Introduction. For most people, speech is the primary means of conveying thoughts, intentions, and feelings to others. Speech is enabled by the rapid manipulation of many articulators in the vocal tract (Simonyan et al., 2016). These articulators give rise to various configurations of the tract, which determine a speech sound's resonances and anti-resonances. Speech sounds produced by the vocal tract are most often understood in terms of their frequency content, at least in part, because the perceptual system operates in the frequency domain. This is particularly true for vowels, which are generally defined by their first two resonances, or formant frequencies (Hermann, 1894).

Vowel production can be modeled with tubes, which have well-defined resonances and can capture the most salient spectral shaping performed by the vocal tract. These tubes can be seen as the 'filters' of source-filter theory in which the glottal sound source and the vocal tract acoustic filter are separable (Fant, 1960). In this view, speech sounds are the result of a frequency-domain multiplication of the sound source and vocal tract filter.

In this short experiment, our aims were twofold: 1) to demonstrate source-filter theory with a set of 3D printed models of the vocal tract and 2) to quantify the degree to which an ellipse tube model derived from a full 3D model of an /a/ vocal tract can capture the formant frequencies of the vowel.

Methods. Audio playback and recording was performed on a MacBook laptop with a JOUNIVO USB desktop microphone and KIXAR USB portable speaker. All signals were sampled at 22,050 Hz. A cone-shaped acoustic coupler was designed in the AutoDesk Fusion 360 to directly couple the output of the speaker to the vocal tracts, minimizing noise leakage and thereby maximizing SNR. Two 3D models of the vocal tract were printed from the Dresden Vocal Tract Dataset (Birkholz et al., 2020). The two models are shown in **Figure 1A.** /i/ and /a/ were selected to print from the dataset because they are maximally distinct in F1-F2 space. We then derived a tube model from the /a/ model which preserved the cross-sectional area of the vocal tract. We used AutoDesk Fusion 360 to 1) take a midsagital slice of the model, 2) estimate the midpoint of the tract at 20 locations along the tract, 3) and measure the lengths of the major and minor axes of the ellipse which encircling the cross section at each of the 20 locations (**Figure 1B**). We calculated the FFT (**Figure 2A**, grey) and subsequently used the upper envelope (peaks spaced 10 + Hz from each other) of the spectrum as the frequency-domain representation of the signal. The envelope, $C(\omega)$, is plotted in **Figure 2A** as a black line. To derive an inverse filter, we set low and high frequencies (0-10 Hz and 8000-11025 Hz) to 1. All intervening frequency magnitude values were taken to be $C^{-1}(\omega) = \frac{1}{C(\omega)}$. The vowel source was digitally synthesized in Python 3.7 using the *numpy* and *scipy* packages. We used both pulse-trains and glottal flow time series in our experiments, but only report on the white-noise experiments in results because we feel they demonstrate the source-filter theory most clearly.

Results. The cone transfer function included major resonances over a 40 dB range in the frequency regions of interest (**Figure 2A**). We observed peak— as measured by the FFT envelope (black) and LPC spectrum (orange)—at ~350, 2100, and 3500 Hz. The peaks were effectively removed by our inverse-filtering approach (**Figure 2B**). The 350 Hz resonance was intact but with a much smaller peak: 10+ dB prior to inverse-filtering, and ~4 dB after inverse filtering. The 2100 and 3500 Hz resonances were almost completely leveled, achieving the desired near-white-noise frequency spectrum of the glottal source (**Figure 2B**).

With this flat frequency spectrum as input, we measured the effect of placing the vocal tract filters on top of the cone (**Figure 3A, B**). As hypothesized, we observed different resonant frequencies for the /i/ model (**Figure 3A**) and /a/ model (**Figure 3B**) as these vocal tracts represent different filters which act on the source. For /i/, LPC analyses identified a very low pole/resonance at 50 Hz, which we took to be spurious—perhaps the result of an imperfect white-noise input spectrum. Beyond this, the spectrum aligns well with empirical measures of resonances from real speakers: the first resonance, F1, occurred at 500 and the second, F2, occurred at 1900 Hz. For /a/, we observed F1 and F2 approach one another, with resonances at 800 and 1100 Hz.

Finally, we compared the full /a/ model (**Figure 3B**) with its tube counterpart (**Figure 3C**). The tube model preserved the full model's F1, F2, and F3 remarkable well. Each of these formants (as identified by LPC) are within 100 Hz of each other. The two spectra diverge at F4; the tube model shifts F4 down by about 600 Hz.

Conclusion. Our results directly demonstrated that vocal tract configuration filters a sound source, thereby defining formant frequencies. Further, we showed that tube models can preserve formant frequencies, both in frequency location and magnitude. We also concluded that vocal tract hardness has little effect on the formants by comparing our models' resonances to those from real speakers; the hard corn starch resin walls of the printed vocal tracts did not attenuate or shift the resonances and may have even amplified them.

Figures & Data

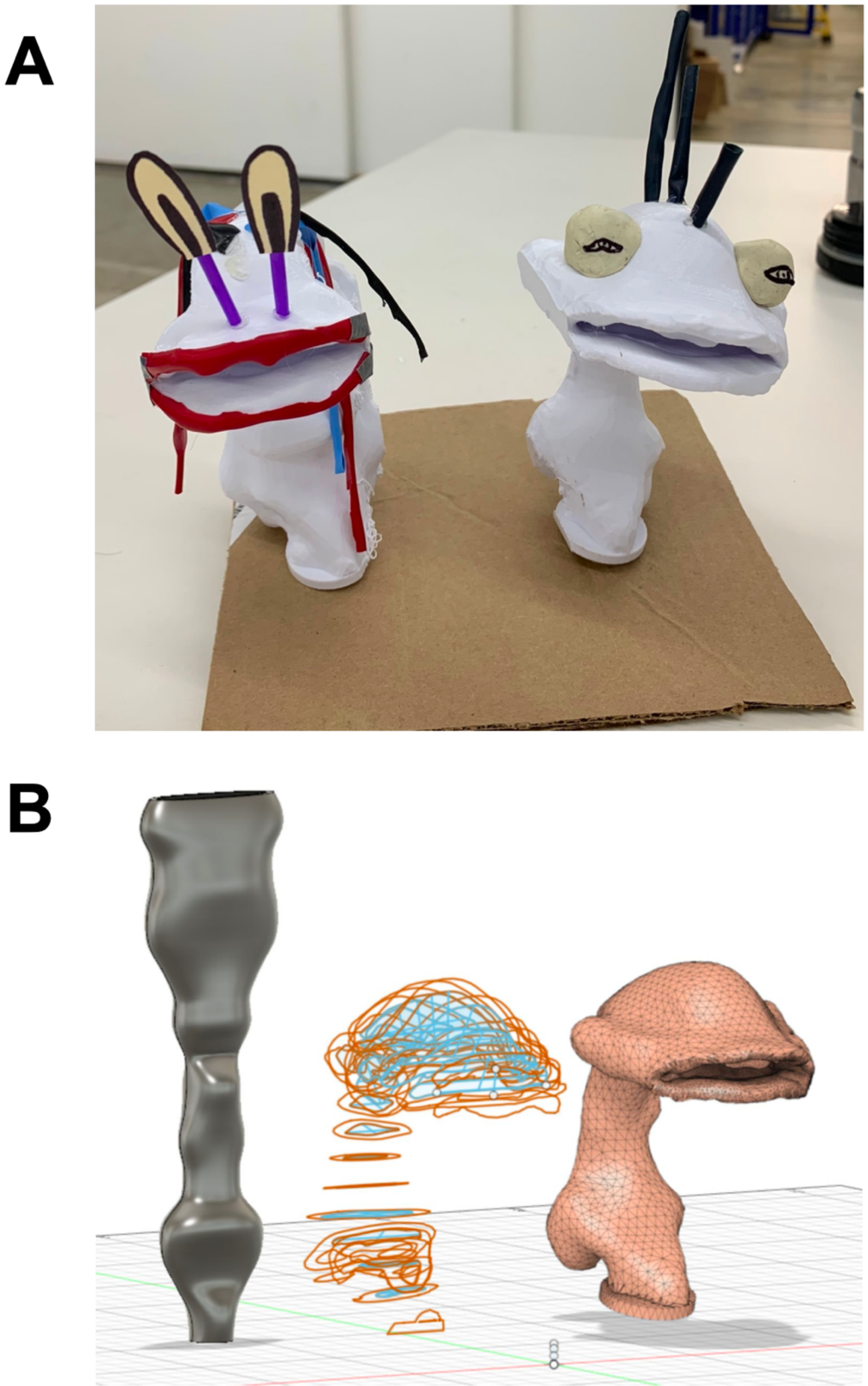

Figure 1: 3D renderings of the /a/ model (right, from the Dresden Vocal Tract Dataset), 20 slices of the model throughout the length of the vocal tract (middle), and the ellipse tube model derived from the sections (left). The tube model was created to closely follow the cross-sectional area at the 20 points along the vocal tract.

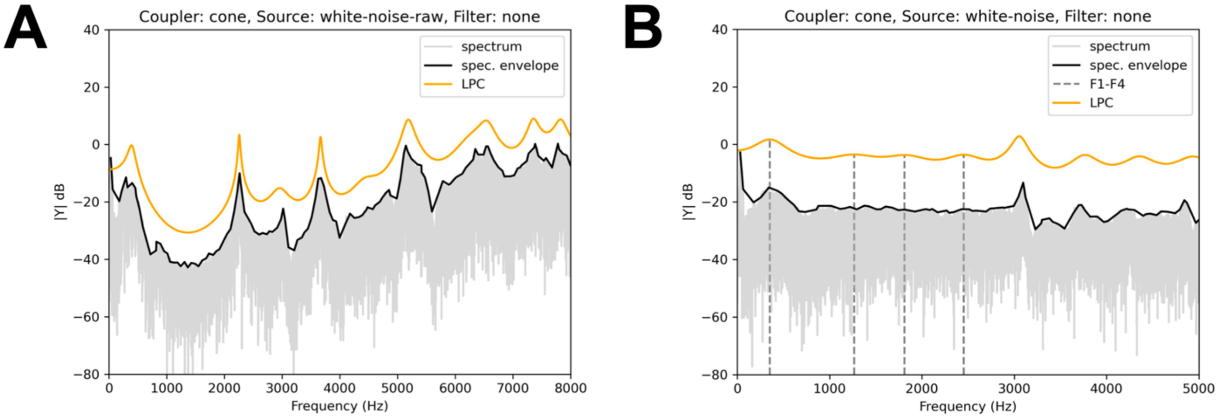

Figure 2: Frequency spectra and LPC-inferred transfer function of the speaker-to-vocal tract coupler prior to (A) and after (B) implementing the inverse filter. Both measurements were taken at the top of the cone without any vocal tract.

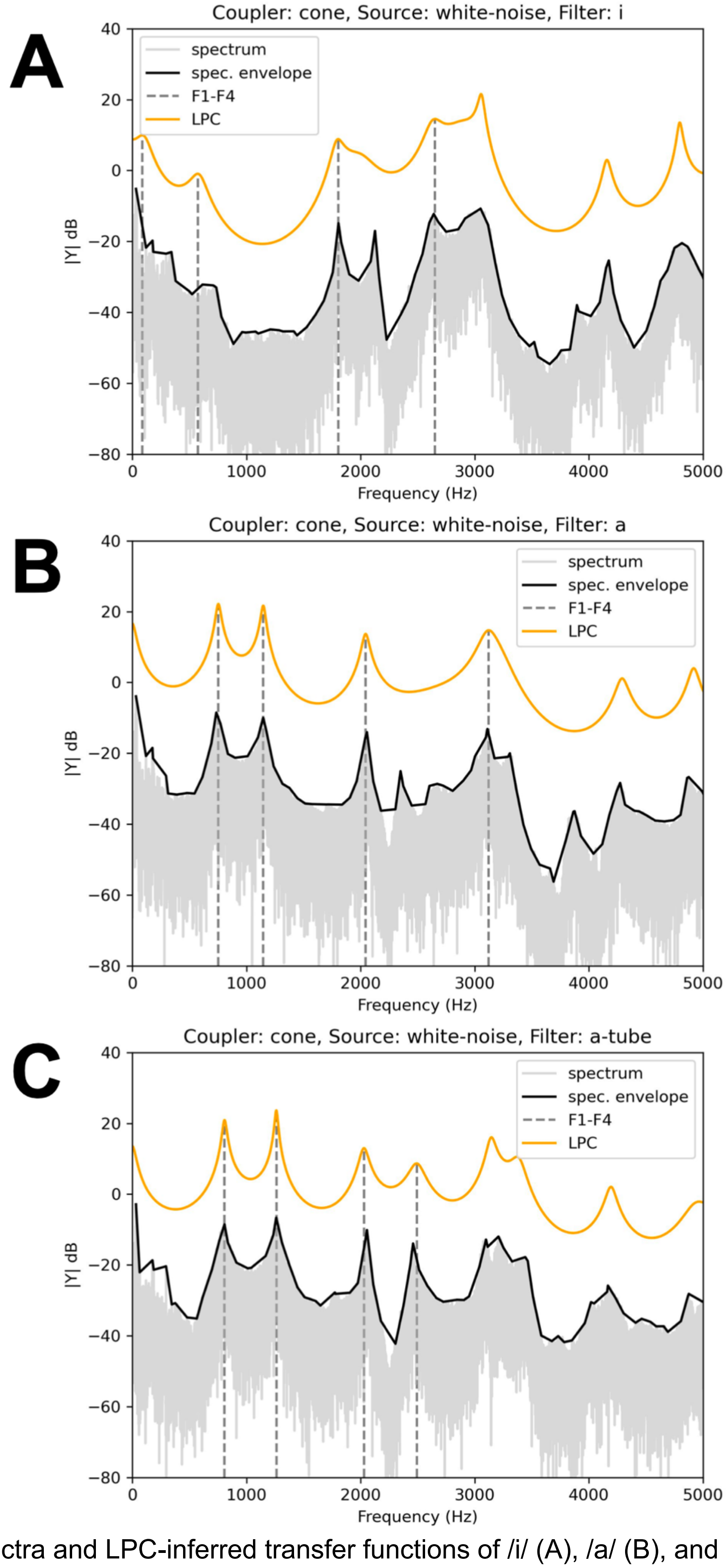

Figure 3: Frequency spectra and LPC-inferred transfer functions of /i/ (A), /a/ (B), and the tube model derived from /a/ (C).

Code for the project can be found at the following Github repository:
https://github.com/prlabu/PhysicalVocalTract

Bibliography

Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., & Fleischer, M. (2020). Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. *Scientific Data*, *7*(1), 255. https://doi.org/10.1038/s41597-020-00597-w

Fant, G. (1960). *Acoustic theory of speech production*. Moulton, the Hague.

Hermann, L. (1894). Phonophotographische Untersuchungen VI. Nachrag zur Untersuchung der Vocalcurven. *Pflüger Archiv, Now European Journal of Physiology*, *43*.

Simonyan, K., Ackermann, H., Chang, E. F., & Greenlee, J. D. (2016). New Developments in Understanding the Complexity of Human Speech Production. *Journal of Neuroscience*, *36*(45), 11440–11448. https://doi.org/10.1523/JNEUROSCI.2424-16.2016