How the peak glottal area affects linear predictive coding-based formant estimates of vowels

Peter Birkholz,[1,a)] Falk Gabriel,[1] Steffen Kürbis,[1] and Matthias Echternach[2]
[1]*Institute of Acoustics and Speech Communication, TU Dresden, 01062 Dresden, Germany*
[2]*Division of Phoniatrics and Pediatric Audiology, Department of Otorhinolaryngology, Munich University Hospital, LMU, Munich, Germany*

(Received 23 January 2019; revised 11 June 2019; accepted 20 June 2019; published online 17 July 2019)

The estimation of formant frequencies from acoustic speech signals is mostly based on Linear Predictive Coding (LPC) algorithms. Since LPC is based on the source-filter model of speech production, the formant frequencies obtained are often implicitly regarded as those for an infinite glottal impedance, i.e., a closed glottis. However, previous studies have indicated that LPC-based formant estimates of vowels generated with a realistically varying glottal area may substantially differ from the resonances of the vocal tract with a closed glottis. In the present study, the deviation between closed-glottis resonances and LPC-estimated formants during phonation with different peak glottal areas has been systematically examined both using physical vocal tract models excited with a self-oscillating rubber model of the vocal folds, and by computer simulations of interacting source and filter models. Ten vocal tract resonators representing different vowels have been analyzed. The results showed that $F1$ increased with the peak area of the time-varying glottis, while $F2$ and $F3$ were not systematically affected. The effect of the peak glottal area on $F1$ was strongest for close-mid to close vowels, and more moderate for mid to open vowels.
 https://doi.org/10.1121/1.5116137

[JFL] Pages: 223–232

I. INTRODUCTION

The formants, or the resonances of the vocal tract, are fundamental parameters in many areas of speech science. According to Fant (1960), formants are the peaks of the spectrum or spectral envelope of speech sounds, and thus depend on both source and filter properties. On the other hand, the resonances are solely a property of the vocal system and correspond to the poles of the considered transfer function (typically assuming a closed glottis). In general, these two concepts should be held apart. Accordingly, different notations have been proposed for formant frequencies (F_1, F_2, F_3) and resonance frequencies (f_{R1}, f_{R2}, f_{R3}), which are also used in the present study (Titze *et al.*, 2015). In practice, the terms formant and vocal tract resonance are often used synonymously. The main reason is the assumption of independence of source and filter. The argument for this assumption is that the average glottal impedance during phonation is high compared to the input impedance of the vocal tract.

Nevertheless, multiple effects of source-filter interaction have been described that contradict a linear source-filter relationship (Childers and Wong, 1994). For example, the acoustic load of the vocal tract affects the skewing of the glottal flow pulses and can produce a "ripple" on the open phase of the flow pulses due to the first resonance. Furthermore, the energy of the first formant is partly dissipated through the glottis during the open phase of the glottal cycle, which increases its bandwidth. The shape of the vocal tract may also affect features of the vocal fold oscillation like f_0 and the phonation threshold pressure (Titze and Palaparthi, 2016).

With regard to formant frequencies, early calculations showed that a typical (finite) glottal impedance would change formant frequencies by just a few percent compared to the closed-glottis resonances, which is insignificant from a perceptual point of view (Badin and Fant, 1984; Flanagan, 1965). However, more recent experimental and numerical studies indicated that source-filter interaction might affect formants more strongly than previously thought. For example, Barney *et al.* (2007) used a driven mechanical shutter to simulate the time-varying glottal area in order to excite an attached rectangular duct with a constant cross-section of 17 × 17 mm^2 and examined how the Linear Predictive Coding (LPC)-based estimate of F_1 was affected by the peak glottal amplitude and the open quotient (OQ). They found that F_1 generally increased with increasing OQ and peak area, and for the highest OQ (80%) and peak area (68 mm^2), F_1 increased up to 40% compared to F_1 with a closed glottis. For a *static* glottal opening with an area of 51 mm^2, they found that the first resonance frequency was raised 13% above the value that would be expected for a closed glottis.

More recently, Uezu and Kaburagi (2017) used computer simulations to examine the effect of the maximum glottal width and the OQ during phonation on LPC-based estimates of F_1 and F_2 for five Japanese vowels. They corroborated the findings by Barney *et al.* (2007) by showing that F_1 tended to increase with an increasing maximum glottal width and OQ. In addition, they found that the relative increase of F_1 compared to the closed-glottis case depended on the vowel, and that the maximum increase of F_1 was

[a)]Electronic mail: peter.birkholz@tu-dresden.de

higher for /i/ (20%) than for /o/ (10%). For F_2, no consistent dependence on maximum glottal width or OQ was observed.

The present study extended the analysis of the effect of source-filter interaction on formant frequencies in the following ways:

- The effect of different peak glottal areas on the first three (instead of one or two) formants was analyzed.
- Ten (instead of one or five) different resonators were examined.
- A self-oscillating rubber-model of the vocal folds was used to excite the physical resonators instead of a driven-shutter model of the vocal folds.
- Both physical models and computer simulations were analyzed for the same set of resonators.

With the greater range of resonator shapes and a more realistic model of the vocal folds, our goal was to identify a possible systematic dependence of the degree of source-filter interaction on the vocal tract shape, and possibly identify systematic effects not only on F_1, but also on F_2 and F_3.

II. METHOD

In brief, we three-dimensionally (3D)-printed a set of ten acoustic resonators representing the vocal tract shape of different vowels, and created a self-oscillating physical model of the vocal folds using silicone rubber. To determine the formants of the resonators for different degrees of source-filter coupling, we used two techniques: (1) For the case of a completely closed glottis (no source-filter coupling), a sine-sweep technique was used to determine the transfer functions of the resonators, from which the resonance frequencies were obtained by peak picking. (2) To determine the formants for increasing degrees of source-filter coupling, the resonators were connected to the silicone vocal folds driven by successively higher values of subglottal pressure, namely 0.8, 1.0, 1.2, 1.4, and 1.6 kPa. Increasing the subglottal pressure led to an increased peak glottal area, as determined from high-speed camera recordings of the vocal folds, and hence to a stronger coupling. An LPC-based analysis of the vowel sounds generated was used to estimate the formants. In addition to these measurements, the physical setup has been reproduced and simulated using an aeroacoustic computer model of the vocal system. Details of the method are provided in the following subsections. The CAD files for the resonators, the CAD files for the physical vocal fold model, and the audio files of the vowels generated with the physical setup and the computer model are available at http://www.vocaltractlab.de/index.php?page=birkholz-supplements.

A. Physical setup and measurements

1. Creation of the physical vocal fold model

Previous research has shown that it is possible to create synthetic vocal folds from flexible silicone compounds that closely reproduce the oscillation patterns of real human vocal folds (Mendelsohn and Zhang, 2011; Murray and Thomson, 2012; Xuan and Zhang, 2014). For realistic oscillations, it is necessary to reproduce the layered structure of human vocal folds. Often, the vocal fold is simplified into a two-layer body-cover structure with a relatively stiff body layer and a softer cover layer. The body layer includes the muscular layer and the deep layer of the lamina propria, and the cover layer includes the intermediate and superficial lamina propria (Zhang, 2016). Synthetic vocal fold models can be created that reproduce this structure by casting silicone compounds with different Young's moduli (Murray and Thomson, 2012). Since the cover layer of such models needs to be very soft to be similar to the mucosa of human vocal folds, it is beneficial to add a third very thin layer of a stiffer silicone to mimic the epithelium of real human folds. This epithelium layer not only helps the soft cover layer to withstand the periodic impact during phonation, but also helps the model to generate a complete glottal closure along the anterior-posterior direction of the glottis leading to a stronger acoustic excitation at higher frequencies (Murray and Thomson, 2012; Xuan and Zhang, 2014).

Based on these considerations, a three-layer vocal fold model was created in this study. Figure 1(a) shows the central cross-section of the model, which is based on the M5 model by Scherer *et al.* (2001). The vocal fold length was set to 17 mm, which is typical for adult men (Zhang, 2016). The model consists of a body layer [dark gray area in Fig. 1(a)] surrounded by a cover layer with a thickness of 1 mm (light gray area), which is in turn covered by a thin epithelium layer with a thickness of about 70 μm (black line). All three layers were fabricated using addition-cure two-component silicone rubbers by Troll Factory Rainer Habekost e.K., Riede, Germany. The body layer was made of "TFC Silikon Kautschuk Typ 13, Shore 00" with a mixing ratio of 1:1:3 of part A, part B, and silicone oil (thinner), yielding a Young's modulus of 2.2 kPa (measured using a rheometer MCR 301 by Anton Paar). The cover layer was made of "TFC Silikon Kautschuk Typ 13, Shore 00" with a mixing ratio of 1:1:4 of part A, part B, and silicone oil, yielding a Young's modulus of 1.2 kPa (also measured using the rheometer). The epithelium layer was made of "TFC Silikon Kautschuk Typ 1, Shore 20" with a mixing ratio of 1:1 of part A and part B, yielding a Young's modulus of 560 kPa (according to the datasheet). The amount of silicone oil that was added to the mixtures for the body and the cover layers was experimentally determined for a good match of the Young's moduli of the silicone rubbers with real human vocal folds. The 2.2 kPa of the synthetic body layer corresponds well to the 2 kPa measured for the thyroarytenoid muscle by indentation (Chhetri *et al.*, 2011), and the Young's modulus of 1.2 kPa of the synthetic cover layer corresponds well to the 1 kPa measured in the transverse direction of the human vocal fold's lamina propria (Alipour and Vigmostad, 2012).

To manufacture the model, an aluminum frame for each vocal fold was milled in the precision mechanics workshop of the TU Dresden [left object in Fig. 1(b)]. The three silicone layers representing a vocal fold were successively attached to the corresponding frame and left on it. To make the body layer (i.e., the first layer), a computer-aided design (CAD) model of the *negative* of the body layer was designed and manufactured on a 3D printer (Ultimaker 3) using the water-soluble material polyvinyl alcohol (PVA), as shown in Fig. 1(b) (right object). The negative was glued to the

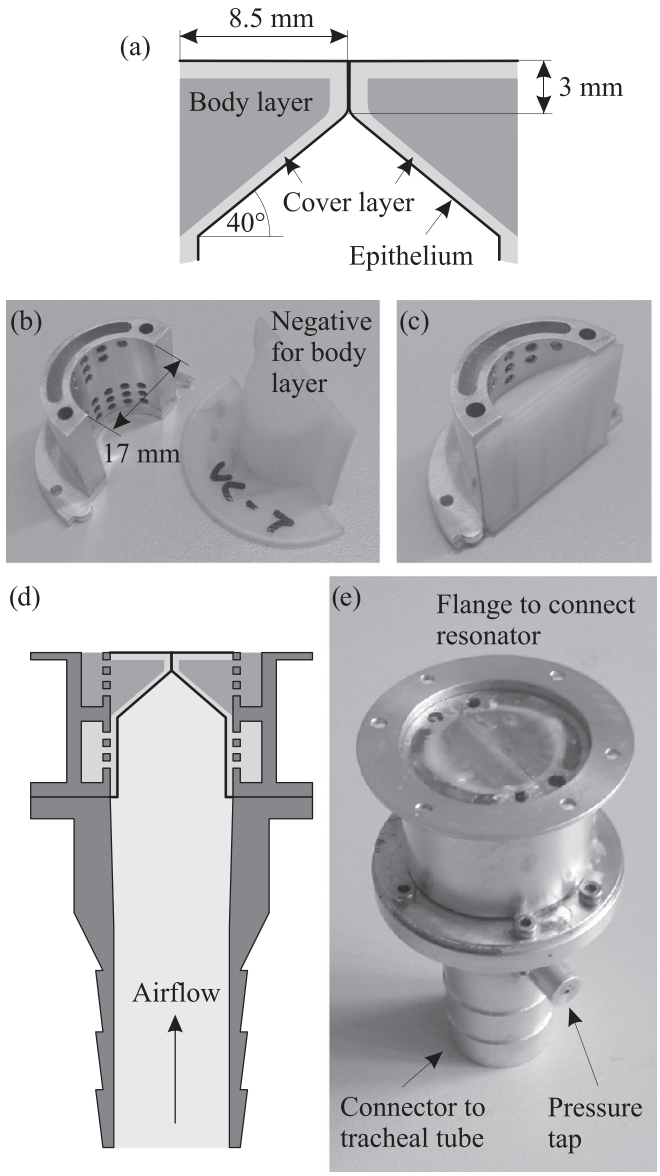

FIG. 1. The physical three-layer model of the vocal folds. (a) Central coronal cross-section through the silicone layers. (b) Geometries of the aluminum frame for one vocal fold and the (water-soluble) negative for the body layer. (c) The negative has been glued to the frame to form the mold for the body layer. (d), (e) Coronal cross-section and photo of the fully assembled model with connectors to the vocal tract resonators and to the tracheal tube.

aluminum frame to form a mold [Fig. 1(c)] into which the silicone mixture for the cover layer was poured. The silicone cured for 12 h at room temperature. Through the small holes in the aluminum frame, the liquid silicone also seeped into the compartments in the wall of the frame [visible as the curved slot in the frame in Fig. 1(c)] providing a stable connection between the cured body layer and the frame. After the silicone had cured, the PVA negative was completely dissolved in water. Analogously to the body layer, a second PVA negative was designed, 3D-printed, and glued to the frame to form the mold for the cover layer. The silicone mixture for the cover was poured into the mold and cured for 12 h at room temperature, after which the PVA was dissolved in water. After the PVA negative was dissolved, the model was tempered at 100 °C for 3 h. Finally, the epithelium layer was created by slowly pouring the corresponding silicone mixture over the cover layer. Most of the liquid silicone drained away and only a very thin film adhered to the cover layer; this was cured for 12 h at room temperature. To increase the thickness of the epithelium layer, this process was repeated once. A coronal cross-section and a photo of the complete vocal fold model with the connectors to the tracheal tube and to the vocal tract resonators are shown in Figs. 1(d) and 1(e).

2. Creation of physical resonators

All ten physical resonators were designed as straight tubes with circular cross-sections. For nine of the resonators, the lengths and area functions were adopted from the German vowels /a, e, i, o, u, ɛ, ø, y, ə/ defined in the software VocalTractLab 2.2 (VTL) (Birkholz, 2013). The acoustic side branches for the nasal cavity and the piriform fossae have been omitted for the physical resonators. The "epilaryngeal tube" of the physical resonators was slightly widened to allow a seamless connection with the physical vocal folds. Therefore, the radius of the tubes was made to change linearly from 8.5 mm at the glottal end to 3.5 mm at a position 1.8 cm above the glottis. Figure 2 shows the tube radii along the tube axes for all nine resonators. The dashed line for /a/ shows the radius in the epilarngeal area as defined in VTL, which is identical for all nine resonators. The tenth resonator was designed with a constant cross-sectional area of 2.27 cm^2 (corresponding to a radius of 8.5 mm) and a length of 16.54 cm (same as for the /ə/ resonator), i.e., as a cylindrical tube. The (constant) radius of this cylindrical resonator was equal to the radii of the other nine resonators at the glottal end, so that it was not necessary to taper its radius to fit on the physical vocal fold model. All resonators were designed with 3 mm thick walls and a flange at the glottal end for the connection with the physical vocal folds. The resonators were 3D-printed on an Ultimaker 3 printer using the material polylactide (PLA) with an infill ratio of 100%.

3. Measurement of resonances with a closed glottis

Using the method by Fleischer *et al.* (2018), the volume velocity transfer function was measured for each resonator, i.e., the complex ratio of the volume velocity at the lips divided by the volume velocity at the glottis. This transfer function corresponds to the case of a closed glottis and an ideal volume velocity source at the glottis. The method excites the resonators from outside (from 25 cm in front of the "mouth opening") with a loudspeaker (VISATON speaker, type FR10-8 Ohm in a custom-made cylindrical enclosure) emitting 100–10 000 Hz sine sweeps into the open end at the "lips," and measures the sound pressure $P_1(\omega)$ inside the resonators at the level of the glottis using a 1/4 in. measurement microphone (type MK301E/MV310, www.microtechgefell.de). The microphone was positioned such that its membrane was flush with the upper surface of the "vocal folds." For the same external sine sweep excitation, a second (reference) measurement $P_2(\omega)$ was performed with a probe microphone (ER-7C, www.etymotic.com) positioned right in front of the *closed* mouth opening. The mouth openings of the models were closed with a flat disk of modeling

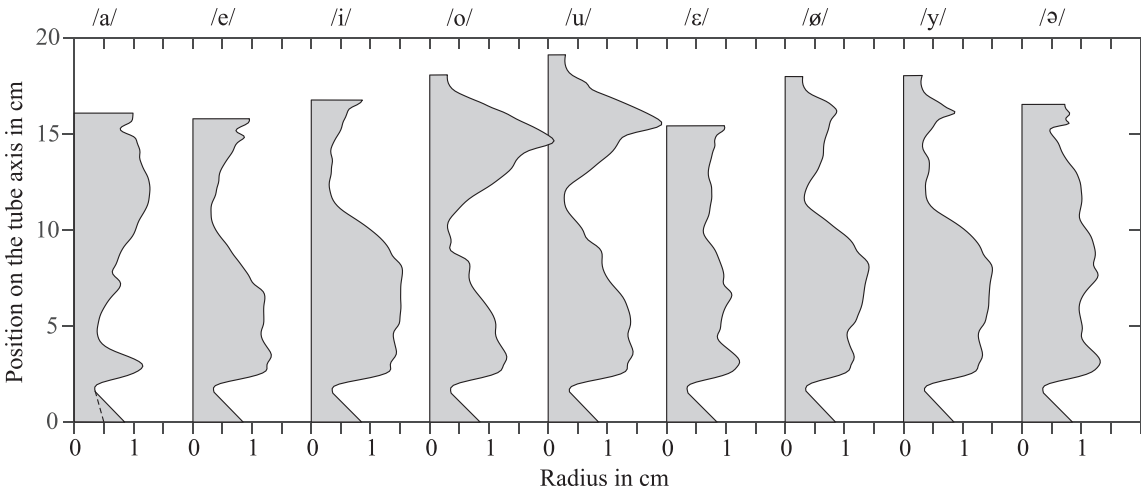

FIG. 2. Radii of the nine vowel resonators as a function of the position along the tube axis (the tenth cylindrical tube is not shown). The dashed line in the function for /a/ indicates the radius of the original cross sections in VTL.

clay with a thickness of about 5 mm. As shown by Fleischer *et al.* (2018), the ratio $P_1(\omega)/P_2(\omega)$ corresponds exactly to the closed-glottis volume velocity transfer function, and can be measured with a high signal-to-noise ratio with this method. Both the emitted sine sweep signal and the recorded microphone signals were sampled at 44 100 Hz and 16 bit. In contrast to the similar measurement method by Delvaux and Howard (2014), we did not explicitly remove potential harmonic distortions from the recorded signals. In the used setup, the risk of harmonic distortions mainly increases with the volume of the loudspeaker that emits the excitation signal. Here, the volume of the loudspeaker was carefully adjusted to be low enough such that its response was essentially linear (as checked by its response to pure sine signals).

The transfer functions were measured with a spectral resolution of 1 Hz, and the first three resonance frequencies of each resonator were determined by picking the peaks of the magnitude of the transfer function. All measurements were performed in the large anechoic chamber of the TU Dresden at a temperature of 21 °C and air humidity of 45%. Figure 3 shows examples of the transfer functions for the /a/ and /i/ resonators.

4. Measurement of generated sounds and peak glottal areas using the physical vocal folds and resonators

To measure the vowel sounds generated by the physical vocal fold model exciting the physical resonators, and to determine the peak glottal area for different subglottal pressures, the experimental setup shown in Fig. 4 was used. The "subglottal" system consisted of a compressor (air blower Medo LA 100A by Nitto Kohki), which was connected to an expansion chamber with a 60 cm long hose, and another 200 cm long hose that connected the expansion chamber to the vocal fold model.[1] The expansion chamber had the shape of a box with a volume of 30 × 30 × 50 cm and was made of 2 cm thick wooden plates, covered inside by sound absorbing foam (NOMA ACOUSTIC 25 mm by NMC). As the pressure output of the compressor was not adjustable, we controlled the subglottal pressure using a manual one-inch shut-off valve connected to the expansion chamber. The valve, the expansion chamber, and the compressor were located in a separate soundproof cabin to prevent the noise of the compressor or the airstream leaving the valve to interfere with the acoustic measurements of the resonators. The subglottal pressure was monitored using a pressure sensor

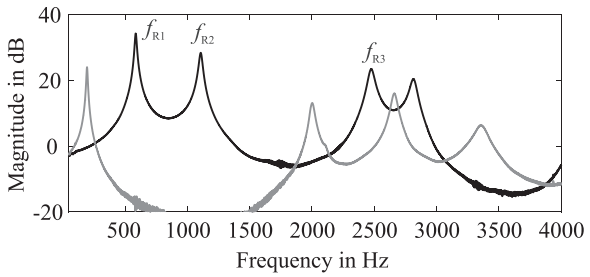

FIG. 3. Measured vocal tract transfer functions of the 3D-printed resonators for /a/ (black) and /i/ (gray) with a closed glottis using the method by Fleischer *et al.* (2018). The resonance frequencies f_{R1}, f_{R2}, and f_{R3} were determined as the peaks of the spectral magnitude, as shown for /a/.

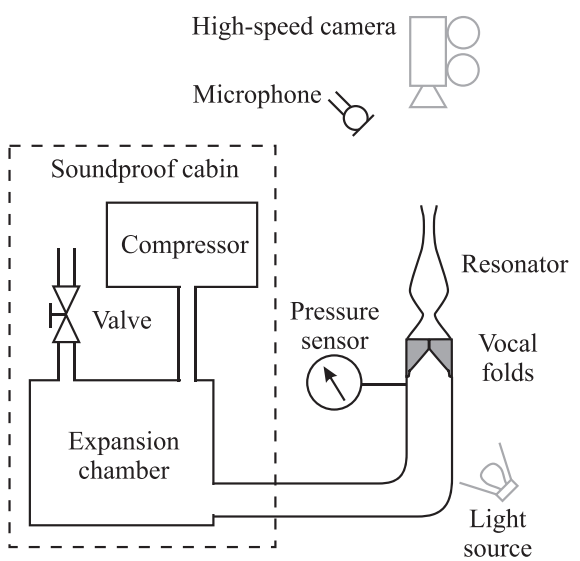

FIG. 4. Setup for the measurement of the sound produced by the ten physical resonators excited by the silicone vocal folds. The same setup was used to measure the peak glottal area as a function of subglottal pressure.

(pressure transmitter DMU4 by Kalinsky Sensor Elektronik & Co. KG, Germany) connected to a pressure tap below the glottis (see Fig. 1). The resonators were firmly screwed on top of the vocal fold model. The audio signal emitted from the resonators was recorded using a measurement microphone (MV210 with microphone capsule MK250 by RFT VEB Mikrofontechnik) connected to a USB audio interface (E-MU 0404-USB by Creative Professional). The USB audio interface was connected to a standard laptop computer with Windows 8.1 as operating system, where the audio data were recorded with a sampling rate of 44.1 kHz and 16 bit quantization using the software Audacity 2.0.2. The microphone was positioned 30 cm above the "lips" of the resonators and about 10 cm sideways of the tube axis to prevent the airflow from directly impinging on the microphone membrane.

Using this setup, the acoustic output of each of the ten resonators was recorded for five subglottal pressure values: 0.8, 1.0, 1.2, 1.4, and 1.6 kPa. The pressure of 0.8 kPa was the lowest value required by the vocal fold model to oscillate in a stable manner with all resonators. For each pressure setting and resonator, the formants were estimated as described in Sec. II C

To find out how the vocal fold oscillation was affected by different subglottal pressures, a strong light source and a high-speed camera (MQ013CG-ON by XIMEA GmbH with 1/2 in. Format Mega Pixel lens) located 15 cm above the resonators were used to film the oscillating vocal folds (gray symbols in Fig. 4). To capture the vocal fold oscillations for the case of a connected acoustic load (resonator), the camera had to be able to "see" the vocal folds through the resonator. The complete glottis could only be seen through the cylindrical tube, so only that resonator was used here. Furthermore, it was not possible to illuminate the vocal folds strongly enough from above through the resonator (because of the shadowing of the camera). Therefore, the light source illuminated the glottis from below through the semi-transparent hose connecting the expansion chamber and the glottis. In this way, the glottis appeared as a bright area in the camera images.

Using this modified setup, high-speed films of the oscillating vocal folds were taken for the same five subglottal pressure settings used for the acoustic recordings. The films were captured with the software Ximea CamTool 4.14 by XIMEA GmbH, Germany, running on a laptop computer, to which the camera was connected. The frame rate was 800 frames/s and the size of the frames was 256 × 256 pixels. All frames were converted to 8-bit grayscale images and the grayscale was inverted so that the glottis was dark and the surrounding structures bright. The open-source software GlottalImageExplorer 1.0 (Birkholz, 2016) was used to extract the glottal areas in the individual frames. Because of the slight undersampling of the oscillations (800 frames/s at an f_0 of about 90 Hz), the peak glottal area for each of the five pressure settings was determined as the maximum glottal area detected in a sequence of 800 frames. Figure 5 shows that there was a linear relationship between subglottal pressure and the peak glottal area, with the peak area increasing from 11 mm^2 at 0.8 kPa to 60 mm^2 at 1.6 kPa. Based on the study by Titze and Palaparthi (2016), we assumed that this

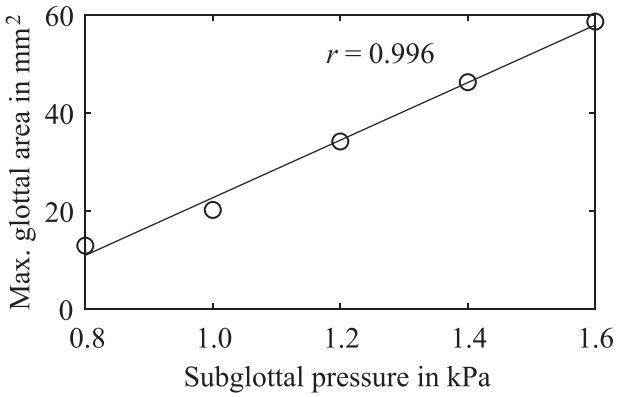

FIG. 5. Measurements of the peak glottal area of the physical vocal fold model exciting the cylindrical resonator with different subglottal pressures (circles). The solid line is the regression line showing that both quantities are highly correlated.

relationship holds for all ten resonators. Titze and Palaparthi found that, for a fixed subglottal pressure, the peak glottal area only varied between 14.1 and 17.6 mm^2 across a set of 11 vowels, suggesting that the peak glottal area is roughly independent from the vowel.

Besides the peak glottal area, the OQ of the vocal folds was calculated for each pressure value as $OQ = N_{open}/N_{total}$, where $N_{open}$was the number of frames where the glottal area was greater than a small threshold (5% of the maximum glottal area), and N_{total}= 4000 (= 5 s) was the total number of consecutive frames analyzed for each pressure. The obtained open quotients were 0.34, 0.36, 0.36, 0.39, and 0.38 for the lowest to the highest pressure values, and hence roughly constant.

B. Computer simulation of the physical setup

1. Acoustic model of the vocal system

To verify the results obtained with the physical models of the vocal tract and the vocal folds, the physical system has been simulated with a computer model based on an extension of the software VocalTractLab 2.2 (Birkholz, 2017). The vocal system was modeled in terms of a discrete, one-dimensional acoustic tube model, i.e., a concatenation of a number of short cylindrical tube sections that represent the combined area function of the trachea, the glottis, and the vocal tract (65 tube sections in total) (Birkholz, 2005). To conform to the physical setup, all side cavities (i.e., the nasal cavity and the piriform fossae) were omitted in the simulation. The tube model was represented in terms of an inhomogeneous acoustic transmission line with lumped elements, as illustrated in Fig. 6, terminated with a radiation impedance at the lips, and accounting for nonlinear losses due to turbulence at the glottis (Birkholz and Jackèl, 2004; Birkholz *et al.*, 2007). As the physical resonators were made of hard plastic, an infinite wall impedance was used in the simulations, and sound radiation from the "skin" was omitted. The area functions of the vocal tract used in the simulations were the same as those of the resonators used with the physical setup (apart from a small deviation at the "epilaryngeal tube," as described in Sec. II A 2).

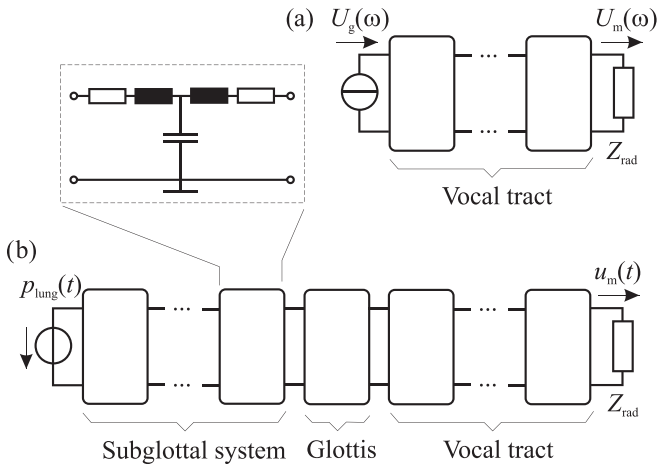

FIG. 6. (a) Acoustic network of the vocal tract driven by a volume velocity source at the glottal end. (b) Acoustic network of the vocal system including the subglottal system, the glottis, and the vocal tract, driven by a pulmonary pressure source.

2. Calculation of resonances with a closed glottis

To determine the resonances of the ten vocal tract shapes for the case of a closed glottis, the corresponding volume velocity transfer functions $H(\omega) = U_m(\omega)/U_g(\omega)$ were calculated in the frequency domain according to the acoustic network in Fig. 6(a), where $U_g(\omega)$ is the volume velocity through the glottis, $U_m(\omega)$ is the volume velocity through the lips, and ω is the angular frequency. The calculations were based on the chain matrices of the individual tube sections with a frequency resolution of 1 Hz (Birkholz and Jackèl, 2004). For each vocal tract shape, the first three resonance frequencies were identified as the peaks in the corresponding transfer function.

3. Synthesis of sustained vowels

To simulate the effect of source-filter interaction for the ten vocal tract shapes, the network in Fig. 6(b) was used to calculate the acoustic field in the discrete time domain. This network represents the complete vocal system comprising the trachea, the glottis, and the vocal tract, driven by a "lung" pressure source. The glottis was modeled as a single tube section with a length of 3 mm and a cross-sectional area $A_g(t)$ that was modulated according to the equation

$$A_g(t) = \max\{A_{peak} \cdot \sin(2\pi f_0 t),\ 0\},$$

where the fundamental frequency f_0 was set to a constant value of 90 Hz (to correspond to the f_0 of the physical vocal fold model), and the peak glottal area $A_{peak}$was calculated as a linear function of the simulated lung pressure p_{lung} according to

$$A_{peak} = 11.27\,\text{mm}^2 + (p_{lung} - 800\,\text{Pa}) \cdot 0.0606\,\text{mm}^2/\text{Pa}.$$

This equation corresponds to the regression line between subglottal pressure and peak glottal area measured with the physical setup (see Fig. 5). Aerodynamically, it was assumed that the glottal flow obeys Bernoulli's law at the entrance of the glottis, and that the dynamic pressure in the glottis is completely lost at the glottal exit (no pressure recovery). The discrete-time simulation of the acoustic field in the vocal system was performed with a sampling rate of 44100 Hz, and the radiated speech signal was approximated as the temporal derivative of the volume velocity through the lips (Birkholz and Jackèl, 2004). For each of the ten vocal tract shapes in combination with each of the five lung pressure values (the same five pressure values as in the physical setup), the radiated sound pressure was calculated for 1 s simulation time, downsampled to 22 050 Hz, and saved as a 16 bit WAV file.

In addition, we also simulated the speech waveforms for the different vocal tract shapes for the case of a closed glottal end, at which a train of glottal flow pulses was injected by means of an "ideal" flow source with an infinite impedance [corresponding to Fig. 6(a)]. This case corresponds to perfect source-filter separation and allowed us to assess the similarity of LPC-based formant estimates of these simulated closed-glottis speech samples with the corresponding resonance frequencies obtained from the calculated closed-glottis transfer functions. The train of glottal flow pulses was created using the flow pulse model by Fant *et al.* (1985) with a fundamental frequency of 90 Hz. The shape of the glottal flow waveform was created with an open quotient of 0.5, a skewing quotient of 3.0, and a tilt parameter of 0.02.

C. LPC-based estimation of formants

For all vowel sounds generated with the physical models and the computer simulations, the first three formant frequencies were estimated with the formant tracker implemented in the software Praat version 6.0.28 (Boersma and Weenik, 2017). This algorithm computes the LPC coefficients on a frame-by-frame basis with the algorithm by Burg, as given by Childers (1978), and translates these into the formant frequencies. Two problems of standard LPC-based formant analysis are that the formant estimates are biased by the pitch harmonics (Shadle *et al.*, 2016; Vallabha and Tuller, 2002), and that the optimal number of LPC poles depends on the speaker and even on the speech sound (Kathiresan *et al.*, 2017). The first problem is mainly relevant for high-pitched female voices, for which specific variants of the linear prediction (LP) analysis were developed (Alku *et al.*, 2013; Gowda *et al.*, 2017). In the present study, the f_0 of the generated sounds was around 90 or 135 Hz (see below), so that the standard LP analysis in the Praat software was deemed sufficient.

With regard to the second problem, the standard practice is to adjust the number of poles manually. The criteria used for the selection of the number of poles for the individual speech samples were based on the proposals by Kathiresan *et al.* (2017). For each vowel sample (each of 1 s duration), we first varied the number of poles, checking that the first three formant estimates were in the expected frequency ranges for the corresponding vowel, and that there were no "spurious" formants. From the cases passing these tests, we selected the number of poles for which the variance of the formant estimates were smallest across all frames. Hence, an

TABLE I. Parameters used for the LPC-based formant analysis with the software Praat.

Parameter	Value
Maximum formant (Hz)	5500.0
Number of formants	Manually adjusted
Window length (s)	0.1
Dynamic range (dB)	30.0
Dot size (mm)	1.0
Method	Burg
Pre-emphasis from (Hz)	50.0
Time step strategy	Automatic

individual optimal number of LPC poles was used for each vowel sample. The length of the LPC analysis window was always set to 100 ms. Apart from the number of poles and the window length, all analysis parameters were set to the default values in Praat (see Table I). For each sample, the values for F_1, F_2, and F_3 were determined as the average values across all frames in the 0.5 s in the middle of the sample. In a similar way, the average f_0 was determined for each sample using the pitch tracker in Praat.

For some of the samples measured with the physical setup, it was not possible to obtain reliable estimates for f_0 or individual formants using the above strategy. The corresponding values have been marked as invalid and excluded from the further analysis (e.g., F_3 of /ø/). In total, from the 50 samples (10 resonators times 5 pressure values), we failed to get reliable estimates for f_0 in four cases, for F_2 in one case, and for F_3 in 11 cases (all samples allowed the reliable estimation of F_1). The problems were exclusively observed for the five resonators that caused the vocal folds to oscillate at the (high) f_0 of around 135 Hz (the other five resonators caused an f_0 of around 90 Hz; for a discussion of this phenomenon see Sec. III). Furthermore, the problems with the determination of f_0 occurred only for low pressures of 800 or 1000 Pa, where the vocal fold oscillation was not (yet) sufficiently periodic. This indicates that the vocal folds require a higher lung pressure for higher f_0 values to generate a stable and periodic oscillation. The difficulty of obtaining reliable formant frequencies with LPC-based analysis for higher pitches is in line with previous findings (Alku *et al.*, 2013).

III. RESULTS AND DISCUSSION

The measured f_0 values and formant frequencies of the vowel samples generated with the physical models of the vocal folds and the resonators are summarized in Fig. 7. With regard to f_0, the five resonators for /i, a, ɛ, ə/ and the uniform cylinder caused the vocal folds to oscillate at a frequency of around 90 Hz, and the other five resonators for /e, o, u, ø, y/ caused the vocal folds to oscillate at around 135 Hz. Hence, the acoustic properties of the resonators had a distinct effect on the oscillation of the vocal folds. The grouping of the measured frequencies around the two values of 90 and 135 Hz was a surprise, which is hard to explain solely on the basis of the acoustic input impedance of the resonators. We assume that the "natural" frequency of the vocal folds was 90 Hz, and the "jump" to 135 Hz was the consequence of the entrainment with a subglottal resonance in combination with certain supraglottal load conditions. The resonance frequencies of the subglottal system of the physical setup were mainly determined by the length of the hose that connected the expansion chamber (open end) and the vocal folds (closed end). Given the length of the hose of $L = 2$ m and the sound velocity $c = 350$ m/s, the resonance frequencies were approximately $f_{Rn} = (2n-1)\cdot c/(4L) = (2n-1)\cdot 43.75$ Hz, where $f_{R2} = 131$ Hz might have caused the entrainment.

With regard to the LPC-based formant estimates in Fig. 7, F_1 tends to increase with increasing pressure, and accordingly with increasing peak glottal area. The frequency values denoted as "c.g." correspond to the resonance frequencies f_{R1} of the resonators with a completely closed glottis. Figure 8(a)

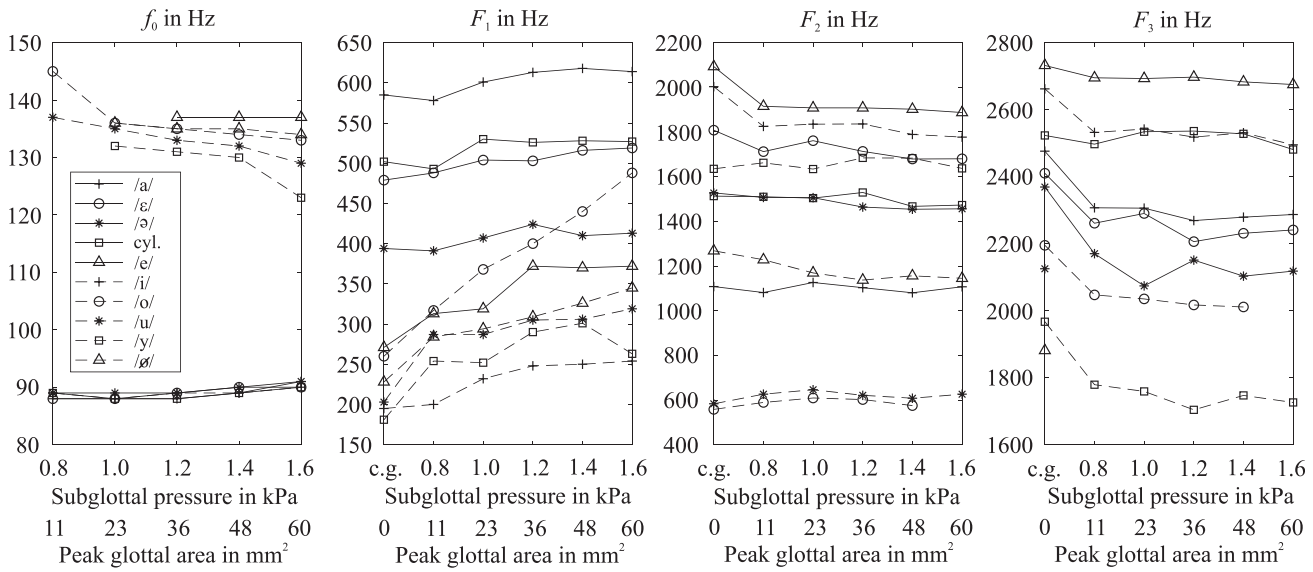

FIG. 7. Estimates of f_0, F_1, F_2, and F_3 for the vowel samples generated with the physical models of the vocal folds and the vocal tract for different subglottal pressures, which are correlated with the peak glottal areas. The abbreviation "c.g." stands for "closed glottis." The corresponding formant values represent the resonance frequencies $f_{R1/2/3}$ for an infinite glottal impedance.

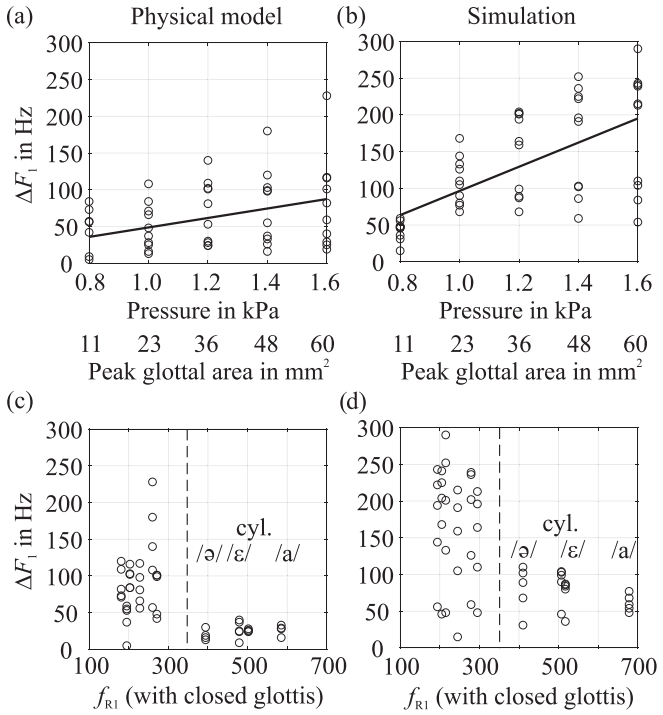

FIG. 8. (a) and (b) $\Delta F_1 = F_1 - f_{R1}$ as a function of the subglottal pressure for the measurements with the physical models in (a) and the simulations in (b), pooled across all resonators (f_{R1} is the first closed-glottis resonance frequency). (c) and (d) ΔF_1 over the closed-glottis resonance frequency f_{R1} for the measurements with the physical models in (a) and the simulations in (b), pooled across all subglottal pressures. The vertical dashed lines separate the resonators for which F_1 is only weakly affected by source-filter interaction ($f_{R1} > 350$ Hz, i.e., the resonators for /ə, ɛ, a/ and the cylindrical tube) from the resonators for which F_1 is strongly affected ($f_{R1} \leq 350$ Hz, i.e., all other resonators). For the weakly affected resonators, the data points have been labeled with the according phonetic symbols. The abbreviation "cyl." stands for the cylindrical tube.

emphasizes the positive correlation between F_1 and subglottal pressure by showing the difference ΔF_1 between the LPC-based F_1 estimate and f_{R1} as a function of the subglottal pressure across all resonators. The Pearson correlation coefficient for these data points is $r = 0.379$, and the correlation is significant ($p = 0.0066$).[2] Hence, the increase of F_1 with increasing peak glottal area is consistent across the different resonators.

However, the rate of increase of F_1 differs between the resonators. Figure 8(c) shows ΔF_1 over f_{R1} for all resonators and pressure values, indicating that there is one group of resonators for which F_1 strongly increases with increasing peak glottal area, and another group for which F_1 increases only slightly with increasing peak glottal area. The group that was strongly affected by source-filter interaction contains the close-mid to close vowels /e, i, o, u, ø, y/, which are characterized by a rather low f_{R1}, and the less affected group contains the mid to open vowels /ə, ɛ, a/ and the cylindrical tube. A boundary that separates the two groups was drawn as a vertical dashed line at $f_{R1} = 350$ Hz in Fig. 8(c). According to Fig. 7, the /o/ resonator was most strongly affected by the subglottal pressure, mainly because the curve for F_1 does not flatten towards higher pressure values as it is the case for the other resonators. We double-checked the resonator tube, the measured audio signals, and the formant analysis, but could not find an obvious explanation for this behavior.

In contrast to F_1, the second and third formant frequencies did not consistently change as a function of the peak glottal area (Fig. 7). When $\Delta F_2 = F_2 - f_{R2}$ and $\Delta F_3 = F_3 - f_{R3}$ are defined analogously to ΔF_1, the Pearson correlation coefficients between the subglottal pressure and $\Delta F_{1/2}$ across all resonators are $r = -0.177$ and $r = -0.125$, respectively, but neither correlation is significant ($p = 0.223$ and $p = 0.45$ for ΔF_1 and ΔF_2, respectively).

The formant frequencies obtained from the computer simulations are summarized in Fig. 9. Here, f_0 is not shown because it was explicitly set to 90 Hz in the simulations. Generally, the data from the simulations were similar to the measured data. There was a significant positive correlation between subglottal pressure and ΔF_1 across all resonators [$r = 0.638$ and $p = 6.2 \times 10^{-7}$; see Fig. 8(b)], but no significant correlation for ΔF_2 and ΔF_3 ($r = 0.174$ and $p = 0.227$ for F_2, and $r = 0.247$ and $p = 0.084$ for F_3). This confirms that an increasing degree of source-filter interaction due to an increasing peak glottal area systematically affects F_1 across all vowels, while F_2 and F_3 are not systematically affected. Figure 8(d) shows furthermore that the simulated vowels can be grouped just like the physical resonators with respect to the strength of the effect.

However, in the simulation results, the average increase of F_1 with increasing peak glottal area was stronger than for the physical models, which is also evident from the slope of the regression line in Fig. 8(b). Related to this, all resonance frequencies f_{R1}, f_{R2}, and f_{R3} were consistently higher for the simulated resonators than for the physical models. Across all resonators, the average difference was 7.1% for f_{R1}, 9% for f_{R2}, and 16.3% for f_{R3}. We suppose that this was mainly caused by the different geometries of the epilaryngeal tubes, which have a cross-sectional area of 80 mm^2 (at the base) in the simulated resonators, but 226 mm^2 in the 3D-printed resonators (see Sec. II A 2). To check the plausibility of this assumption, we adjusted (i.e., widened) the epilaryngeal tube geometry of the simulated vocal tract shapes of the vowels /a/, /i/, and /u/ to correspond to that of the 3D-printed physical resonators. For all three vowels, f_{R1}, f_{R2}, and f_{R3} decreased by up to 13% compared to the simulations with the narrower epilaryngeal tube, which supports the assumption.

A detailed comparison of the curves in Figs. 7 and 9 shows a couple of other minor differences between the results of the measurements and the simulations. Besides the deviations of the geometry of the epilaryngeal tubes, the following factors may be responsible for this:

- The simulated vocal tract tubes approximate the continuous area functions of the physical resonators with piecewise constant area functions (the cylindrical tube sections).
- The sound velocity was set to 350 m/s in the simulations (default in VTL for a peripheral body temperature of 31 °C) while it was approximately 344 m/s during the measurements with the physical models (for a room temperature of 21 °C). This may partly explain the higher formant frequencies of the simulated resonators. For a cylindrical tube, the higher sound velocity in the simulation would explain a formant frequency increase of 1.7%.

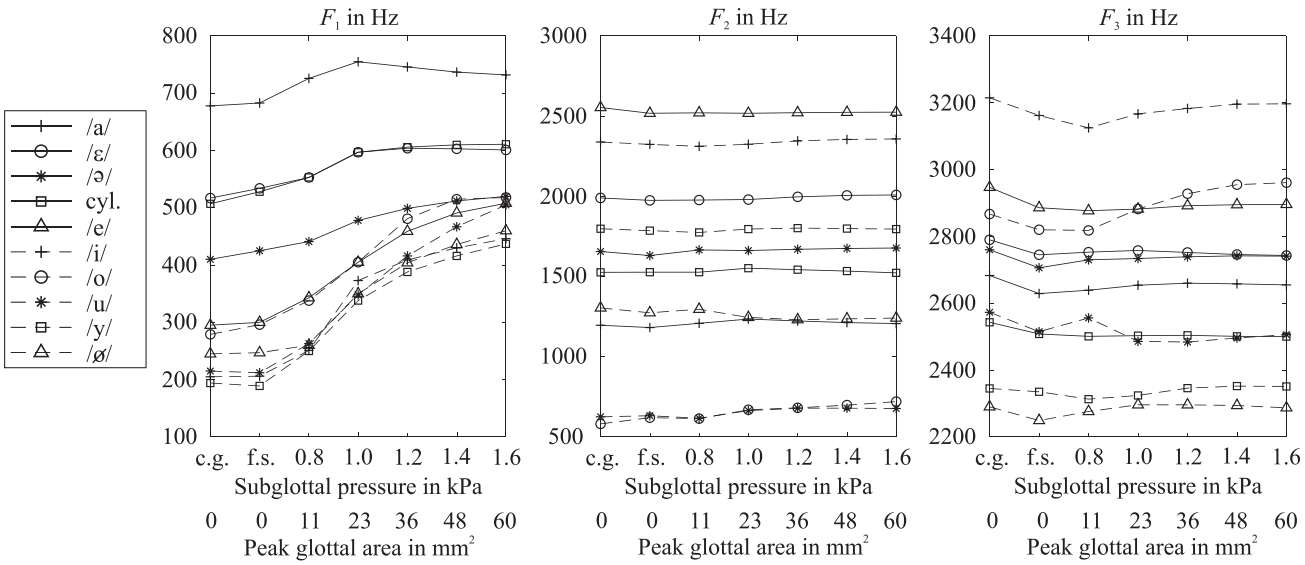

FIG. 9. Estimates of F_1, F_2, and F_3 for the vowel samples generated with the computer models of the vocal folds and vocal tract for different subglottal pressures, which are correlated with the peak glottal areas. The abbreviation "c.g." stands for "closed glottis." The corresponding formant values represent the resonance frequencies $f_{R1/2/3}$ for an infinite glottal impedance. "f.s." stands for the LPC-based formant estimates from the simulations with a flow pulse source and a closed glottis.

- The representation of vocal fold oscillation in the simulations was significantly simplified compared to the oscillation of the physical vocal fold model.

Despite these difference, the major effects of source-filter interaction on the formant frequencies were the same in both the simulations and the physical measurements.

To assess the general reliability of the strategy for LPC-based formant estimation used here (see Sec. II C), we included the time-domain simulations of the vowels excited with an "ideal" flow pulse source at the glottis, which was simulated with an infinite impedance in this case (closed glottis). The estimated formant frequencies for these samples are shown in Fig. 9, where they are denoted as "f.s." (flow source). These formants are very similar to the closed-glottis resonances determined from the corresponding volume-velocity transfer functions for all vowels (denoted as "c.g." in Fig. 9). Hence, the LPC-based formant analysis strategy used here seems generally capable of giving accurate results for the kind of audio samples used here.

The results of the present study generally confirm the results of previous studies saying that there is an effect due to source-filter interaction on F_1 (Barney *et al.*, 2007; Uezu and Kaburagi, 2017). However, with the comparatively realistic physical models and computer models of vowel generation used here, it was demonstrated that, depending on the vowel, the effect of source-filter interaction on F_1 can be much greater than previously reported. In general, it was found that vowels with a low first closed-glottis resonance frequency ($f_{R1} < 350$ Hz) were considerably more affected than vowels with a higher f_{R1}. The vowel /o/ showed the strongest relative increase of F_1 from about 250 Hz for the closed-glottis case to about 500 Hz for a peak glottal area of 0.6 cm^2. Given the threshold of 14 Hz for the perceptual discrimination of F_1 values, even small changes of the peak glottal area during phonation may cause perceptual effects for some vowels (Kewley-Port and Watson, 1994). Why the peak glottal area affects the close vowels more than the open vowels is not clear yet. Possibly, low-frequency resonances (i.e., f_{R1} of close vowels) might generally react more strongly to the (low-frequency) periodic changes of the acoustic boundary condition at the glottis than higher-frequency resonances. However, more research is needed to explain this effect in detail.

IV. CONCLUSIONS

In conclusion, when LPC-based estimates of F_1 are compared across speaker groups or different speaking or singing conditions, the effect of the peak glottal area on F_1 should be taken into account. For example, for shouted or Lombard speech, it has been found that F_1 is systematically higher than for "normal" speech (Summers *et al.*, 1988). Given the new findings, this increase may be due not only to changes in the vocal tract shape, but possibly also to the greater glottal peak area during shouting. Furthermore, when LPC-based formant estimates of vowels of natural speech are used as reference values for vowels in parametric speech synthesizers based on a source-filter model, it should be considered that the values determined for F_1 do not represent the closed-glottis resonances f_{R1}, but are values for one particular voice source setting.

Finally, more research should be done to establish novel formant tracking methods that can discriminate the vocal tract resonances during the open and closed phases of individual glottal cycles during phonation. A promising approach for this is based on the relatively new concept of the reassigned spectrogram (Fulop, 2007; Gardner and Magnasco, 2006; Shadle *et al.*, 2016).

ACKNOWLEDGMENTS

This research was part of the project "Sprechmaschine" (speaking machine) funded by the German Federal Ministry

of Education and Research (BMBF) with the support code 01UQ1601A. We are also grateful to Paavo Alku and Manu Airaksinen for providing their Matlab code for the AME-WLP algorithm for formant estimation to check whether this algorithm would give different formant estimates than the formant tracker built into Praat for our signals (which was not the case).

[1]Note that the 200 cm long "trachea" has resonance frequencies much lower than those of a normal human-length trachea. This most probably had an entrainment effect on the f_0 of the vocal fold model, as discussed in Sec. III.

[2]All correlation coefficients and p values reported here were calculated with the function corrcoef in Matlab R2017b, and the significance level was assumed to be $\alpha = 0.05$.

Alipour, F., and Vigmostad, S. (**2012**). "Measurement of vocal folds elastic properties for continuum modeling," J. Voice **26**(6), 816.e21–816.e29.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., and Story, B. H. (**2013**). "Formant frequency estimation of high-pitched vowels using weighted linear prediction," J. Acoust. Soc. Am. **134**(2), 1295–1313.

Badin, P., and Fant, G. (**1984**). "Notes on vocal tract computation," STL-QPSR **2–3**, 53–108.

Barney, A., De Stefano, A., and Henrich, N. (**2007**). "The effect of glottal opening on the acoustic response of the vocal tract," Acta Acust. united Ac. **93**(6), 1046–1056.

Birkholz, P. (**2005**). *3D-Artikulatorische Sprachsynthese* (Logos Verlag, Berlin).

Birkholz, P. (**2013**). "Modeling consonant-vowel coarticulation for articulatory speech synthesis," PLoS One **8**(4), e60603.

Birkholz, P. (**2016**). "GlottalImageExplorer—An open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*, edited by O. Jokisch (TUDPress, Dresden, Germany), pp. 39–44.

Birkholz, P. (**2017**). "VocalTractLab [computer software]" http://www.vocaltractlab.de (Last viewed 2 July 2019).

Birkholz, P., and Jackèl, D. (**2004**). "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Proceedings of Interspeech 2004-ICSLP*, October 4–8, Jeju, Korea, pp. 1125–1128.

Birkholz, P., Jackèl, D., and Kröger, B. J. (**2007**). "Simulation of losses due to turbulence in the time-varying vocal system," IEEE Trans. Audio Speech Lang. Process. **15**(4), 1218–1226.

Boersma, P., and Weenik, D. (**2017**). "Praat: Doing phonetics by computer [computer program]" http://www.praat.org/ (Last viewed 2 July 2019).

Chhetri, D. K., Zhang, Z., and Neubauer, J. (**2011**). "Measurement of Young's modulus of vocal folds by indentation," J. Voice **25**(1), 1–7.

Childers, D. G. (**1978**). *Modern Spectrum Analysis* (IEEE Computer Society Press, Piscataway, NJ).

Childers, D. G., and Wong, C.-F. (**1994**). "Measuring and modeling vocal source-tract interaction," IEEE Trans. Biomed. Eng. **41**(7), 663–671.

Delvaux, B., and Howard, D. (**2014**). "A new method to explore the spectral impact of the piriform fossae on the singing voice: Benchmarking using MRI-based 3D-printed vocal tracts," PLoS One **9**(7), 1–15.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, the Netherlands).

Fant, G., Liljencrants, J., and guang Lin, Q. (**1985**). "A four-parameter model of glottal flow," STL-QPSR **4**, 1–13.

Flanagan, J. L. (**1965**). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Berlin).

Fleischer, M., Mainka, A., Kürbis, S., and Birkholz, P. (**2018**). "How to precisely measure the volume velocity transfer function of physical vocal tract models by external excitation," PLoS One **13**(3), 1–16.

Fulop, S. A. (**2007**). "Phonetic applications of the time-corrected instantaneous frequency spectrogram," Phonetica **64**(4), 237–262.

Gardner, T. J., and Magnasco, M. O. (**2006**). "Sparse time-frequency representations," Proc. Natl. Acad. Sci. **103**(16), 6094–6099.

Gowda, D., Airaksinen, M., and Alku, P. (**2017**). "Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation," J. Acoust. Soc. Am. **142**(3), 1542–1553.

Kathiresan, T., Maurer, D., Suter, H., and Dellwo, V. (**2017**). "Enhancing the objectivity of interactive formant estimation: Introducing Euclidean distance measure and numerical conditions for numbers and frequency ranges of formants," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, edited by J. Trouvain, I. Steiner, and B. Möbius (TUDPress, Dresden, Germany), pp. 130–137.

Kewley-Port, D., and Watson, C. S. (**1994**). "Formant-frequency discrimination for isolated English vowels," J. Acoust. Soc. Am. **95**(1), 485–496.

Mendelsohn, A. H., and Zhang, Z. (**2011**). "Phonation threshold pressure and onset frequency in a two-layer physical model of the vocal folds," J. Acoust. Soc. Am. **130**(5), 2961–2968.

Murray, P. R., and Thomson, S. L. (**2012**). "Vibratory responses of synthetic, self-oscillating vocal fold models," J. Acoust. Soc. Am. **132**(5), 3428–3438.

Scherer, R. C., Shinwari, D., De Witt, K. J., Zhang, C., Kucinschi, B. R., and Afjeh, A. A. (**2001**). "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees," J. Acoust. Soc. Am. **109**(4), 1616–1630.

Shadle, C. H., Nam, H., and Whalen, D. (**2016**). "Comparing measurement errors for formants in synthetic and natural vowels," J. Acoust. Soc. Am. **139**(2), 713–727.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S., and Wolfe, J. (**2015**). "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," J. Acoust. Soc. Am. **137**(5), 3005–3007.

Titze, I. R., and Palaparthi, A. (**2016**). "Sensitivity of source–filter interaction to specific vocal tract shapes," IEEE/ACM Trans. Audio Speech Lang. Process. **24**(12), 2507–2515.

Uezu, Y., and Kaburagi, T. (**2017**). "A simulation study on the effect of glottal boundary conditions on vocal tract formants," in *Proceedings of Interspeech 2017*, August 20–24, Stockholm, Sweden, pp. 2292–2296.

Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**(1–2), 141–160.

Xuan, Y., and Zhang, Z. (**2014**). "Influence of embedded fibers and an epithelium layer on the glottal closure pattern in a physical vocal fold model," J. Speech Lang. Hear. Res. **57**(2), 416–425.

Zhang, Z. (**2016**). "Mechanics of human voice production and control," J. Acoust. Soc. Am. **140**(4), 2614–2635.