Building a Vowel: a Bottom-up Guide for Phonetic and Language Learning Sciences

Axel Ekström[1,2,*], Runhui Song[2,3] and Jens Edlund[2]

[1]Centre for Cultural Evolution, Department of Psychology, Stockholm University, Stockholm, Sweden
[2]Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[3]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

Abstract

We present a guide for building a vowel "from the ground up" using software developed within the HumInfra infrastructure. Using this guide, a user may recreate in a simulation conditions for the production of a given vowel quality. It is summarized step-wise how a midsagittal view of a speaker's vocal tract may be reduced to a simplistic sequence of two-dimensional segments, which – input into the simulation software TubeN – predicts and recreates the given vowels quality. Our hands-on guide is of interest to students in the linguistic and teaching sciences, and to those teaching speech science in a broader sense.

Keywords

educational software, language learning, acoustics, phonetics

1. Introduction

Vowels arise from movements of articulators – including the jaw, lips, tongue, and velum. Such movements shape the acoustic signal in predictable ways by alternating the resultant spectral peaks, typically termed formants, generally held as the basis for vowel synthesis and perception [1]. The modern interpretation of this relationship stems from "source–filter theory" [1], which explains speech production as the interaction between a voice *source* (supplied through vocal fold oscillation) and a *filter* (the supralaryngeal vocal tract, beginning at the larynx and terminating and the mouth opening). While foundational in phonetic science and early speech synthesis, the framework has historically been studied within engineering disciplines [2, 3, 4, 5]. However, knowledge of speech production is not merely of use to those in speech sciences – but also to, for example, students of linguistics and language learning. Here, we present a step–by–step "ground-up" approach to constructing a simple, accessible model of vowel production, using attainable data and freely available software [6] developed within the *HumInfra* national infrastructure. This guide is intended for students in the broader humanities and educational sciences. While simplistic, the model yields predictable and replicable results, and it has already proven effective in workshops with phonetics students [7].

2. Building a vowel, step by step

2.1. Obtain vocal tract data

The goal of the present paper is a step-wise guide by which formant patterns observed in natural speech can be "reverse engineered" computationally, without an engineering background.[1] The first step toward such a model is data derived from articulation of the desired vowel – typically captured using magnetic resonance imaging (MRI) techniques. In Sweden, several universities now possess imaging equipment

HumInfra Conference 2025 (HiC 2025), Stockholm, 12–13 November 2025.
*Corresponding author.
axeleks@kth.se (A. Ekström)
0000-0002-6739-0838 (A. Ekström); 0000-0001-9327-9482 (J. Edlund)

[1]We do not describe nuances of formant estimation here, as sources toward this end are widely available elsewhere.

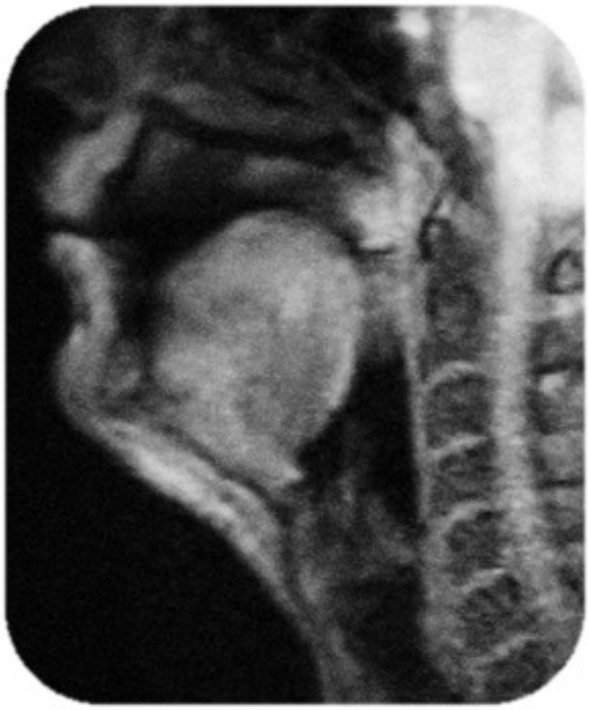

Figure 1: Adult male native Swedish speaker, aged 28, speaking [uː]. Data was collected during a magnetic resonance imaging (MRI) single-subject case study at the Stockholm University Brain Imaging Center. The scanner was a Siemens Prisma 3 Tesla whole-body MRI. Data collection used the "FLASH" sequence- a spoiled gradient echo with a flip angle of 5 degrees, an echo time of 1.7 ms, and a repetition time of 4 ms. The field of view was 280(freq) x 278(phase) mm, with a resolution of 224(freq) x 156(phase) and a bandwidth = 500 Hz / px.

necessary to record such data. These facilities include the Stockholm University Brain Imaging Center (SUBIC) and the Lund University Bioimaging Center. However, note that databases of such data are also available [8, 9]; as such, it is not necessary that new data be acquired. For the sake of illustration, we selected from data previously recorded for other research purposes, an occurrence at which the speaker (an adult male) produced long close back rounded vowel [uː] (Figure 1).

2.2. Segmentation

A midsagittal view of a speaker's vocal tract at the moment of vowel realization can be simplistically reduced to a sequence of smaller segments, defined by their length and area. For [uː] – the above stated test case – an initial *narrow* section (the rounded and protruded lips) is followed by an expansive *open* section (the anterior oral cavity, where the tongue has been retracted, leaving more open space), etc. To achieve useable segment lengths and areas, the vocal tract shape may be traced and isolated from its surrounding anatomy. A line may then be traced at the midpoint of the distance from one wall to the other (up-to-down, or left-to-right, depending on the position in the tract). By then tracing equidistant sections in the structure, segments defined by their *distance in the sagittal plane* may be derived for computational implementation. In theory, the lengths of such segments may be varied. However, both 1cm and 0.5cm segment lengths are commonplace in the literature.

2.3. Converting distances to areas

Articulatory-acoustic dimension reduction techniques developed over several decades [3, 1, 10] have illustrated that the vocal tract – a three-dimensional complex of biological elements – can be simplistically modeled as a two-dimensional sequence of segments defined by their length and area. Without direct

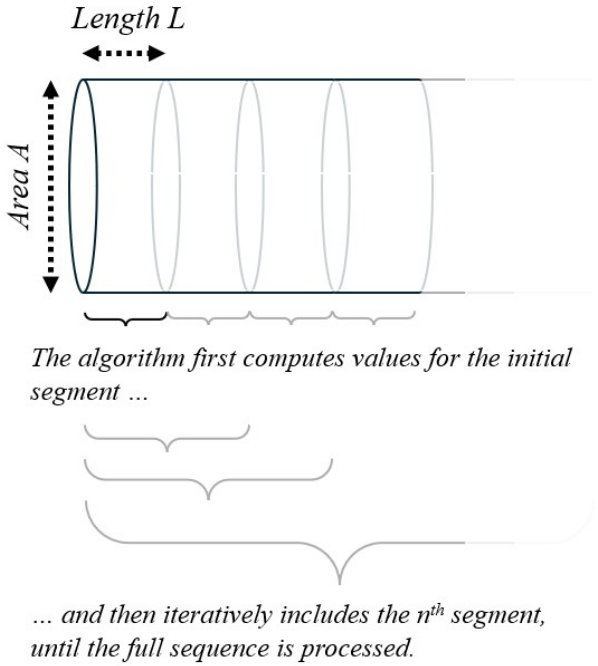

Figure 2: Workflow of the Liljencrants and Fant algorithm, as implemented in *TubeN*. The program makes no restrictions on the number or lengths of the total sequence.

access to three-dimensional vocal tract data – which requires more invasive, and more specialized methodology – it is necessary to convert cross-sectional relationships to area estimates for each segment. Fant [11] stipulates such an equation which is appropriate for the present purposes, according to which conversions from distance *d(x)* in the sagittal plane to cross-sectional area *A(x)* may be computed as

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

Specifically, however, because this relationship is not constant throughout the vocal tract, it is necessary to adapt the power function based on what part of the image is currently being processed.

For the lip section, Fant [11] stipulates values $a = 1.8$ and $b = 2.5$ where $d > 1.7$ cm, and $a = 1.8$ and $b = 2.5$ where $d < 1.7$ cm.

For the volume of the oral cavity, Fant [11] suggests including corrections for air columns on both sides of the tongue, which is retracted during sustained production of [u:]. As such, the user may implement $a = 2.4$ and $b = 1.4$, before adding a correction of an additional 35% to the final volume [11].

Finally, Fant [11] notes that the power function should be subdivided for application to the pharynx, as its midsagittal distance-to-volume relationships varies significantly. He suggests that where $d < 1.75$, $a = 2$ and $b = 1.6$; and where $1.75 < d < 2.5$, $a = 2.8$ and $b = 1$.

2.4. Computing the properties of an acoustic tube

The TubeN software [6] implements an algorithm developed by Liljencrants and Fant [12], which predicts, for any sequence of tube segments, the resultant spectral peaks (functionally equivalent to formants).[2] That is, if completed appropriately, formants predicted by TubeN for a shape corresponding to the speaker's vowel production, should closely match the vowel quality produced by that speaker at that moment. The TubeN software is publicly available and can be accessed online through the relevant GitHub repository.[3]

[2]For the sake of brevity, the mathematical representation of the program is not included here. Readers are directed to the original publication [12, 6].
[3]https://github.com/jbeskow/tuben

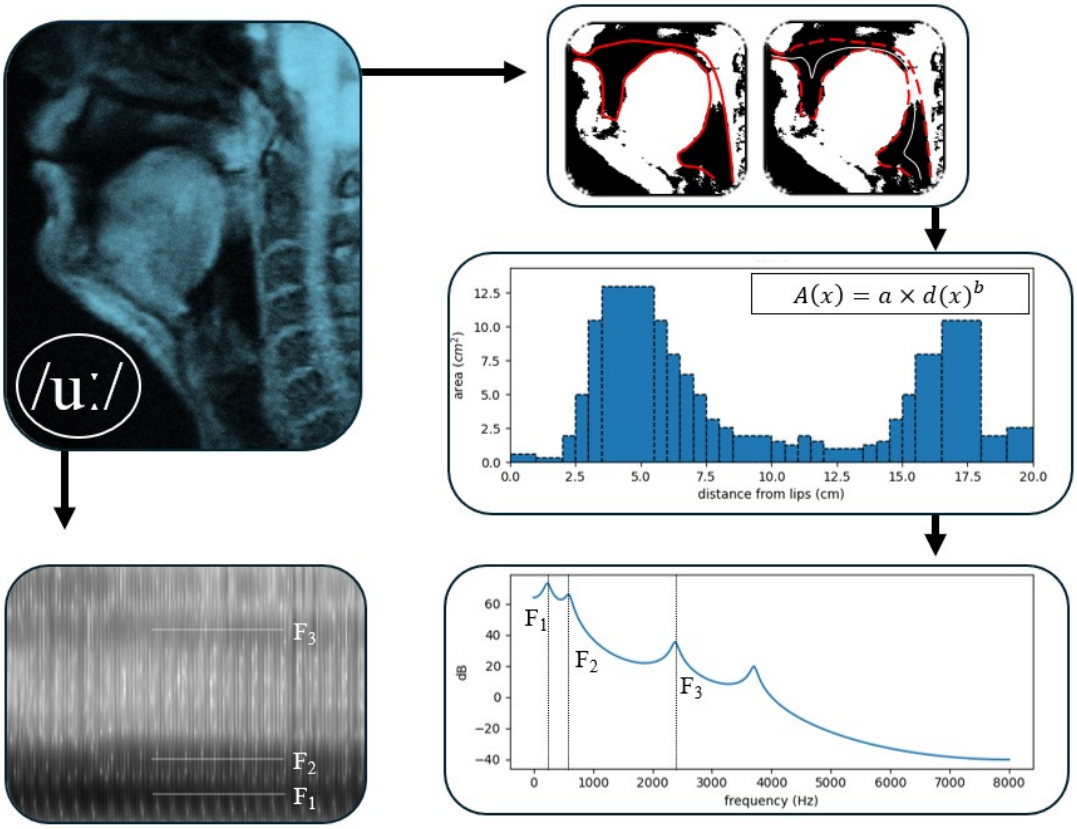

Figure 3: Note that segments are not annotated. In theory, segments can be of variable length and – provided relationships between them are mostly preserved – predicted vowel qualities will overlap approximately.

3. Concluding comments

In brief, the "vowel reverse engineering" exercise described above involves four steps. They are, in order, (1) obtain or collect midsagittal view of a speaker vocal tract mid vowel production; (2) trace the effective tract, from glottis to lips; (3) segment said outline into equidistant "slices" and (4) convert them into appropriate cross-sectional area estimates; (5) input lengths and areas of segments into the *TubeN* software, which automatically generates predicted resonant frequencies (or formants); and finally, (6) match with formants estimated during live production of the same vowel. We look forward to exploring the potential of this procedure as a teaching tool in broader phonetics and language pedagogical education, and are open to collaborations with any interested parties toward this end.

Acknowledgments

The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017–00626). We thank Gláucia Laís Salomão and Jonathan Berrebi at the Stockholm University Brain Imaging Center for their assistance in recording our MRI data. We extend additional thanks to Jonas Beskow for discussion about the program, and to the attendants of the Swedish Linguistics (SLING) meeting of 2023, and the Fonetik 2024 Conference, where the software and its application was extensively discussed.

References

[1] G. Fant, Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations, Mouton, 1971.

[2] B. E. Lindblom, J. E. Sundberg, Acoustical consequences of lip, tongue, jaw, and larynx movement, The Journal of the Acoustical Society of America 50 (1971) 1166–1179.
[3] H. K. Dunn, The calculation of vowel resonances, and an electrical vocal tract, The Journal of the Acoustical Society of America 22 (1950) 740–753.
[4] P. Badin, G. Fant, Notes on vocal tract computation, STL QPSR 2 (1984) 53–108.
[5] K. N. Stevens, S. Kasowski, C. G. M. Fant, An electrical analog of the vocal tract, The Journal of the Acoustical Society of America 25 (1953) 734–742.
[6] R. Song, J. Beskow, J. Edlund, M. Tronnier, R. Tu, K. Zhang, A. Ekström, Open source software for tube vocal tract modeling, resonance prediction, illustration, and 3D printing, BioRXiv (2025). doi:https://doi.org/10.1101/2025.10.15.682256.
[7] M. Tronnier, A. G. Ekström, Teaching speech acoustics through vocal tract modeling, in: G. Ambrazaitis, Raschellà, N. J. Young (Eds.), Proceedings from FONETIK 2025, Linnaeus University, 2025, pp. 83–84.
[8] T. Sorensen, Z. I. Skordilis, A. Toutios, Y. C. Kim, Y. Zhu, J. Kim, S. S. Narayanan, Database of volumetric and real-time vocal tract mri for speech science, in: Proceedings of Interspeech, 2017, pp. 645–649.
[9] Y. Lim, A. Toutios, Y. Bliesener, et al., A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images, Scientific Data 8 (2021) 187. doi:10.1038/s41597-021-00976-x.
[10] R. Carré, P. Divenyi, M. Mrayati, Speech: A dynamic process, Walter de Gruyter GmbH & Co KG, 2017.
[11] G. Fant, Vocal tract area functions of swedish vowels and a new three-parameter model, in: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 1992), ISCA, 1992, pp. 807–810. doi:10.21437/ICSLP.1992-262.
[12] J. Liljencrants, G. Fant, Computer program for vt-resonance frequency calculations, STL-QPSR 16 (1975) 15–21.