creating it. The code works by creating an envelope that goes from 0 to 1 over 1/8th of the total samples. This envelope can then be inserted at any point during the system with ones filling in the remaining space. The mouth end is then updated with the value of the envelope multiplied by the segment diameter.

The table has audio examples of plosives at different positions, the excitation type was “LF” and the vowel synthesised was &epsilon; (bird).

After creating plosives and excitation, the next feature to look at was creating diphthongs. These are created by transitioning from one monophthong to another mid-way through. The code allows the user to specify two monophthongs and transition between the two. This allows the user to create sounds that might not be normally produced.

* [1]"The Project - NESS", NESS, 2021. [Online]. Available: https://www.ness.music.ed.ac.uk/project.
* [2]"AI Generated Voices ~ Resemble AI", Resemble AI, 2021. [Online]. Available: https://www.resemble.ai/.
* [3] Loy, G., & Chowning, J. (2007). Sound Synthesis. In Musimathics, Volume 2: The Mathematical Foundations of Music (pp. 363-452). Cambridge, Massachusetts; London, England: The MIT Press. doi:10.2307/j.ctt5hhm8g.14
* [4] B. Story and I. Titze, "Parameterization of vocal tract area functions by empirical orthogonal modes", Journal of Phonetics, vol. 26, no. 3, pp. 223-260, 1998. Available: 10.1006/jpho.1998.0076
* [5] J. Mullen, D. Howard and D. Murphy, "Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 964-971, 2006. Available: 10.1109/tsa.2005.858052