# TOWARDS A SOFT FLUIDIC ELASTOMER TONGUE FOR A MECHANICAL VOCAL TRACT

*Peter Birkholz*[1]*, Christian Kosmas Mayer*[2]*, Patrick Häsner*[1]

[1]*Institute of Acoustics and Speech Communication, TU Dresden,* [2]*Schaufler Lab@TU Dresden*
*peter.birkholz@tu-dresden.de*

**Abstract:** Most mechanical vocal tract models use rigid-bodied robotic components to represent soft articulators such as the tongue and the velum. However, such rigid articulators have limited capabilities to deform realistically during contact with the vocal tract walls, to generate flow-induced vibrations for trill consonants, or to dampen the acoustic resonances of the vocal tract. Here we present the first prototype of a soft fluidic elastomer tongue based on the ideas of soft robotics. The tongue was designed as a soft silicone structure with three water-filled chambers for the anterior, middle, and posterior parts of the tongue. The tongue shape is controlled by the volumes of water in the three chambers. Integrated into a simplified vocal tract model, we explored the potential of this design to reproduce human tongue shapes for vowels, and synthesized the corresponding speech sounds using a vibrating reed source at the glottis. While it was possible to create plausible tongue shapes for the vowels /a, i, e, ɛ/, the vowels /o/ and /u/ could not be reproduced. The F1-F2 formant space of the synthesized sounds was considerably smaller than the human formant space. Consequently, follow-up designs of this type of tongue would benefit from finite-element simulations to optimize the possible variations in shape.

## 1 Introduction

Mechanical vocal tract models with movable articulators are valuable tools for speech research and education [1]. However, current designs mostly use rigid-bodied robotic parts to represent what are in reality soft articulators like the tongue, the velum, or the lips. For example, Howard [2] modelled the 2D sagittal shape of the vocal tract in terms of movable rigid bodies for the tongue, the lips, and the velum, and Umeda [3] modelled the vocal tract rather abstractly in terms of a straight tube with movable cuboids to change its cross sections. In contrast, Fukui et al. [4] and Endo et al. [5] used soft rubber to model the *surface* of the articulators, in particular of the tongue, but used movable rigid bodies attached to the soft surfaces to manipulate their shape. These approaches neither allow a realistic deformation of the articulators, when they come into contact with the vocal tract walls or with each other (e.g. during most consonants), nor flow-induced vibrations for trill consonants. Furthermore, hard materials do not sufficiently dampen the acoustic resonances compared to the soft walls of the human vocal tract [6].

In this pilot study, we built upon the ideas from the field of soft robotics [7], and designed and evaluated a soft fluidic elastomer tongue as part of a mechanical vocal tract. The goal was to create a tongue model that can reproduce the sagittal tongue shapes for vowels, using only a few degrees of freedom. Sec. 2 will first detail the design of the vocal tract model and the tongue, and then describe the experimental assessment of the model performance. The results are presented and discussed in Sec. 3.

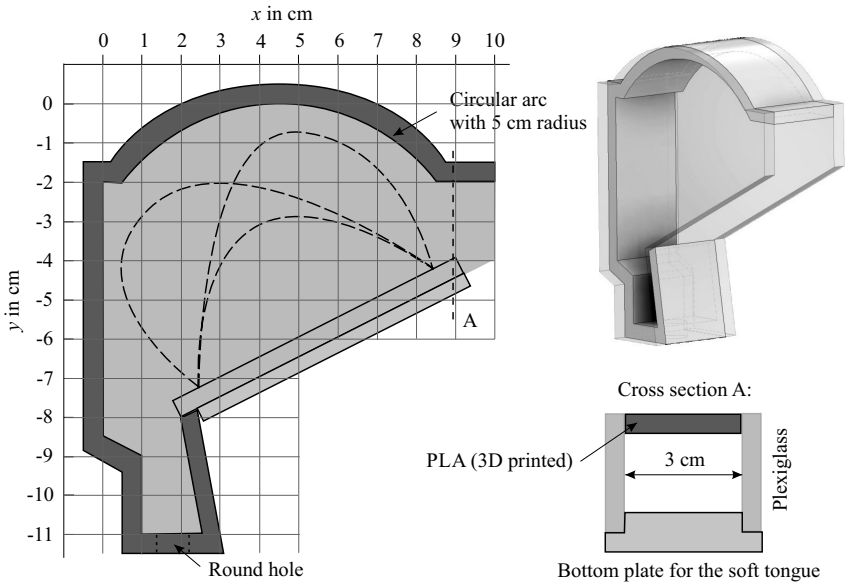


**Figure 1** – Setup of the mechanical vocal tract. Its size is based on the MRI data of the male speaker s1 in the Dresden Vocal Tract Dataset [8]. The dashed curves illustrate some of the targeted tongue shapes for vowels.

# 2 Method

## 2.1 Vocal tract walls

To enable a proper assessment of a soft model of the tongue to be made, it needs to be embedded in the confined space of a suitable matching vocal tract. Here we created a simplified model of the vocal tract using boundaries as shown in Figure 1. Its gross dimensions are based on measurements taken from the MRI data of the vocal tract of subject 1 of the Dresden Vocal Tract Dataset [8]. The side walls of the model were made from two parallel plexiglass plates separated by a distance of 3 cm, to ensure the tongue inside was visible. The superior, posterior, inferior, and anterior boundaries of the model replicate the sagittal profiles of the hard and soft palate, the posterior pharynx, and the larynx. These boundaries were 3D-printed in one piece on an Ultimaker 3 printer using polylactic acid (PLA) with 100% infill ratio. A hole of 1 cm diameter at the glottal end allowed the injection of an acoustic source signal to enable the synthesis of speech sounds. The anterior inferior side of the model allowed the insertion of a tongue model between the plexiglass plates. The anterior side of the model remained open, corresponding to a constant mouth opening. Since the focus of the study was on the tongue, no other articulators were replicated in detail.

## 2.2 Soft fluidic silicone tongue

The basic idea of fluidic elastomer actuators is to create motion by inflating specially designed chambers with soft elastomer walls using pressurized fluids. In our case, the actuator needed to be able to assume the same shapes as the human tongue for the generation of speech sounds. As we restrict ourselves to vowels in this study, the range of shape variation is defined by the corner vowels /a/, /i/, and /u/, which are shown in Figure 2. For /i/, the tongue is as far forward and up as possible (without creating significant frication noise), for /u/ as far back and up as possible, and for /a/ in as low and back a position as possible. For all other vowels, the tongue shapes lie

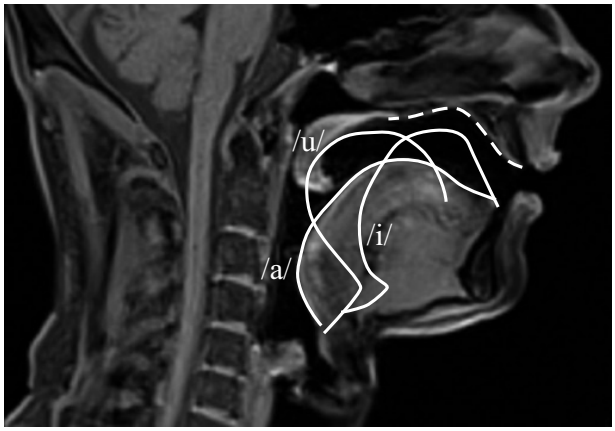


**Figure 2** – Tongue shapes for the vowels /a/, /i/, and /u/ based on the MRI data of speaker 1 of the Dresden Vocal Tract Dataset [8].

somewhere between these extremes and can be considered as the weighted superposition of the shapes for /a/, /i/, and /u/ [9].

The approach taken here was to design a soft fluidic tongue with three chambers for the anterior, middle, and posterior parts of the tongue. Each chamber, when pressurized, was supposed to expand mainly in the direction of one of the extreme positions for the three corner vowels. It should then be able to generate intermediate shapes with specific combinations of fluid pressures in the three chambers. With three fluid chambers that can be individually pressurized, this tongue has three degrees of freedom (DOF). This agrees with the finding that the gross deformation of the human tongue can be described by 2–3 DOF or principal components [10].

While many soft fluidic actuators uses pressurized air, we used water here to give the artificial tongue a density that is comparable to the human tongue. We believe this is important for realistic acoustic conditions in the vocal tract model.

To create the three-chamber actuator, we used the technique of lamination-based casting [7]. For the soft main part of the actuator, a mold was designed with the CAD software Inventor 2022 and 3D-printed with an Ultimaker 3 printer using PLA. Figure 3a shows an exploded view of the multiple parts of the mold, and Figure 3b shows the assembled mold (apart from the cover on one side). A soft silicone rubber with a high tear resistance (Dragon Skin 10NV by KauPo Plankenhorn, shore hardness A10) was mixed of its two liquid components, degassed in a vacuum chamber, and poured into the mold in Figure 3b (lying on its closed side). Then the missing side wall was screwed to the mold. After curing, the mold was disassembled to expose the silicone object, which is shown as the upper element in Figure 3c. The PLA blocks inside the three chambers were removed through the openings in the bottom of the object. The thickness of the silicone walls was 1 mm, and all edges were rounded to prevent tearing upon inflation.

The bottom part of the actuator consisted of a rigid 3D-printed plate, with a layer of a harder silicone (Zhermak ZA 50 LT by Troll Factory Rainer Habekost, shore hardness A50) on top. These two flat structures are shown in the exploded view in Figure 3c, and as the black and blue parts in Figure 3d. The silicone layer of the bottom part had a little hump at the anterior end to achieve a slightly elevated tongue tip. Six metal feed-through nozzles were screwed to the bottom part as shown in Figure 3c and connected the two layers. There were two nozzles per chamber, so that the air in the chambers could escape through one nozzle when they were completely filled with water through the other nozzle. The main part and the bottom part were

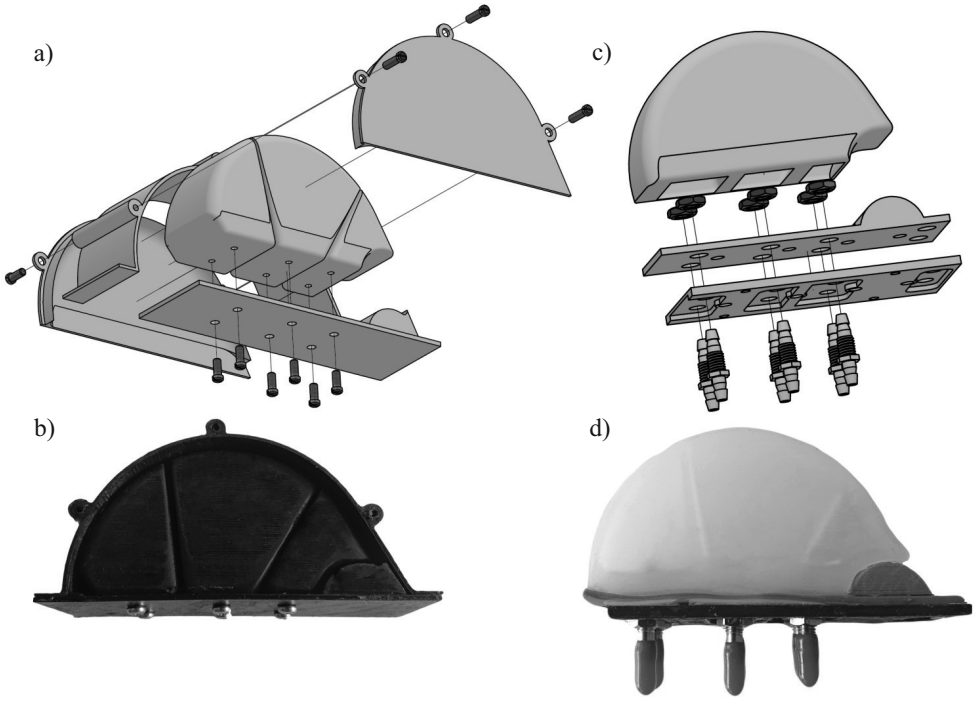


**Figure 3** – a) Exploded view of the mold used to cast the soft fluidic tongue. b) The mold assembled from its 3D-printed parts. c) From top to bottom: the cast main part of the tongue, the upper and lower layers of the bottom part of the tongue, and the feed-through nozzles to fill the chambers with water. d) Photo of the finished tongue model.

laminated together with Dragon Skin 10NV.

Figure 3d shows a photo of the fully assembled actuator with red caps on the feed-through nozzles. Due to the slightly transparent silicone of the main component, the two inner walls that separate the three chambers from each other can be clearly seen. The 3D-printable CAD files for the tongue mold as well as the CAD files for the vocal tract walls are contained in the supplemental material at `https://www.vocaltractlab.de/index.php?page=birkholz-supplements`.

### 2.3 Experiment

To assess the capability of the soft fluidic tongue to reproduce human-like tongue shapes for vowels, the actuator was integrated into the vocal tract model, and its three chambers were filled with water. Three 60 ml water-filled syringes were connected to the ports of the three chambers via silicone tubes (the second port of each chamber was closed with a cap) and used to control the volume of water, and hence the pressure, in the chambers.

By the systematic variation of the water volumes, a range of deformations of the actuator was created. In this process, the volumes in the chambers were both increased and decreased relative to the neutral shape (Figure 3d) in multiple steps. The chambers were filled only to the extent that no critical constriction or occlusion occurred in the vocal tract. The sagittal contours of the generated tongue shapes were visually compared to tongue contours of vowels in magnetic resonance images of the vocal tract of a male speaker (subject 1 in the DVTD [8]). The generated shapes that were deemed similar to the human tongue shapes of vowels, or represented certain special cases (e.g. with maximally inflated chambers before the formation of a critical constriction) were first photographed and subsequently used to synthesize the

corresponding sound.

For the synthesis, a vibrating reed source was used to generate a source signal that was injected at the glottis of the vocal tract model [11]. Although silicone vocal fold models can generate more natural-sounding synthetic vowels [12], the vibrating reed source proved to be more robust for longer operations. The generated audio signals were recorded about 50 cm from the mouth of the model with a high-quality studio microphone (M 930 by Microtech Gefell) connected to a laptop computer via a USB audio interface (TASCAM UH-7000). The recording software used was Audacity 2.3.3, and the signals were sampled at 44100 Hz with 16 bit quantization. All recordings were performed in an audio studio with sound-absorbing walls. The audio recordings of the synthetic sounds were each trimmed to a length of 1 s, normalized with respect to their peak amplitude, and faded in and out at the beginning and end over 25 ms each. The corresponding WAV files are contained in the supplemental material.

### 2.4 Formant analysis

To characterize the formant space of the vocal tract model with the soft fluidic tongue, the formant frequencies $F_1$ and $F_2$ were determined for all samples using the Praat (version 6.1.09) software. The standard Praat settings were used for this analysis except for the parameter "Number of formants". The number of formants was individually adjusted between 6 and 7.5 (in steps of 0.5) for each sound to obtain the most accurate formant estimates analogous to [13, 14]. The final values for $F_1$ and $F_2$ were determined as the mean values across the middle 0.5 s of the samples.

## 3 Results and discussion

Six of the generated tongue shapes were selected, because they were considered similar to human tongue shapes for vowels and sufficiently different from each other, or represented interesting special cases. Figure 4 shows photos of these shapes, where they are denoted as #1, ...#6. Shape #1 is the "neutral" shape in which the actuator was fabricated. It was designed for an approximately uniform cross-sectional area of the vocal tract that corresponds to the vowel schwa. The shapes #2, ...#6 were created by adding (positive $\Delta V$) or removing (negative $\Delta V$) water from the three chambers, as indicated by the bar graphs. The bars 1, 2, and 3 denote the posterior, middle, and anterior chamber. The tongue shape #2 resembles that of /a/. Here, the posterior chamber was inflated for a pharyngeal (non-critical) constriction, and the other two chambers were deflated for a wide-open oral cavity. The tongue shapes #3 and #4 resemble those of /e/ and /i/, respectively. For these, the anterior chamber was inflated for a constriction at the hard palate, and the two other chambers were deflated for a wide pharyngeal cavity. The shape #5 was an attempt to create an upper posterior constriction, which is required for /u/ (see Figure 2), by a strong inflation of the middle chamber. However, in this case, the middle chamber (and so the whole tongue) not only expanded into the desired direction, but also towards the anterior and posterior sides. Hence, a plausible tongue shape for /u/ (as well as /o/) could not be achieved. Finally, shape #6 represents the case where the middle chamber was strongly deflated, and the other two were strongly inflated. This shape resembles the tongue for /a/ with an additionally raised tongue blade, like for /ʃ/ in the context of /a/.

Figure 5 shows the first two formant frequencies obtained with the six tongue shapes. For comparison, the formant frequencies of German vowels according to Pätzold and Simpson [15] are shown in the same Figure. It can be seen that the area of the formant space enclosed by the synthesized vowels is smaller than the area enclosed by the natural vowels. As expected from the visual analysis of the generated tongue shapes, the vowels /u/ and /o/ could not be syn-

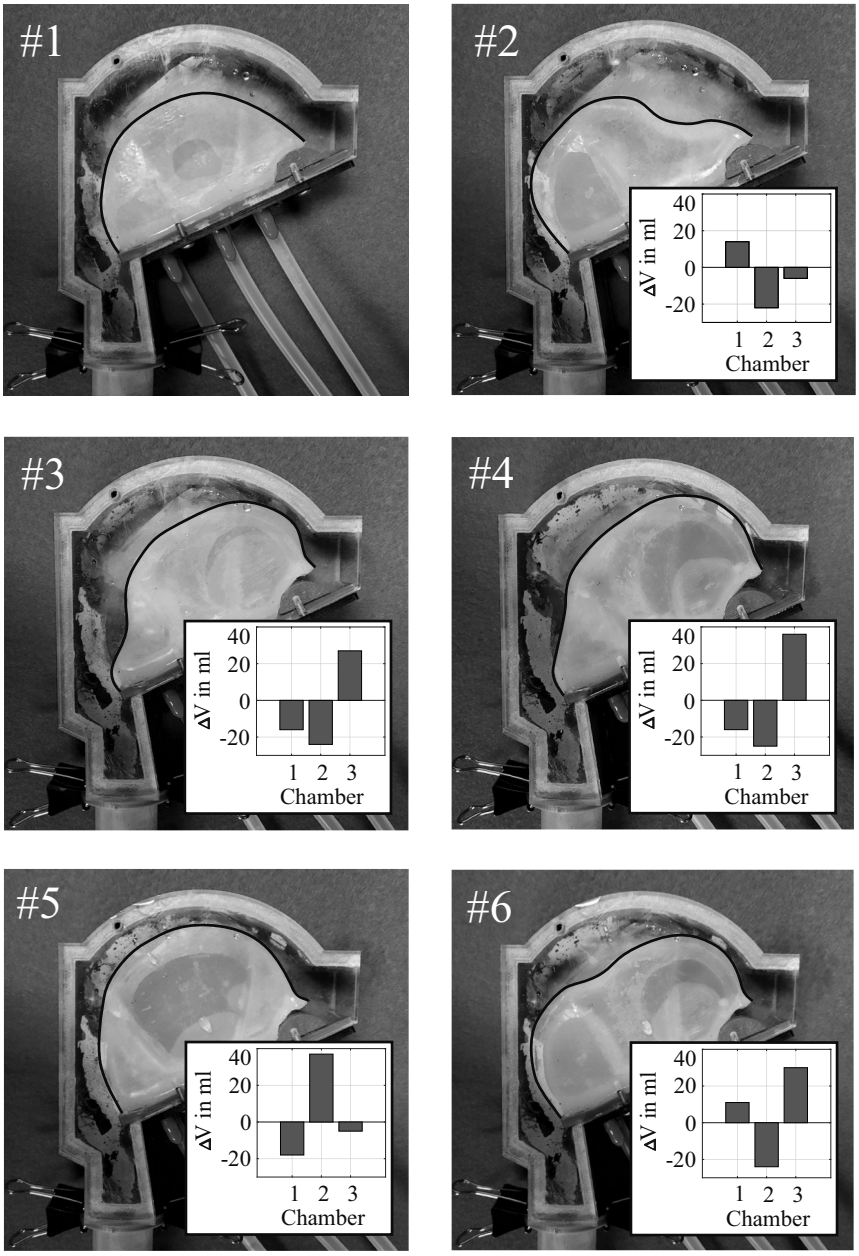


**Figure 4** – Six different tongue shapes generated with different fluid volumes in the three tongue chambers. Shape #1 represents the neutral tongue shape. The shapes #2 to #6 differ from the neutral shape due to different volumes of water in the chambers. The deviations from the neutral volumes are given as bar graphs, where "1", "2", and "3" denote the posterior, middle, and anterior chamber, respectively. In each photo, the contour of the tongue was traced with a black line to make its shape more clearly visible despite the fogged Plexiglass pane.

thesized. However, the expected formants for /i/ and /e/ could not be achieved either, although the tongue shapes for these vowels could be plausibly produced. These sounds probably strictly require spread lips and a higher larynx position, which could not be adjusted in the simplified

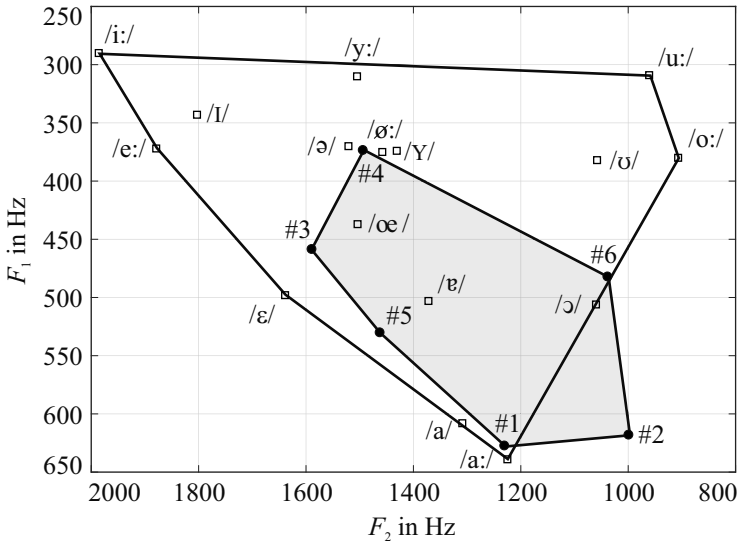


**Figure 5** – Formant map with the German vowels according to Pätzold and Simpson [15] as the reference (empty squares), and with the six vowel-like sounds generated with the fluidic silicone tongue (filled circles, denoted as #1, #2, ...#6). For both sets of sounds, the convex hull is shown, with a shaded background for the synthetic sounds.

vocal tract. It is furthermore interesting to note that the formants generated with the neutral tongue shape #1 correspond to the vowel /aː/ according to the reference formant data. However, the subjective quality of the synthesized sound rather resembles /ɐ/, while the sound with tongue shape #2 sounds much more like /a/. This may indicate that $F_2$ of /aː/ is too high in the reference data.

The results demonstrate that a fluidic elastomer tongue can generate a range of realistic tongue shapes and offers an interesting alternative for future mechanical or robotic vocal tract models. To be able to reproduce the tongue shapes of all speech sounds, the design of the soft fluidic tongue needs to be improved. For this, finite-element modeling tools like Ansys or COMSOL could help to simulate the possible deformations before the tongue model is manufactured [16]. Finally, additional articulators like the lips and the velum should be included as movable structures in the vocal tract model.

## Acknowledgments

This work was supported by the Schaufler Lab@TU Dresden (2020/21). We would also like to thank Ian Howard for proofreading this manuscript.

## References

[1] ARAI, T.: *Mechanical vocal-tract models for speech dynamics*. In *Interspeech 2010*. Makuhari, Japan, 2010.

[2] HOWARD, I. S.: *Robotic actuation of a 2D mechanical vocal tract*. In J. TROUVAIN, I. STEINER, and B. MÖBIUS (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pp. 25–32. TUDpress, Dresden, 2017.

[3] UMEDA, N.: *Phonemic feature and vocal feature-synthesis of speech sound, using an*

*acoustic model of vocal tract*. *The Journal of the Acoustical Society of Japan*, 22, pp. 195–203, 1965.

[4] FUKUI, K., T. KUSANO, Y. MUKAEDA, Y. SUZUKI, A. TAKANISHI, and M. HONDA: *Speech robot mimicking human articulatory motion*. In *Interspeech 2010*, pp. 1021–1024. Makuhari, Japan, 2010.

[5] ENDO, N., T. KOJIMA, H. ISHIHARA, T. HORII, and M. ASADA: *Design and preliminary evaluation of the vocal cords and articulator of an infant-like vocal robot lingua*. In *IEEE-RAS International Conference on Humanoid Robots 2014*, pp. 1063–1068. Madrid, Spain, 2014.

[6] BIRKHOLZ, P., P. HÄSNER, and S. KÜRBIS: *Acoustic comparison of physical vocal tract models with hard and soft walls*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022)*. Singapore, submitted.

[7] MARCHESE, A. D., R. K. KATZSCHMANN, and D. RUS: *A recipe for soft fluidic elastomer robots*. *Soft Robotics*, 2(1), pp. 7–25, 2015.

[8] BIRKHOLZ, P., S. KÜRBIS, S. STONE, P. HÄSNER, R. BLANDIN, and M. FLEISCHER: *Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties*. *Scientific Data*, 7(1), pp. 1–16, 2020.

[9] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLOS ONE*, 8(4), p. e60603, 2013.

[10] FUCHS, S. and P. PERRIER: *On the complex nature of speech kinematics*. *ZAS Papers in Linguistics*, 42, pp. 137–165, 2005.

[11] ARAI, T.: *Sound sources used in speech production research with physical models of the human vocal tract*. In *Proc. of the 3rd International Workshop on the History of Speech Communication Research (HSCR 2019)*, p. 79–84. Vienna, 2019.

[12] BIRKHOLZ, P., S. STONE, and S. KÜRBIS: *Comparison of different methods for the voiced excitation of physical vocal tract models*. In P. BIRKHOLZ and S. STONE (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 84–94. TUDpress, Dresden, 2019.

[13] BIRKHOLZ, P., F. GABRIEL, S. KÜRBIS, and M. ECHTERNACH: *How the peak glottal area affects linear predictive coding-based formant estimates of vowels*. *The Journal of the Acoustical Society of America*, 146(1), pp. 223–232, 2019.

[14] KATHIRESAN, T., D. MAURER, H. SUTER, and V. DELLWO: *Enhancing the objectivity of interactive formant estimation: Introducing Euclidean distance measure and numerical conditions for numbers and frequency ranges of formants*. In J. TROUVAIN, I. STEINER, and B. MÖBIUS (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pp. 130–137. TUDPress, Dresden, 2017.

[15] PÄTZOLD, M. and A. P. SIMPSON: *Acoustic analysis of German vowels in the Kiel Corpus of Read Speech*. *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung Univ. Kiel*, 32(1978), pp. 215–247, 1997.

[16] XAVIER, M. S., A. J. FLEMING, and Y. K. YONG: *Finite element modeling of soft fluidic actuators: Overview and recent developments*. *Advanced Intelligent Systems*, 3(2), p. 2000187, 2021.