COMPARISON OF DIFFERENT METHODS FOR THE VOICED EXCITATION OF PHYSICAL VOCAL TRACT MODELS

Peter Birkholz, Simon Stone, Steffen Kürbis

Institute of Acoustics and Speech Communication, Technische Universität Dresden
peter.birkholz@tu-dresden.de

Abstract: Physical models of the vocal tract need a suitable voice source to generate vocalic sounds. Here we constructed and tested three types of voice sources: a self-oscillating silicone model of the vocal folds, a vibrating reed source, and a source based on an electro-mechanical loudspeaker. Each of the sources was used to excite eight physical resonators representing the German tense vowels. The excitation signals generated by the sources were analyzed and the intelligibility of the generated vowel sounds was perceptually evaluated. The spectral slope of the excitation was steepest for the silicone vocal folds, least steep for the reed source, and in-between for the loudspeaker source. The intelligibility of the vowels was highest for the excitation with the silicone vocal folds, and lowest for the loudspeaker source.

1 Introduction

Physical models of the vocal tract are useful teaching implements to demonstrate the principles underlying speech production [1], and can potentially also be used as research tools to study speech-related phenomena. These models can be as simple as straight tubes with varying cross-sections along the tube axis [2] and as complex as fully articulated robotic vocal systems [3]. All of these models need a mechanism for the voiced excitation to generate vocalic sounds. Existing types of voice sources include self-oscillating, physical rubber models of the vocal folds (e.g., [4]), driven brass shutters representing the vocal folds [5], vibrating reeds [6], loudspeakers or horn drivers emitting a periodic excitation signal (e.g., [7]), ionophones [8], or impedance heads [9]. In this study we created three of these source types (silicone vocal folds, a reed source, and a loudspeaker source), and analyzed the produced excitation signals and their effect on the intelligibility of vowels generated by attaching physical resonators to the sources. The goal was to identify potential strengths and weaknesses of the sources.

2 Vocal tract resonators

For each of the tense German vowels /a, e, i, o, u, ɛ, ø, y/ one physical resonator was designed as a straight tube with circular cross-sections. The lengths and area functions were adopted from the corresponding vowels defined in the software VocalTractLab 2.2 (VTL, `www.vocaltractlab.de`) [10]. The acoustic side branches for the nasal cavity and the piriform fossae were omitted for the physical resonators. Furthermore, the "epilaryngeal tube" of the physical resonators was slightly widened to allow a seamless connection with the physical vocal folds (see [11] for details). All resonators were designed with 3 mm thick walls and a flange at the glottal end for the connection with the different types of sources. The resonators were 3D-printed on an Ultimaker 3 printer using the material polylactide (PLA) with an infill ratio

Figure 1: The 3D-printed vocal tract resonators for the tense German vowels /a, e, i, o, u, ɛ, ø, y/ (from left to right).

of 100%, and are shown in Figure 1. It is important to note that these resonators only approximate the corresponding vowels, because they have hard walls and omit any side cavities like the piriform fossae. Accordingly, their formant frequencies and bandwidths may deviate from perceptually optimal values.

3 Vowel synthesis with different sources

Each of the 3D-printed resonators was excited with three different types of voice sources to generate different samples of the vowels /a, e, i, o, u, ɛ, ø, y/: a self-oscillating silicone model of the vocal folds, a vibrating reed source, and an enclosed loudspeaker emitting periodic sound pulses. The silicone vocal folds and the reed source were driven with three subglottal pressures each. In addition, the eight vowels were synthesized with the articulatory synthesizer VTL using a two-mass model of the vocal folds. The different sources are shown in Figure 2 and briefly described in the following subsections, along with the synthesis methods. The CAD files needed for 3D-printing the resonators and the audio samples generated with the different sources are available as supplemental material from `http://www.vocaltractlab.de/index.php?page=birkholz-supplements`.

3.1 Silicone vocal folds

Using silicone rubber, it is possible to create physical vocal folds that closely resemble human vocal folds in anatomy and function (e.g., [12, 4]). For a realistic behaviour, the layered structure of the human vocal folds should be reproduced. Here we designed a three-layer model comprising a body layer (vocalis muscle), a cover layer of about 1 mm thickness (lamina propria), and a very thin protecting epithelium layer of about 50 μm thickness. All three layers were fabricated by molding addition-cure two-component silicone rubbers with different fractions of silicone oil, using individually created molds (see [11] for details of the fabrication process). The silicones for the body layer and the cover layer were created with Young's moduli of 2.2 kPa and 1.2 kPa, respectively, to resemble the according layers of human vocal folds

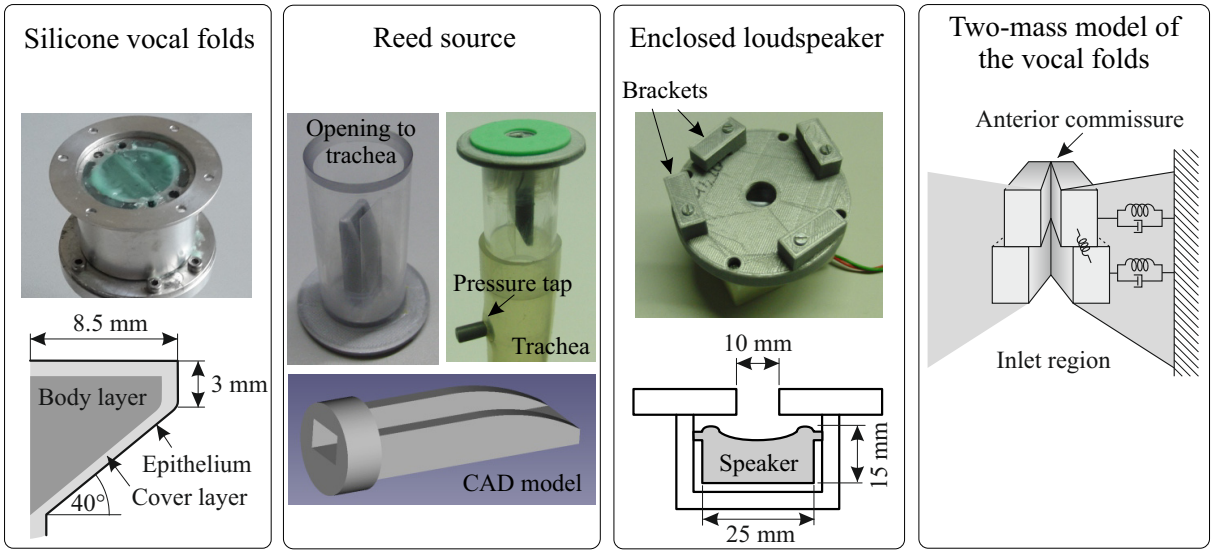


Figure 2: The sources used for the voiced excitation of the resonators. The two-mass model of the vocal folds was only used in the computer simulation.

[13, 14]. The vocal fold length was set to 17 mm, and the geometry of the central coronal cross-section is shown in Figure 2.

The measurement setup and procedure were analogous to the one described in [11]. The pressure supply and the subglottal system consisted of a compressor (air blower Medo LA 100A by Nitto Kohki), which was connected to an expansion chamber (30 cm x 30 cm x 50 cm) with a 60 cm long hose, and another 200 cm long hose that connected the expansion chamber with the vocal fold model. A manual one-inch shut-off valve connected to the expansion chamber was used to control the subglottal pressure, which was monitored by a pressure sensor connected to a pressure tap below the glottis. The valve, the expansion chamber, and the compressor were located in a separate soundproof cabin to prevent any noise interfering with the acoustic measurements. The resonators were attached to the top of the glottis model and the generated sounds were recorded with a measurement microphone 30 cm above the resonators with a sampling rate of 44.1 kHz and 16 bit quantization. Using this setup, the acoustic output of each of the resonators was recorded for three subglottal pressure values: 0.8, 1.2, and 1.6 kPa. The lowest pressure of 0.8 kPa was just above the threshold where the vocal folds oscillated in a stable fashion with all resonators, and the highest value of 1.6 kPa was below the pressure where oscillations became chaotic with some resonators.

3.2 Vibrating reed source

The vibrating reed source was a design by T. Arai and is an improved version of the design published in [6]. The central element is the channel-like "shallot" shown as a CAD model in Figure 2, which was kindly provided to us by T. Arai. This shallot has been extended by a circular "base plate" and was 3D-printed with an Ultimaker 3 printer using the material PLA. The "tongue" was cut out of a plastic sheet (used for overhead projectors) and glued to the upper edges of the left and right sides of the groove, so that the distal part of the tongue was free to vibrate. A plexiglas tube of 25 mm inner diameter and 50 mm length was glued to the base plate to encompass the shallot.

The assembled reed source was connected to the same "subglottal" pressure source as the silicone vocal folds. As with the silicone vocal folds, the range of pressures at which the reed oscillated in a periodic way depended on the attached resonator. For most of the resonators,

there was a stable oscillation of the reed between 300 and 600 Pa. Exceptions were the resonators for /o, u, ø, y/, which needed pressures higher than 300 Pa to initiate the oscillation. For pressures above 600 Pa, the oscillation started to become chaotic for some resonators. Accordingly, we selected three discrete pressure values to drive the reed source and record the generated vowel sounds, namely 300 Pa, 450 Pa, and 600 Pa. When the reed did not oscillate for a certain combination of resonator and pressure value, the corresponding sample was omitted from the subsequent analysis.

3.3 Enclosed loudspeaker

The loudspeaker source was constructed by putting a miniature speaker (type 32KC08-1 by Veco Vansonic, 8 Ω, 3 W, 2 cm diaphragm diameter) into a 3D-printed enclosure as shown in Figure 2. The sound generated by the loudspeaker was injected into the "glottal" end of the resonators through a hole with a diameter of 10 mm. The signal $x(t)$ emitted by the loudspeaker was a sequence of temporally differentiated glottal flow pulses, i.e., $x(t) = \mathrm{d}u_\mathrm{g}(t)/\mathrm{d}t$ at a constant f_0 of 90 Hz. The glottal flow pulses $u_\mathrm{g}(t)$ were designed according to the pulse shape "B" proposed by Rosenberg [15] with relative opening and closing times of 40% and 16%, respectively. The signal $x(t)$ was played back using the software Audacity 2.0.2. The equalizer included in the software was used to compensate the free-field frequency response of the loudspeaker, and the volume was adjusted such that were no measurable nonlinear distortions.

3.4 Simulated two-mass model of the vocal folds

Besides the three physical sources, a modified two-mass model of the vocal folds [16] was used to computationally synthesize the eight vowels using the articulatory speech synthesizer VocalTractLab 2.2 [17]. The vowels were synthesized in the time-domain using the same vocal tract shapes that the physical models were based on. To make the simulations as similar as possible to the setup with the physical tube models, the computer model of the vocal tract assumed an infinite wall impedance, no sound radiation via the "skin", and omitted any side cavities (nasal cavity and the piriform fossae). The parameters of the modified two-mass model were adjusted for modal phonation with a subglottal pressure of 800 Pa, a fundamental frequency of 120 Hz, rest displacements of the lower and upper mass elements of 0.05 mm, no extra area between the arytenoid cartilages, and an aspiration strength of -40 dB.

4 Analysis of excitation signals

4.1 Method

One way to obtain the excitation signals generated by the different types of sources would be to measure the free-field sound pressure signals radiated from the sources when no resonator is attached. However, this approach would neglect the effect of the vocal tract filter on the sources. In contrast to the widespread assumption of the independence of source and filter, the effects are usually not negligible: One major effect is that the input inductance of the vocal tract causes the glottal volume velocity waveform to be skewed to the right with respect to the glottal area [18]. The main acoustic effect of this skewing is an increased strength of the excitation. Furthermore, the input inductance of the vocal tract facilitates the oscillation of the vocal folds and reduces the oscillation threshold pressure [19]. This was also observed for the silicone vocal folds here: while they were able to oscillate at a "lung" pressure of 800 Pa when a resonator was attached, a much higher pressure was needed for self-sustained oscillation when no resonator was attached.

Based on these considerations, we decided to obtain the excitation signals of the different source types by inverse filtering (based on Linear Predictive Coding, LPC) [20] the sound pressure signals that were radiated when the sources were connected to a cylindrical tube as resonator. The cylindrical tube had a length of 16.54 cm and a cross-sectional area of 2 cm^2 and was manufactured like the other resonators by 3D-printing. A cylindrical tube was used because its resonance frequencies are widely and regularly spaced. This makes the LPC-based estimation of its transfer function robust and reliable. Hence, the cylindrical tube was excited by the different source types and pressure values analogously to the vowel synthesis in Section 3, and the radiated pressure signals were recorded, inverse filtered, and analyzed with the software Praat (version 6.0.28 [21]).

For each recorded sample, the following steps were performed in Praat:

- The audio signal was resampled to a frequency of 11000 Hz using the function *Convert* → *Resample*, and 0.5 s from the middle of the sample were extracted for further processing.
- The LPC coefficients were determined using the function *Analyse spectrum* → *To LPC (burg)* using the standard settings (except the prediction order). The prediction order, i.e., the number of LPC poles, was set to 14 for the loudspeaker source, and to 12 for all other sources. These numbers were manually determined for an optimal correspondence between the poles and the peaks in the spectrogram. The two additional poles for the loudspeaker source were needed because of an additional formant in the corresponding sample (see below and Figure 4).
- The sample was inverse filtered by selecting both the audio sample object and the LPC object and clicking *Filter (inverse)*.
- From the resulting excitation signal, the pitch-corrected long-term average spectrum [22] (LTAS) was calculated with the function *Analyse spectrum* → *To Ltas (pitch-corrected)* with the standard settings, i.e., a bin width of 100 Hz. This calculation was not possible for the sample of the silicone vocal folds at 800 Pa, because the signal was not as periodic as necessary for the calculation of the pitch-corrected LTAS.
- From the LTAS the slope was calculated as the difference of the average energy (in dB) in the frequency bands 0 . . . 1 kHz and 1 . . . 4 kHz using the function *Query* → *Get slope* on the LTAS object.

4.2 Results and discussion

Figure 3 shows the excitation signals obtained by inverse filtering in terms of the time signals (left) and the long-term average spectra (right). According to source-filter theory, the time waveforms correspond to the temporal derivatives of the glottal flow signals. The signal shape varied considerably between the source types and is difficult to interpret in a meaningful way. The waveform that is most reminiscent of the *theoretical* shape of the glottal flow derivative during phonation (see, e.g., [23, ch. 2]) is the one generated by the two-mass model.

In contrast to the time signals, all LTAS do resemble the expected shape of excitation spectra. All spectra are broadband with a negative slope, and fluctuations are mostly below 10 dB. According to Table 1, the spectral slopes of the excitation signals of the silicone vocal folds and the two-mass model are steeper than those of the reed source, meaning that the reed source produces stronger high-frequency components. The spectral slope of the excitation with the enclosed loudspeaker is intermediate.

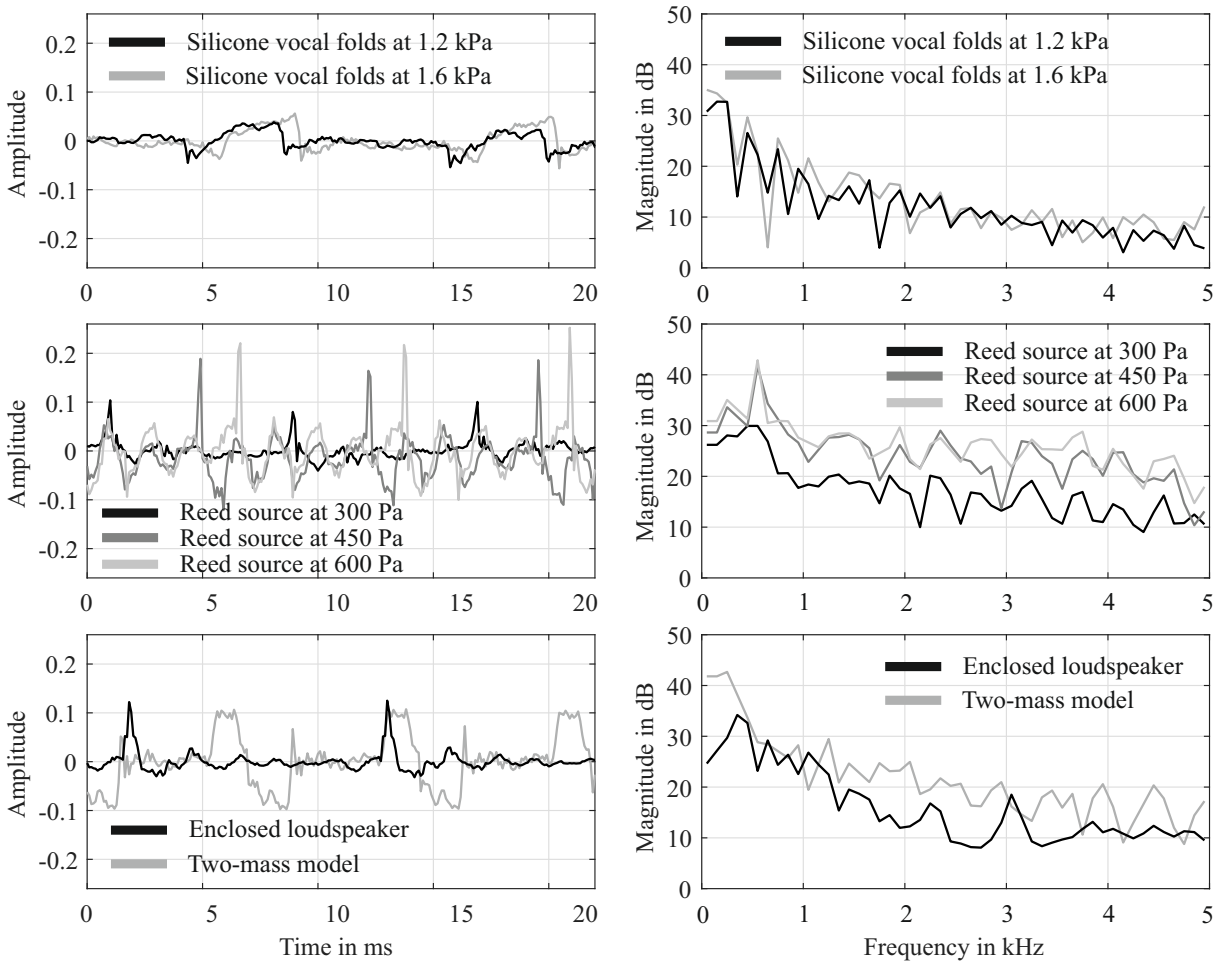


Figure 3: Comparison of the time signals (left) and the corresponding long-term average spectra (right) of the excitation signals generated by the different source types. The excitation signals were obtained by inverse filtering the sound pressure signals radiated by a cylindrical resonator excited with the respective sources.

With regard to the formant frequencies that were obtained with the LPC analysis and used for the inverse filtering, we found that the sample generated with the enclosed loudspeaker differed markedly from all other samples. For the loudspeaker sample, the formant frequencies were not close to the expected resonance frequencies of the cylindrical tube with a closed glottal end, i.e., $F_n = (2n-1)c/(4L) = 518$ Hz, 1554 Hz, 2590 Hz, ..., where $L = 16.54$ cm is the tube length and $c = 343$ m/s is the sound velocity at 20°C. Instead, they were strongly shifted towards lower frequencies. Figure 4 illustrates this by means of the LTAS of the loudspeaker sample in contrast to the LTAS of a sample generated with the silicone vocal folds. While F_1 and F_2 are still close to each other in both samples, the higher formants differ considerably. This formant shift must be attributed to the coupling of the resonator tube with the air-filled cavity between the loudspeaker diaphragm and the "glottis plane", which effectively increases the volume of the resonator. The smaller spacing of the formants was also the reason why an LPC order of 14 (instead of 12) was required for the loudspeaker source to correctly represent the formant structure.

Table 1: Spectral slope of the excitation signals in dB

Excitation signal	Slope in dB
Silicone vocal folds at 1200 Pa	-15.7
Silicone vocal folds at 1600 Pa	-15.8
Reed source at 300 Pa	-9.7
Reed source at 450 Pa	-9.4
Reed source at 600 Pa	-8.9
Enclosed loudspeaker	-11.8
Two-mass model	-16.2

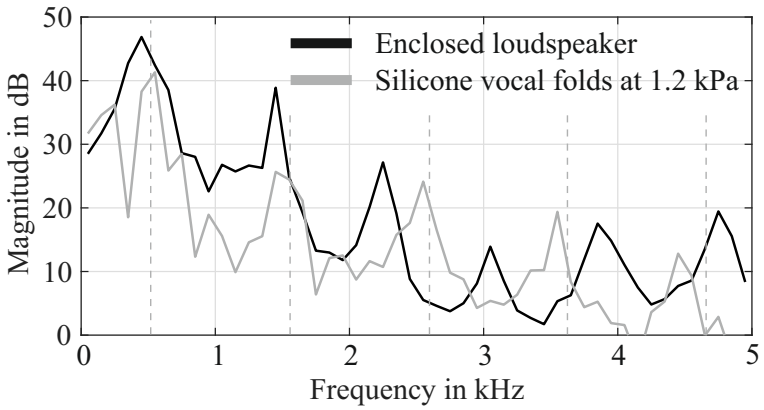


Figure 4: Long-term average spectra of the cylindrical resonator excited with the enclosed loudspeaker (black) and the silicone vocal folds (gray). The dashed lines indicate the theoretical resonance frequencies of the cylindrical resonator.

5 Perception experiment

5.1 Method

The perception experiment was conducted to find out how the type of voice source affects the intelligibility of the vowels. Often, the intelligibility of speech is mainly attributed to the (supraglottal) articulation instead of the voice source characteristics. However, the artificial sources used here might introduce unknown effects that affect the intelligibility as well.

The test material consisted of the audio samples generated with the four types of sources in Section 3. There, for each of the eight resonators, three samples were generated with the silicone vocal folds (at 800, 1200, and 1600 Pa), three samples were generated with the reed source (at 300, 450, and 600 Pa), one sample was generated with the loudspeaker source, and one sample with the two-mass model, for a total of theoretically 64 samples. However, as the reed source failed to oscillate for the vowels /o, u, ø, y/ at 300 Pa, and for the vowels /o, u/ at 450 Pa, only 58 samples could be used for the perception experiment. To generate the stimuli for the experiment, all samples were symmetrically cropped around their center to a length of 600 ms, peak-normalized and windowed with a Tukey window of the same length and $\alpha = 0.03$, resulting in a fade-in and fade-out of 9 ms each. Finally, the stimuli were padded with 200 ms of silence at the beginning and end.

The listening test was conducted as a multiple forced-choice experiment using the software Praat [21]. Three blocks of the 58 stimuli were presented to the subject for a total of 174 trials. The order of the stimuli was randomized within each block and for each subject. In every trial,

the subject had to listen to the stimulus and assign it to one of the eight vowels by clicking on one of eight buttons on a computer screen labeled with the German letters corresponding to the vowels. Subjects were allowed to repeat each stimulus up to three more times before having to make a decision.

After both 60 and 120 trials the subjects were asked to take a short break before continuing with the test. The subjects were 20 German native speakers (13 male, 7 female) between 23 and 65 years (median age 33.5 years), who reported no known hearing impairments. The stimuli were played back in a quiet room using a Terratec "Aureon X Fire 8.0 HD" external sound card and AKG K240 Monitor semi-open headphones.

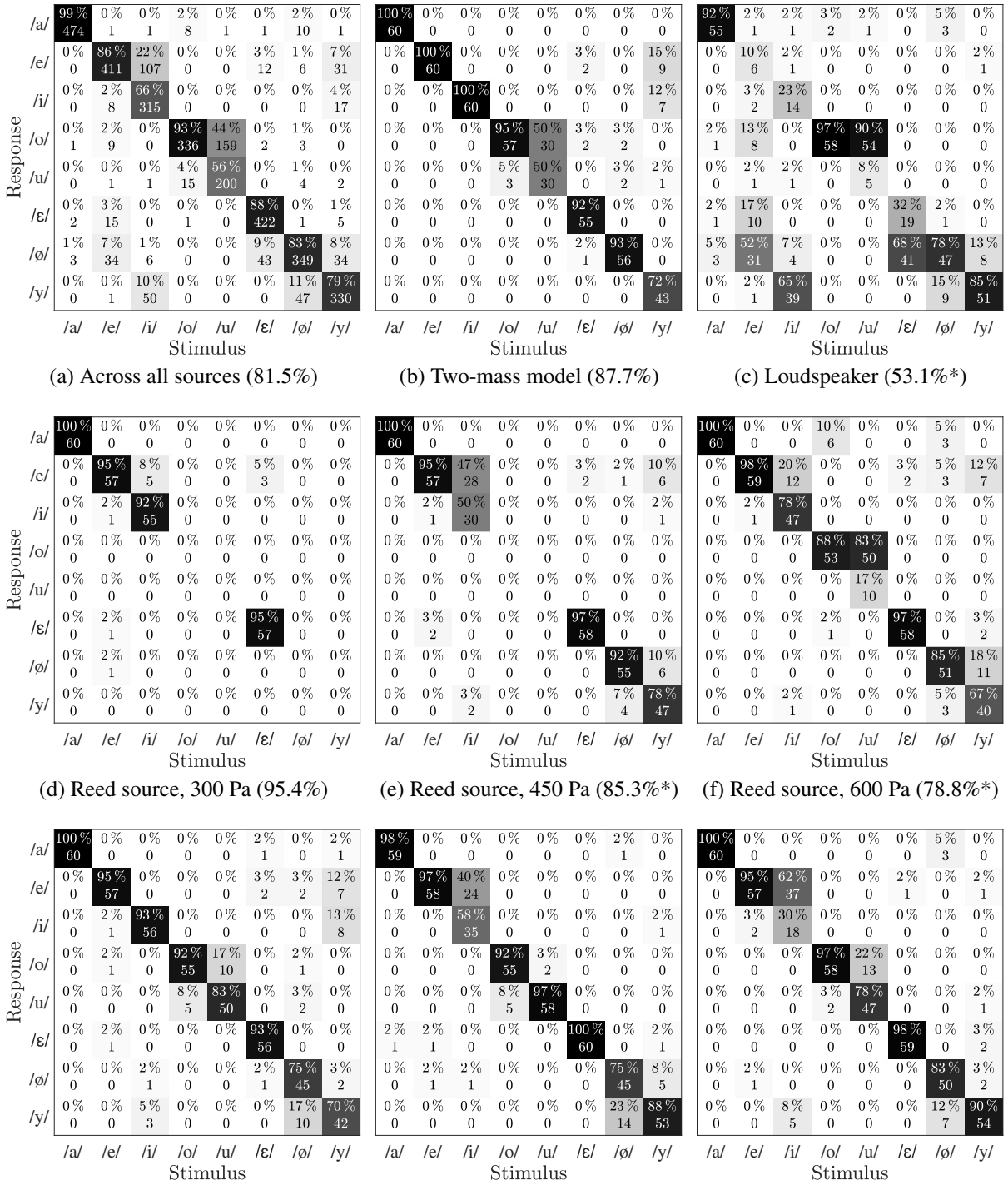
5.2 Results and discussion

The results of the listening test are shown in Figure 5. The first take-away from the experiment is the global recognition rate of the vowel sounds across all sources shown in Figure 5(a): Most confusions occurred for /u/ and /i/ and, to a less severe extent, for /y/. This is most likely an indicator for shortcomings of the resonator geometries for these sounds, independent of the excitation. This assumption is further supported by the fact that even the corresponding VTL simulations did not score as high as might be expected, even though they use an ideal (simulated) excitation source. As noted in Sections 2 and 3, the physical resonators had hard walls and no side branches, and the VTL simulation parameters were adjusted to emulate these conditions. However, the main tube geometries of the vowels in VTL were acoustically optimized under the condition of soft walls and attached side branches. Hard walls and the neglect of the side cavities accordingly altered the vowel qualities to some extent (e.g., F_1 of /u/ was shifted by as much as 13%). Hence, the results of the perception experiment for the physical sources should not be regarded as absolute but related to the baseline set by the VTL stimuli.

When comparing the identification accuracy (the number of correct responses divided by the number of stimuli) of the physical sources with the baseline set by VTL, the loudspeaker source and the reed source at 450 Pa and 600 Pa are significantly worse according to a paired two-sample t-test. With regard to the loudspeaker source, the low accuracy of 53.1% is probably due to the strong acoustic coupling of the resonators and the source, as noted in Section 4.2. This coupling likely affected the formant frequencies and hence the quality of individual vowels. A possible solution to the problem would be to use a smaller hole in the loudspeaker enclosure. However, when the opening gets too small, the increase of the acoustic impedance above the loudspeaker diaphragm might easily cause nonlinear distortions of the source.

With regard to the reed source, the absence of some samples at 300 Pa and 450 Pa leads to a recognition accuracy that may be misleading at first, and for comparison, the respective sample subset generated with the two-mass model in VTL needs to be considered: For example, the reed source at 300 Pa scored a global accuracy of 95.4%, but the accuracy across the corresponding VTL subset /a, e, i, ɛ/ was 98%. Because of the lack of data points, it is not entirely clear whether the reed source at 300 Pa is significantly worse than the baseline or not. However, for 450 and 600 Pa, the samples from the reed source were identified significantly worse than the baseline.

For the samples generated with the silicone vocal folds, the recognition score did not significantly differ from the baseline for either of the three pressure values. Broken down by sound, the most notable differences are that the /u/ stimuli were more reliably identified by the subjects using the silicone vocal folds, while the accuracy for /i/ fell with rising subglottal pressure for the silicone vocal folds. Since the VTL stimuli were only generated with one fixed pressure of 800 Pa, it remains to be investigated whether this effect could also be observed in the baseline.



(a) Across all sources (81.5%) (b) Two-mass model (87.7%) (c) Loudspeaker (53.1%*)

(d) Reed source, 300 Pa (95.4%) (e) Reed source, 450 Pa (85.3%*) (f) Reed source, 600 Pa (78.8%*)

(g) Silicone vocal folds, 800 Pa (87.7%) (h) Silicone vocal folds, 1200 Pa (88.1%) (i) Silicone vocal folds, 1600 Pa (84.0%)

Figure 5: Confusion matrices of the listening test results. The * marks a recognition accuracy that is significantly different from the accuracy achieved with the VocalTractLab reference stimuli (two-mass model) according to a paired two-sample t-test with $\alpha = 0.05$.

6 Conclusions

The three examined physical source types do affect the perception of the tense German vowels in different ways. The vowel identification score was lowest for the loudspeaker source, probably caused by the strong source-filter coupling. When a source similar to this is considered for the excitation of physical resonators, this problem needs to be be addressed. For the self-oscillating silicone vocal fold model, the identification scores were highest, and self-sustained oscillations were possible for a wide range of subglottal pressures ($800 - 1600$ Pa) across all resonators. Hence, the silicone vocal folds appear to be an ideal voice source for physical resonators. The identification scores of the vowels generated with the reed source were somewhat lower than for the silicone vocal folds (possibly due to the less steep spectral tilt of the source signal), and the range of pressures that allowed periodic self-sustained oscillations was smaller than for the silicone vocal folds. On the other side, the reed source is easier to manufacture than the silicone vocal folds.

References

[1] T. Arai, "Mechanical vocal-tract models for speech dynamics," in *Interspeech 2010*, Makuhari, Japan, 2010.

[2] ——, "Education system in acoustics of speech production using physical models of the human vocal tract," *Acoustical Science and Technology*, vol. 28, no. 3, pp. 190–201, 2007.

[3] K. Fukui, T. Kusano, Y. Mukaeda, Y. Suzuki, A. Takanishi, and M. Honda, "Speech robot mimicking human articulatory motion," in *Interspeech 2010*, Makuhari, Japan, 2010.

[4] Y. Xuan and Z. Zhang, "Influence of embedded fibers and an epithelium layer on the glottal closure pattern in a physical vocal fold model," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 2, pp. 416–425, 2014.

[5] A. Barney, A. De Stefano, and N. Henrich, "The effect of glottal opening on the acoustic response of the vocal tract," *Acta Acustica united with Acustica*, vol. 93, no. 6, pp. 1046–1056, 2007.

[6] T. Arai, "Education in acoustics and speech science using vocal-tract models," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2444–2454, 2012.

[7] A. Hannukainen, J. Kuortti, J. Malinen, and A. Ojalammi, "An acoustic glottal source for vocal tract physical models," *Measurement Science and Technology*, vol. 28, no. 11, p. 115902, 2017.

[8] F. Fransson, "The STL ionophone sound source," *STL-QPSR*, vol. 6, pp. 27–30, 1965.

[9] J. Wolfe, D. T. W. Chu, J.-M. Chen, and J. Smith, "An experimentally measured source-filter model: Glottal flow, vocal tract gain and output sound from a physical model," *Acoustics Australia*, vol. 44, no. 1, pp. 187–191, 2016.

[10] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.

[11] P. Birkholz, F. Gabriel, and S. Kürbis, "How the peak glottal area affects LPC-based formant estimates of vowels," submitted.

[12] A. H. Mendelsohn and Z. Zhang, "Phonation threshold pressure and onset frequency in a two-layer physical model of the vocal folds," *Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2961–2968, 2011.

[13] D. K. Chhetri, Z. Zhang, and J. Neubauer, "Measurement of Young's modulus of vocal folds by indentation," *Journal of Voice*, vol. 25, no. 1, pp. 1–7, 2011.

[14] F. Alipour and S. Vigmostad, "Measurement of vocal folds elastic properties for continuum modeling," *Journal of Voice*, vol. 26, no. 6, pp. 816.e21–816.e29, 2012.

[15] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 583–590, 1971.

[16] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Interspeech 2011*, Florence, Italy, 2011, pp. 2681–2684.

[17] P. Birkholz, "VocalTractLab [software]," 2017. [Online]. Available: http://www.vocaltractlab.de

[18] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, 1994.

[19] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1536–1552, 1988.

[20] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[21] P. Boersma and D. Weenik, "Praat: doing phonetics by computer [software]," 2017. [Online]. Available: http://www.praat.org/

[22] P. Boersma and G. Kovacic, "Spectral characteristics of three styles of Croatian folk singing," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1805–1816, 2006.

[23] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1998.