Speech Communication 132 (2021) 96–105

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Effects of the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis

Peter Birkholz [*], Susanne Drechsel

Institute of Acoustics and Speech Communication, TU Dresden, Germany

ARTICLE INFO

Keywords:
Transvelar coupling
Piriform fossae
Vocal tract model

ABSTRACT

Acoustic models of the vocal tract for articulatory speech synthesis often neglect a range of acoustic effects that are known to exist in the human vocal tract. Here we extended a basic acoustic vocal tract model by three features: the piriform fossae, transvelar acoustic coupling of the oral and nasal cavities, and sound radiation from the skin of the neck. The main goal was to find out how these features affect the naturalness of the synthesized speech. To this end, ten German words were synthesized with different combinations of the additional features, and listeners compared the naturalness of these stimuli. Surprisingly, all three features reduced the perceived naturalness, although they should make the synthesis more realistic. A closer analysis revealed that all new features emphasized the low frequencies compared to the high frequencies of the synthetic speech, leading to slightly more muffled speech with the used glottal excitation. An additional perception experiment with synthetic stimuli with a slightly more tense voice revealed no perceptual preference for the synthesis with or without the piriform fossae. These results indicate that the examined features play a minor role for the naturalness of articulatory synthesis compared to the voice source characteristics.

1. Introduction

Compared to other speech synthesis methods, articulatory speech synthesis is considered as one of the most challenging approaches, because it requires realistic models of the vocal tract and the vocal folds, an aerodynamic and acoustic simulation, and an articulatory control model. Each of these components comes in a range of types. Vocal tract models include geometric (Mermelstein, 1973; Iskarous et al., 2003; Kröger et al., 2004; Birkholz, 2013; Stone et al., 2018), statistical (Maeda, 1990; Beautemps et al., 2001; Badin et al., 2002; Toutios et al., 2011; Story et al., 2018), and biomechanical (Payan and Perrier, 1997; Dang and Honda, 2004; Stavness et al., 2011) models. Vocal fold models can be roughly divided into geometric (Titze, 1989; Cranen and Schroeter, 1996; Birkholz et al., 2019) and self-oscillating biomechanical (Birkholz, 2011; Erath et al., 2013) models. Aerodynamic and acoustic simulations are mostly performed in the time domain (Maeda, 1982; Birkholz and Jackèl, 2004; Birkholz et al., 2007; van den Doel and Ascher, 2008; Elie and Laprie, 2016; Birkholz and Pape, 2019), but hybrid time–frequency domain simulations are also possible (Sondhi and Schroeter, 1987; Teixeira et al., 2005). Control models mostly use a kind of gestural specification of an utterance that is then translated into trajectories of the articulatory parameters (Saltzman and Munhall, 1989; Kröger, 1993; Xu et al., 2006; Birkholz, 2007; Birkholz et al., 2011b; Alexander et al., 2019), but there are also alternative approaches (Bouabana and Maeda, 1998; Deng and dynamic, 1998; Okadome and Honda, 2001; Story and Bunton, 2019).

With appropriate models for all the components in place, articulatory synthesis should be able to generate highly natural speech with the highest level of flexibility with regard to the generated voices and speaking styles (Shadle and Damper, 2001). However, despite considerable progress in the last years (see https://www.vocaltractlab.de/index.php?page=vocaltractlab-examples for some recent examples), the speech quality of articulatory synthesizers cannot yet fully compete with state-of-the-art unit-selection synthesizers (e.g. Liu et al., 2017) or neural synthesizers (e.g. Shen et al., 2018). The key to the success of articulatory synthesis is to identify the critical details of the models and simulations that make the synthesized speech sound more natural.

This study explored specific features of an articulatory synthesizer at the acoustic level. Most synthesizers assume (one-dimensional) plane acoustic waves in an oropharyngeal tube with varying circular cross-sectional areas that has a separate branch for the nasal cavity (Birkholz, 2005; Elie and Laprie, 2016; Flanagan et al., 1975; Kröger, 1998; Maeda, 1982; Murphy et al., 2015; Sondhi and Schroeter, 1987; Teixeira et al., 2005; van den Doel and Ascher, 2008). There are also different methods for the *three-dimensional* acoustic simulation of the vocal tract (Blandin et al., 2015; Fleischer et al., 2015; Freixes et al.,

[*] Corresponding author.
E-mail address: peter.birkholz@tu-dresden.de (P. Birkholz).

https://doi.org/10.1016/j.specom.2021.06.002
Received 11 February 2021; Received in revised form 22 April 2021; Accepted 4 June 2021
Available online 10 June 2021

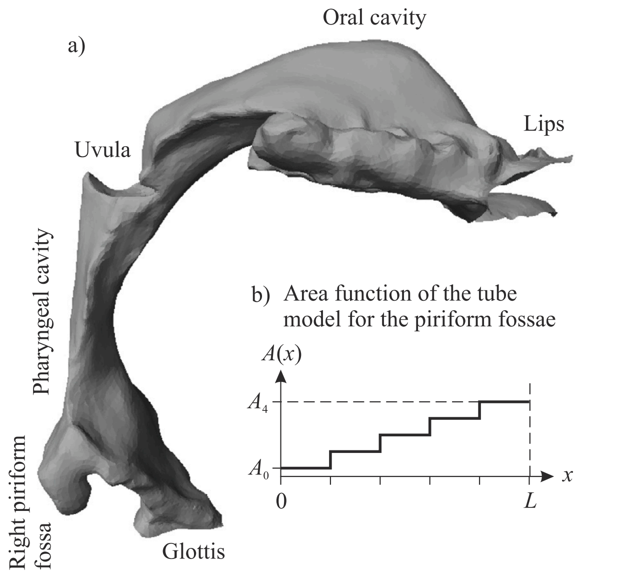

Fig. 1. (a) Rendering of the inner surface of the vocal tract for the vowel /a/ extracted from MRI data (from speaker s1 of the Dresden Vocal Tract Dataset Birkholz et al., 2020) showing the position of the piriform fossae near the glottal end. (b) The modelled area function of the (combined) piriform fossae that consists of five tube sections.

2019; Pont et al., 2020; Takemoto et al., 2010; Vampola et al., 2020) that are also accurate beyond the 5 kHz limit of plane-wave approaches. However, with current computer hardware they are far too computationally expensive for practical applications. With the overall goal of articulatory text-to-speech synthesis in mind, we therefore used a one-dimensional model in the present study. While the frequently used plane-wave models reflect the basic acoustic properties of the vocal tract quite well, they often neglect some potentially important features of real vocal tract acoustics. Here we consider three of these features.

The first feature is the acoustic consideration of the piriform fossae, which are two cavities lateral to the laryngeal vestibule near the glottis (Fig. 1 at the bottom left). They act as side branches to the main vocal tract which create a pronounced spectral trough in the speech spectrum between 4–5 kHz and shift the formants to lower frequencies (Dang and Honda, 1997; Delvaux and Howard, 2014; Fujita and Honda, 2005; Kitamura et al., 2005; Takemoto et al., 2013; Vampola et al., 2015; Zhang et al., 2016, 2019). Here we were mainly interested in the effect of the piriform fossae on the high frequencies, because they significantly affect the spectral envelope between 4–5 kHz and above, which may be perceptually relevant according to Monson et al. (2014).

The second feature is the acoustic coupling of the oral and nasal cavities through the soft tissue of the velum (Dang et al., 2016; Suzuki et al., 1990). This coupling mechanism is considered to generate sound in the nasal cavity even for non-nasalized sounds with a tightly closed velo-pharyngeal port. Dang et al. (2016) argue that this is actually one of the most important sources for the voice bars of voiced stops.

The third feature is the sound radiated by the vocal tract walls. From all positions on the surface of the neck and the face, the surface vibration is strongest at a level just above the larynx (Fant et al., 1976). This laryngeal wall vibration originates from the mechanical vibrations of the vocal folds or from the air pressure fluctuations directly above the vocal folds that are transmitted to the neck surface through the thin laryngeal tissue. This sound is usually considered as the main source for the voice bars of voiced stops and has therefore a similar effect as transvelar coupling.

As described above, all three features have different effects on the synthesized speech: while the piriform fossae cause a pronounced spectral dip in the 4–5 kHz region, transvelar coupling and laryngeal wall vibration increase the sound intensity at low frequencies and cause the voice bar during the closure phases of voiced plosives. So far, the effects of these features have been studied only at the acoustic level and for a limited range of speech sounds (e.g. Delvaux and Howard, 2014; Dang and Honda, 1996a).

The goal of the present study was twofold: (1) to propose mechanisms to include the three features into an acoustic transmission-line model of the vocal tract, and (2) to explore the effect of these features on the perceived naturalness of *connected* synthesized speech. With regard to the latter, we were interested in whether the increased realism introduced by these features would improve the naturalness of the synthesized speech. With regard to the first goal, the three features have been integrated into the acoustic model of the articulatory speech synthesizer VocalTractLab (www.vocaltractlab.de) as an extension to version 2.2, which is described in Section 2. The synthesizer was then used to synthesize ten German words with different combinations of included features, which were evaluated in perceptual experiments presented in Sections 3 and 4. Conclusions are drawn in Section 5.

2. The articulatory speech synthesizer

2.1. Overview

VocalTractLab is a complete and open source articulatory speech synthesizer implemented in C++. Its core components are a 3D geometric model of the vocal tract (Birkholz, 2013), multiple (exchangeable) models of the vocal folds (Birkholz et al., 2011a, 2019; Ishizaka and Flanagan, 1972), an aeroacoustic simulation in the time domain (Birkholz and Jackèl, 2004; Birkholz et al., 2007; Birkholz, 2014; Birkholz and Pape, 2019), and a gestural control model (Birkholz, 2007; Birkholz et al., 2011b). From an aeroacoustic point of view, the vocal system is represented as a time-varying branched tube, where the individual branches are formed by concatenations of short cylindrical tube sections with variable lengths and cross-sectional areas. The main branch comprises the subglottal system, the glottis, the pharyngeal cavity, and the oral cavity with 23, 2, 16, and 24 tube sections, respectively. Between the pharyngeal and oral cavities, the tube for the nasal cavity branches off. It consists of 19 tube sections and has additional branchings to four Helmholtz resonators that represent the paranasal sinuses (Dang and Honda, 1996b). The cross-sectional areas of the subglottal and nasal tubes are assumed to be time-invariant and based on anatomical data (except for the most posterior tube sections of the nasal branch, which vary with the velum height). On the other hand, the cross-sectional areas of the pharyngeal and oral tube sections are derived from the time-varying shape of the vocal tract model, and the two glottal tube sections take the glottal areas at the lower and upper edges of the vocal folds from the vocal fold model. In this study, the triangular glottis model (Birkholz et al., 2011a) was used for the simulations. The control parameter trajectories of the vocal tract and vocal fold models are generated by the gestural score, which is based on the ideas of articulatory phonology (Browman and Goldstein, 1992).

For the aeroacoustic simulation, the branched tube model is transferred into an acoustic transmission-line model with lumped elements as shown in Fig. 2. In this circuit, each tube section is represented as a T-type two-port network as sketched in the top-left box in Fig. 2. The network components are calculated based on the geometric dimensions of the tube sections (Birkholz and Jackèl, 2004). For the consideration of aerodynamic losses due to turbulence at tube expansions (e.g., at the exit of the glottis), non-linear resistors are added (Birkholz et al., 2007; Birkholz and Pape, 2019). The main energy source for the simulation is a pulmonary pressure source connected to the “lower” end of the subglottal tube. Additional noise sources are automatically inserted into the circuit as random pressure sources based on the aerodynamic and geometric conditions in the individual tube sections (Birkholz, 2014). At the mouth and the nostrils, the circuit is terminated with radiation impedances. The time-varying pressures and volume velocities in the whole acoustic network are numerically simulated in the time domain

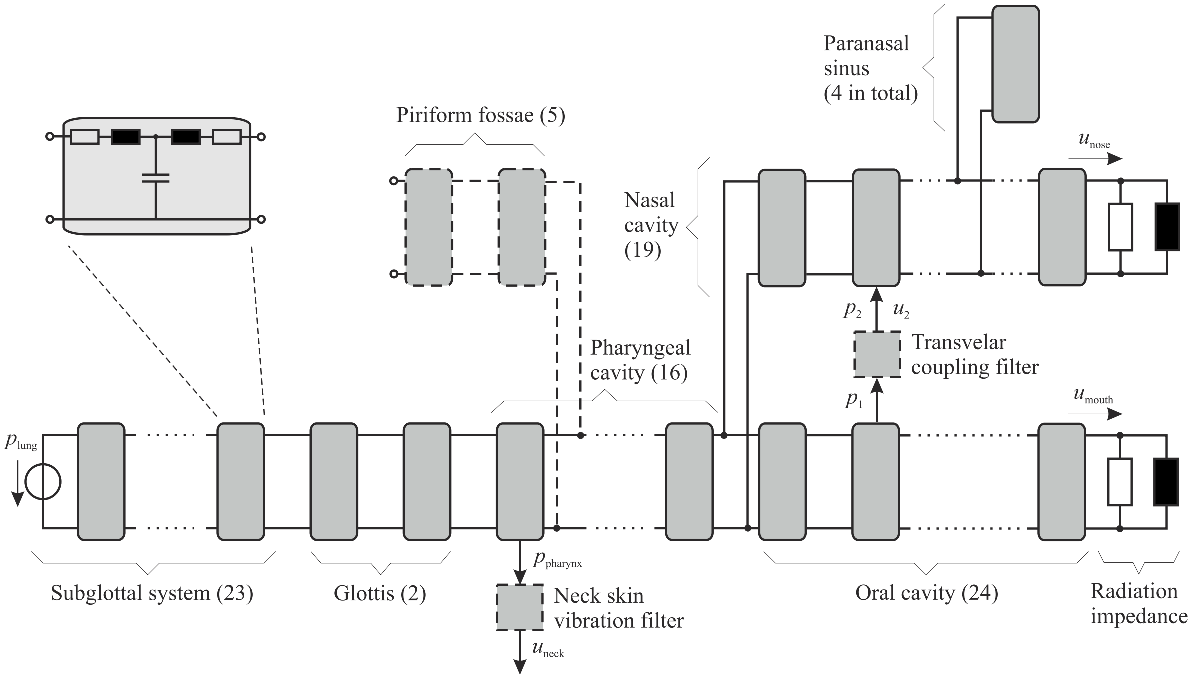

Fig. 2. Schema of the acoustic transmission line model of the vocal tract. The elements for the piriform fossae, laryngeal wall vibration, and transvelar acoustic coupling analysed in this study are indicated by the grey boxes with dashed border lines. The numbers in brackets behind the labels indicate the numbers of tube sections that represent the different vocal tract parts.

with a sampling rate of 44 100 Hz (Birkholz and Jackèl, 2004; Birkholz, 2005). The radiated sound pressure p_{rad} is calculated as the time-derivative of the volume velocities through the mouth and nostrils, i.e., $p_{\text{rad}} = d(u_{\text{mouth}} + u_{\text{nose}})/dt$, low-pass filtered at 8 kHz with an 8-pole digital Chebyshev filter, and resampled to 22 050 Hz.

The network components drawn with solid borders in Fig. 2 represent the currently predominant structure for acoustic models of the vocal tract (e.g. Elie and Laprie, 2016; Maeda, 1982; Sondhi and Schroeter, 1987; Teixeira et al., 2005). The additional acoustic components that were implemented and analysed in the present study are drawn as dashed boxes and are detailed in the following subsections.

2.2. Modelling the piriform fossae

As mentioned above, the piriform fossae are two small pear-shaped side cavities of the vocal tract located left and right from the epilaryngeal tube. So far there are only a few studies, each including only a few subjects, that provide quantitative data on the piriform fossae during speech production. All studies agree that the piriform fossae cause a pronounced spectral dip in the 4–5 kHz region of the speech spectrum, that they slightly shift the formant frequencies, that their cross-sectional areas increase approximately linearly from the closed end to the open end, and that their dimensions can substantially differ across subjects. On the other hand, there is no consensus on the degree of (a)symmetry between the left and right cavities, and the dependence on the vocal tract shape. While Dang and Honda (1997) found that the left and right piriform fossae were acoustically symmetric for their four subject, other studies reported left–right asymmetries (Takemoto et al., 2013; Zhang et al., 2016, 2019). Furthermore, some studies found that the shape of the piriform fossae is relatively stable regardless of the produced vowel (Dang and Honda, 1997; Kitamura et al., 2005), while others reported a significant dependence of the cavity volumes on the produced vowel (Delvaux and Howard, 2014; Zhang et al., 2019).

In the present study, we assumed that the left and right piriform fossae are symmetric and that their shapes are independent from the vocal tract state. Assuming symmetric cavities has the advantage that they are acoustically similar to a *single* tube with the combined volume of both cavities. Here, as shown in Fig. 1b, this tube was modelled in terms of $N = 5$ tube sections with the cross-sectional areas

$$A_i = \frac{2V_{\text{tot}}}{L}(i + 1/2)/N, \qquad i = 0 \ldots N-1, \tag{1}$$

where V_{tot} is the total volume of both piriform fossae and L is the total (acoustic) length. The length of each section was L/N, so that the total volume sums up to V_{tot}. In male subjects, the total volume of the piriform fossae ranges from about 1.0 cm^3 to 3.5 cm^3, and the length ranges from 14 mm to 21 mm (Delvaux and Howard, 2014; Zhang et al., 2019). For the incorporation of this branch in the transmission

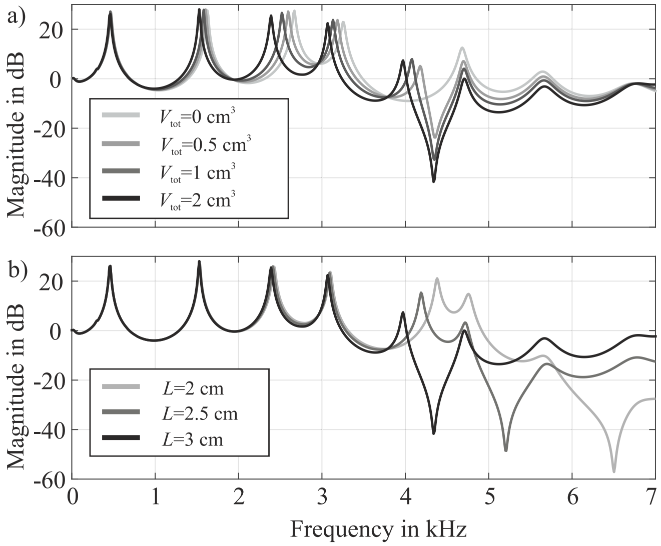

Fig. 3. Effects of different lengths L and (total) volumes V_{tot} of the piriform fossae on simulated volume velocity transfer functions between the (closed) glottis and the lips for the vowel schwa. (a) Transfer functions for different volumes at a constant length of $L = 3$ cm. (b) Transfer functions for different lengths at a constant volume of $V_{\text{tot}} = 3$ cm^3.

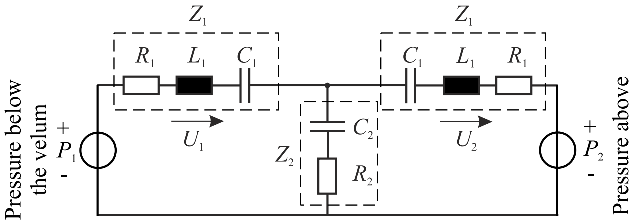

Fig. 4. Equivalent acoustic network for the velum.

line model of the vocal tract, an open-end correction must be applied that makes it acoustically appear about 50% longer (Dang and Honda, 1997). For the present study, we chose the values $V_{\mathrm{tot}} = 2$ cm^3 and $L = 3$ cm.

Fig. 3a and b illustrate the effects of different volumes and lengths of the piriform fossa tube, respectively, in terms of simulated vocal tract transfer functions for the vowel schwa using the acoustic model presented above. They show that the volume mainly determines the depth of the spectral trough (i.e., the bandwidth of the antiresonance), and the degree of shift of the resonances. The greater the volume, the deeper the spectral trough and the stronger the resonances are shifted towards lower frequencies. At the perceptual level, this makes the voice sound more resonant for larger volumes of the piriform fossae (Delvaux and Howard, 2014). The length of the piriform fossa tube mainly affects the *frequency* f_{dip} of the antiresonance with $f_{\mathrm{dip}} \propto 1/L$ (Dang and Honda, 1997). With the length and volume chosen in the present study, the frequency and depth of the antiresonance are in good agreement with measurements (Birkholz et al., 2020).

2.3. Modelling transvelar acoustic coupling

Transvelar acoustic coupling means that intraoral sound pressure variations stimulate mechanical vibrations of the velum, which in turn generate sound in the nasal cavity. By this mechanism, sound can be radiated from the nostrils even when the velopharyngeal port is fully closed. The implementation of this mechanism here is based on the mechanical two-layer diaphragm model of the velum proposed by Dang et al. (2016). The equivalent acoustic circuit of this model is shown in Fig. 4 and differs from the structure of the acoustic circuits that represent the tube sections of the vocal tract model shown in Fig. 2. To preserve the structure of the acoustic model, the velum model was therefore implemented as a digital filter.

The input to this filter were the sound pressures p_1 and p_2 below and above the middle part of the velum, respectively. These values were taken from the centres of the third oral tube section and the third nasal tube section (as counted from the nasal branching point). The output of the filter was the volume velocity u_2 generated by the movement of the superior velum surface, which was added to the acoustic volume velocity that enters the third nasal tube section.

The parameters of the digital filter were obtained from the acoustic model of the velum (Fig. 4) using the matched z-transform as follows. The network elements of the acoustic model are (Dang et al., 2016)

$$R_1 = 8.96 \text{ g cm}^{-4} \text{ s}^{-1}$$
$$L_1 = 0.0343 \text{ g cm}^{-4}$$
$$C_1 = 4.15 \times 10^{-4} \text{ cm}^4 \text{ s}^2 \text{ g}^{-1}$$
$$R_2 = 0.9 \text{ g cm}^{-4} \text{ s}^{-1}$$
$$C_2 = 4.78 \times 10^{-5} \text{ cm}^4 \text{ s}^2 \text{ g}^{-1}$$

and can be combined into the impedances $Z_1 = R_1 + sL_1 + 1/(sC_1)$ and $Z_2 = R_2 + 1/(sC_2)$, where s is the complex frequency variable. Then the current U_2,[1] which enters the nasal cavity via the velum, can be expressed as a function of the two pressures P_1 and P_2 acting on the velum surfaces from below and above, respectively:

$$U_2 = P_1 \frac{Z_2}{2Z_1Z_2 + Z_1^2} + P_2 \frac{-Z_1 - Z_2}{2Z_1Z_2 + Z_1^2}.$$

Based on this equation, we defined the two transfer functions

$$H_1 = \left.\frac{U_2}{P_1}\right|_{P_2=0} = \frac{Z_2}{2Z_1Z_2 + Z_1^2} \quad \text{and}$$
$$H_2 = \left.\frac{U_2}{P_2}\right|_{P_1=0} = \frac{-Z_1 - Z_2}{2Z_1Z_2 + Z_1^2}$$

so that $U_2 = P_1H_1 + P_2H_2$. Expanding Z_1 and Z_2 in these transfer functions and sorting terms by powers of s we obtain

$$H_1(s) = k_1 \frac{sm_1 + s^2}{d_0 + sd_1 + s^2d_2 + s^3d_3 + s^4} = k_1 \frac{M(s)}{D(s)}$$
$$H_2(s) = k_2 \frac{sn_1 + s^2n_2 + s^3}{d_0 + sd_1 + s^2d_2 + s^3d_3 + s^4} = k_2 \frac{N(s)}{D(s)}$$

with the following constants:

$$k_1 = R_2/L_1^2 = 7.65 \text{ cm}^4 \text{ s g}^{-1}$$
$$k_2 = -1/L_1 = -29.15 \text{ cm}^4 \text{ s g}^{-1}$$
$$m_1 = 1/(C_2R_2)$$
$$n_1 = 1/(C_1L_1) + 1/(C_2L_1)$$
$$n_2 = R_1/L_1 + R_2/L_1$$
$$d_0 = 2/(C_1C_2L_1^2) + 1/(C_1^2L_1^2)$$
$$d_1 = 2R_1/(C_2L_1^2) + 2R_2/(C_1L_1^2) + 2R_1/(C_1L_1^2)$$
$$d_2 = 2R_1R_2/L_1^2 + 2/(C_2L_1) + R_1^2/L_1^2 + 2/(C_1L_1)$$
$$d_3 = 2R_2/L_1 + 2R_1/L_1.$$

The roots of the polynomials $M(s)$, $N(s)$ and $D(s)$ were numerically evaluated, with $M(s) = 0$ for $s_{M1} = 0$ and $s_{M2} = -23\,245.0$, with $N(s) = 0$ for $s_{N1} = 0$ and $s_{N2,N3} = -143.7 \pm 812.1j$, and with $D(s) = 0$ for $s_{D1,D2} = -156.9 \pm 1124.9j$ and $s_{D3,D4} = -130.6 \pm 230.6j$. These roots were mapped into the z-domain by $z = e^{s\Delta t}$ according to the matched z-transform method, where $\Delta t = 1/44\,100$ s is the time step of the digital simulation, to obtain the transfer functions

$$H_1(z) = k_1' \frac{z^2 \prod_{i=1}^{2}(z - z_{Mi})}{\prod_{i=1}^{4}(z - z_{Di})}$$

and

$$H_2(z) = k_2' \frac{z \prod_{i=1}^{3}(z - z_{Ni})}{\prod_{i=1}^{4}(z - z_{Di})}.$$

The constants $k_1' = 5.03 \times 10^{-7}$ and $k_2' = 6.59 \times 10^{-4}$ were determined by matching the magnitude response of the analog and digital filters at a frequency of 10 Hz. Note that zeros at $z = 0$ have been added to the digital filters (two for H_1 and one for H_2) to obtain minimum delay filters. Mapping $H_1(z)$ and $H_2(z)$ in the time domain with the time index n finally yields the recursion equation for the volume velocity $u_2[n]$ of the form

$$u_2[n] = \sum_{i=0}^{2} a_i p_1[n-i] + \sum_{i=0}^{3} a_i' p_2[n-i] + \sum_{i=1}^{4} b_i u_2[n-i]$$

with the recursion coefficients a_i, a_i' and b_i.

Fig. 5 shows the magnitude responses of the filters $H_1(s)$ and $H_2(s)$, which are low-pass filters with a cut-off frequency of about 200 Hz. Hence, the transmission of sound through the closed velum

[1] Capital-letter U and P denote the volume velocity and pressure in the frequency domain, while lower-case letters are used for these quantities in the time domain.

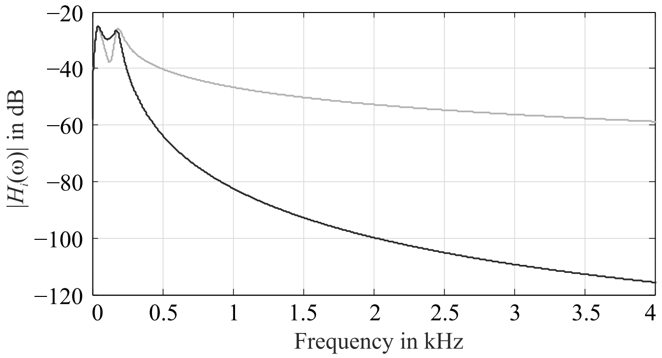

Fig. 5. Magnitude of the transfer functions $H_1(s)$ (black) and $H_2(s)$ (grey).

is essentially restricted to frequencies up to 200 Hz. According to our simulations, typical sound pressures during phonation below the closed velum cause volume velocity waves with a peak amplitude of about 1 cm^3/s above the velum in the nasal cavity. For comparison, the peak amplitude of the glottal volume velocity during phonation is about 400 cm^3/s (Stevens, 1998).

2.4. *Laryngeal wall vibration*

Sound from inside the vocal tract is also transmitted through the skin at other places of the vocal tract than the velum. Since the anterior laryngeal wall is quite thin and the larynx contains the vibrating vocal folds, laryngeal wall vibration contributes most to the sound radiated from the skin (Fant et al., 1976). This sound radiation has rarely been modelled in articulatory synthesizers. To the best of our knowledge, there is also no published data on transfer functions from the sound pressure in the laryngeal region of the vocal tract to the velocity of the outer skin movement. There are a few studies on *similar* transfer functions of the neck tissue with different input and output variables (Meltzner et al., 2003; Švec et al., 2005; Wu et al., 2014), however, they are not straightforward to convert to the presented situation. Since these transfer functions generally show a low-pass effect like the acoustic model for the velum, the corresponding digital filter was used here also for laryngeal wall vibration. Here, the variables p_1, p_2, and u_2 represent the intra-laryngeal sound pressure, the sound pressure in front of the outer laryngeal wall, and the volume velocity of the outer laryngeal wall surface, respectively. The pressure p_2 is very small compared to p_1 and was set to zero here. When laryngeal wall vibration was included in the acoustic simulation, the total radiated sound pressure was calculated as

$$p_{\mathrm{rad}} = d(u_{\mathrm{skin}} + u_{\mathrm{mouth}} + u_{\mathrm{nose}})/dt,$$

where $u_{\mathrm{skin}} = u_2$ is the output of the filter.

3. Perception experiment 1

In the previous section we discussed three new features that could be individually included in or excluded from the acoustic simulation. In experiment 1 we examined

- whether listeners were able to perceive any difference at all between synthesized words with all three features excluded (basic acoustic model) and with the individual features included separately (task 1), and
- how the naturalness of the synthesized speech was affected when the features were included in different combinations (task 2).

Table 1
The 10 words synthesized for the perception experiments.

Word	Transcription	English
Banane	[baˈnaːnə]	Banana
Birne	[ˈbɪʁnə]	Pear
Gurken	[ˈɡʊʁkən]	Cucumbers
Himbeere	[ˈhɪmˌbeːʁə]	Raspberry
Mandarine	[ˌmandaˈʁiːnə]	Mandarin
Melone	[meˈloːnə]	Melon
Mirabelle	[miʁaˈbɛlə]	Mirabelle
Orange	[oˈʁaŋʒə]	Orange
Physalis	[fyːˈzalɪs]	Physalis
Rosine	[ʁoˈziːnə]	Raisin

3.1. *Stimuli*

Each of the 10 German words in Table 1 was synthesized in 8 variants, i.e. for all 8 combinations of the 3 binary features {piriform fossae included vs. excluded} × {transvelar acoustic coupling included vs. excluded} × {laryngeal wall vibration included vs. excluded} (80 stimuli in total).[2] To this end, a gestural score was manually created for each word using the graphical editor in VocalTractLab. The gestural score for a word was independent from the inclusion or exclusion of the three features, because the features do not affect the articulatory movements. To ensure a natural-sounding prosody of the words, the phone durations and f_0 contours in the scores were adjusted to match those of corresponding recorded words from a male German speaker. The subglottal pressure was set to 1 kPa and the time constants of the gestures were adjusted according to the guidelines given in Birkholz et al. (2017). For all stimuli, the vocal fold rest displacement for voiced sounds was set to 0.15 mm. This value was found to be perceptually suitable based on previous synthesis experiments with the basic acoustic model. All stimuli were saved as 16 bit mono WAV files with a sampling rate of 22 050 Hz.

One aspect of the synthesis deserves special attention. The vocal tract target shapes for the speech sounds used for the synthesis were originally optimized for the *basic* acoustic simulation, i.e., for the case that the piriform fossae are not included (Birkholz, 2013). However, as shown in Section 2.2, the inclusion of the piriform fossae can shift the formants to lower frequencies. This can have a negative impact on the quality of the vowels and hence the perception. In this study we were interested in the perceptual effects of the spectral changes that the piriform fossae cause at the higher frequencies, i.e., around the frequency of the spectral trough and above. Therefore, when the piriform fossae were included in the acoustic simulation, a *modified* set of vocal tract shapes was used as targets for the vowels. The modified shapes were obtained by means of parameter optimization (Birkholz, 2013) in such a way that they have the same first three resonance frequencies with the effect of the piriform fossae as the original shapes without the effect of the piriform fossae. Fig. 6a illustrates the difference between the original and modified vocal tract shapes for the vowel /a/. Fig. 6b shows the calculated vocal tract transfer functions of /a/ for the original shape without the piriform fossae (black line) and with the piriform fossae (dashed grey line), and for the modified shape with the piriform fossae (solid grey line).

3.2. *Participants*

Twenty-two native speakers of German (15–64 years old; mean: 33 years; 15 males and 7 females) participated in the experiment after providing informed consent. None of them reported speech or hearing problems. The experiment was conducted according to the ethical principles based on the WMA Declaration of Helsinki.

[2] All stimuli (also for experiment 2) are available as supplemental material at https://www.vocaltractlab.de/index.php?page=birkholz-supplements.

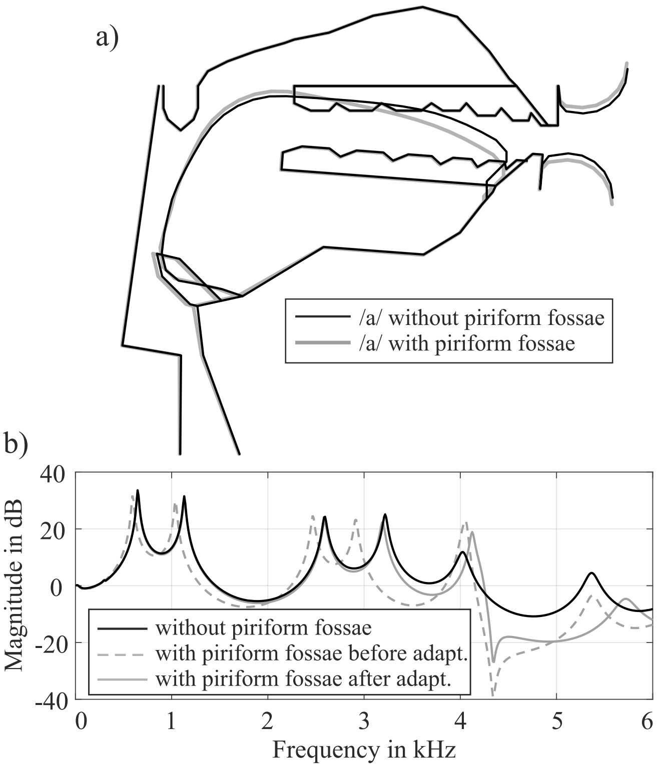

Fig. 6. (a) Vocal tract model shapes for /a/ in the midsagittal plane that generate the same first three acoustic resonances with (grey line) and without (black line) the piriform fossae. (b) Volume velocity transfer functions from the glottis to the lips of the vocal tract in (a) without the piriform fossae (black line), with the piriform fossae but without an adaptation of the articulation (grey dashed line), and with the piriform fossae after an adaptation of the articulation (grey solid line).

3.3. Task 1

The first task was a discrimination task to find out whether two stimuli for the same word that differed in just one feature could be perceptually discriminated. For each of the 10 words, the following three pairs of feature combinations were contrasted:

- (*no* piriform fossae, *no* wall vibration, *no* transvelar coupling) vs. (**piriform fossae**, *no* wall vibration, *no* transvelar coupling)
- (*no* piriform fossae, *no* wall vibration, *no* transvelar coupling) vs. (*no* piriform fossae, **wall vibration**, *no* transvelar coupling)
- (*no* piriform fossae, *no* wall vibration, *no* transvelar coupling) vs. (*no* piriform fossae, *no* wall vibration, **transvelar coupling**)

Hence, the basis variant of each word (no piriform fossae, no wall vibration, no transvelar coupling) was compared to the three variants with exactly one feature enabled for a total of $10 \times 3 = 30$ pairs. Each pair was used in four types of AXB sequences (AAB, BBA, BAA, and ABB) for a total of 120 sequences.

The 120 sequences were presented to the participants in randomized order using a laptop computer with an external sound card (Terratec Aureon XFire 8.0 HD) and high-quality closed headphones (AKG K-240) in a quiet room. The three stimuli of each sequence were played with short pauses of 100 ms between them. The task of the participants was to decide whether X was identical to A or B in each AXB sequence by pressing one of two buttons on the computer screen using a mouse. Each sequence could be repeated once by pressing another button on the screen. After a decision, the stimuli of the next sequence were automatically played.

For the analysis, we performed three binomial tests of the null hypothesis that the subjects were not able to detect the difference due to the respective acoustic features, i.e., that their chance of selecting the correct response was 50%.

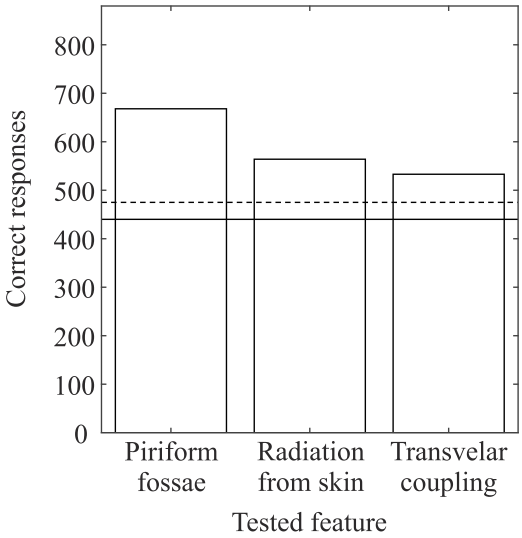

Fig. 7. Results of the discrimination task of the first experiment. The height of the bars shows how often the participants correctly contrasted two stimuli that just differed in one feature. The solid horizontal line indicates the chance level. The dashed line is the upper boundary of the acceptance area of the null hypothesis that the subjects selected their responses by chance. Hence, the acoustic effect of including or excluding any of the three features was perceivable.

3.4. Task 2

The second task was an assessment of the naturalness of the stimuli by forced-choice pairwise comparisons of stimuli. For each of the 10 words, all possible stimuli pairs with different feature combinations were formed, i.e. $7 + 6 + \cdots + 1 = 28$ pairs per word. Hence, for the 10 words, we obtained 280 stimuli pairs.

The 22 participants were divided into two groups of 11 people, and each group performed the test with one of the two possible orders in which the stimuli of each pair can be presented. For each participant, the 280 stimuli pairs were presented one after another in an individually randomized order using the same equipment and room as in task 1. Each pair was presented with a short pause of 100 ms between the two stimuli. After the presentation of a pair, the participant had to decide, which of the two stimuli sounded more natural by pressing one of two buttons on a computer screen using a mouse. Each pair could be repeated once by pressing another button on the screen. After a decision, the stimuli of the next pair were automatically played.

For the analysis, the responses of the participants were collected in a table with 22 rows, one for each participant, and 8 columns, one for each feature combination. Each table cell contained the number how often the respective participant preferred the respective feature combination across all 10 words. Because each of the 8 feature combinations was contrasted with all 7 others, the maximal possible number in a cell was $7 \times 10 = 70$. For each pair of two feature combinations (i.e., two columns of the table), the rankings were compared with the Wilcoxon signed-rank test.

3.5. Results and discussion

With regard to task 1, the binomial tests confirmed with $p < 0.001$ that the rate of correct responses was significantly higher than the chance rate of 50%. The bars in Fig. 7 show the number of correct responses for the three different features, all of which are above the upper boundary of the acceptance area of the null hypothesis (dashed

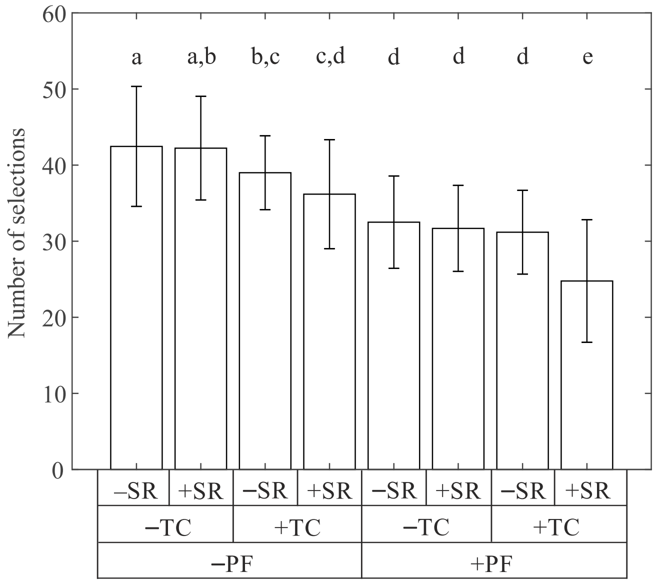

Fig. 8. Results of the assessment of naturalness of experiment 1. The height of the bars indicates how often the participants preferred the stimuli with a certain combination of features over the stimuli with other feature combinations. The vertical lines indicate the range of ±1 standard deviation. The three features are the radiation of sound from the neck skin (SR), transvelar coupling (TC), and the piriform fossa (PF). A preceding plus sign means that a certain feature was included in the simulation, and a minus sign means that it was excluded. Bars that do not share any letter indicate feature combinations that were rated significantly different based on a Wilcoxon signed-rank test at a 5% level of significance.

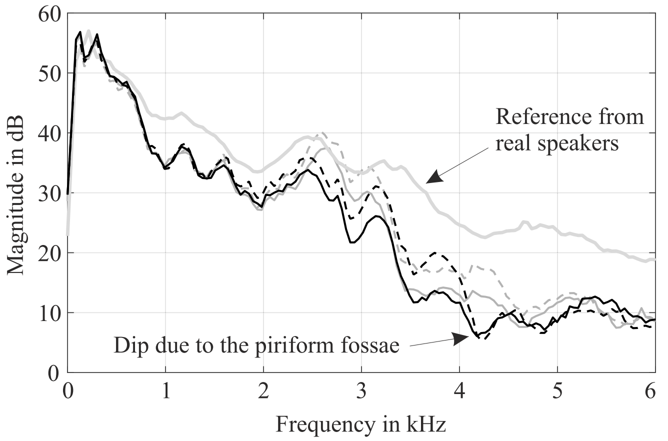

Fig. 9. Long-term average spectra (LTAS) of the ten words for different feature combinations in the synthesis: with the piriform fossae (black curves), without the piriform fossae (thin grey curves), with weak glottal adduction (solid lines), and with stronger glottal adduction (dashed lines). The thick grey curves represents the LTAS of the ten words spoken by six real speakers.

line). This suggests that the listeners were able to reliably discriminate stimuli with and without the tested acoustic features.

With regard to task 2, Fig. 8 shows how often each combination of acoustic features was preferred over the others combinations. The leftmost bar shows the results for the case that all three features are *excluded* (corresponding to the basic acoustic model), and the rightmost bar corresponds to the case that all three features are *included*. Significant differences between the feature combinations are indicated by the letters above the bars: The rankings for feature combinations with no letter in common are significantly different ($\alpha = 0.05$) (Piepho, 2018). A correction of the significance level due to the multiple pairwise comparisons was not conducted due to the exploratory nature of the analysis.

Surprisingly, the stimuli synthesized with the new features tended to sound less natural than the stimuli synthesized without any of the features, although the additional features should make the synthesis more realistic. When the effect of the features is examined individually, i.e., when all stimuli *with* a particular feature (group 1) are compared to all stimuli *without* this feature (group 2), then the difference of the responses is highly significant for the features “piriform fossae” and “transvelar coupling” ($p < 0.001$ with a Wilcoxon signed rank test). For the feature “laryngeal wall vibration”, the two groups are not significantly different ($p = 0.1077$).

The reason for the reduced naturalness with increased realism of the simulation was not immediately clear. However, some participants of the experiment reported that they perceived some of the stimuli as more muffled than others. This motivated us to have a closer look at the long-term-average spectra (LTAS) of the stimuli. The LTAS were calculated by concatenating the audio signals of the words synthesized with a certain combination of features (omitting pauses between the words), and averaging the short-term magnitude spectra with a window length of 23.2 ms and a time step of 10 ms. Because the piriform fossae had the strongest impact on the LTAS, we focus on this feature here. Fig. 9 shows the difference between the LTAS of the stimuli synthesized with the piriform fossae (solid black curve) and without (thin solid grey curve). Obviously, the spectral magnitudes of the stimuli synthesized *with* the piriform fossae were up to 10 dB smaller in the range between 2.5 kHz and 5 kHz. Hence, synthetic stimuli including the piriform fossae have less high-frequency energy and therefore sound somewhat more muffled.

As a reference, we also calculated the LTAS of the ten words spoken by six real speakers, which is shown as the thick grey line in Fig. 9. The spectral slope of this LTAS is less steep than for any of the synthetic stimuli. This indicates that the spectral envelope of the synthetic voice source generally rolled off too fast with increasing frequency.

In summary, when the high-frequency components in synthetic utterances are generally weak, listeners seem to prefer stimuli with higher intensities at higher frequencies (the basic acoustic model in our case) in pairwise comparisons.

4. Perception experiment 2

The second experiment was designed to find out whether the piriform fossae would still reduce the naturalness when a more tense synthetic voice with stronger high-frequency components was used. The more tense voice was generated with more adducted vocal folds in the used self-oscillating vocal fold model (Birkholz et al., 2011a). Because transvelar acoustic coupling and laryngeal wall vibration had a much smaller spectral impact than the piriform fossae, they were not considered in the second experiment.

4.1. Stimuli

Each of the 10 words in Table 1 was synthesized in 4 variants, i.e. all combinations of the binary feature {piriform fossae included vs. excluded} and two settings for the vocal fold rest displacement ξ_g. The settings for the rest displacement were $\xi_g = 0.15$ mm (same as in experiment 1) and $\xi_g = 0.05$ mm (for more adducted vocal folds and hence a more tense voice). The synthesis procedure was the same as for experiment 1. Transvelar acoustic coupling and laryngeal wall vibration were disabled for all stimuli.

For the new stimuli with the more tense voice, two LTAS spectra were calculated: one for the stimuli where the piriform fossae were included, and one for the stimuli were they were excluded. These spectra are shown as the dashed lines in Fig. 9, where the black dashed line corresponds to the stimuli *with* the piriform fossae. Both spectra show clearly higher intensities between 2–5 kHz than the LTAS of the

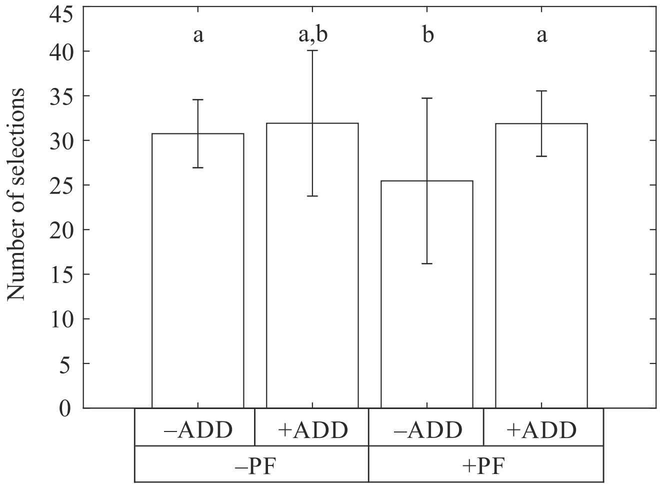

Fig. 10. Results of the assessment of naturalness of experiment 2. The height of the bars indicates how often the participants preferred the stimuli with a certain combination of features over the stimuli with other feature combinations. The two features are the increased adduction of the vocal folds (ADD) and the piriform fossa (PF). A preceding plus sign means that a certain feature was included in the simulation, and a minus sign means that it was excluded. Bars that do not share any letter indicate feature combinations that are significantly different based on a Wilcoxon signed-rank test at a 5% level of significance.

stimuli with more abducted vocal folds (the thin solid lines). Hence, the spectral characteristics of the stimuli synthesized with the more tense voice became more similar to the characteristics of the natural utterances (thick grey line), although significant differences remained.

4.2. Participants

Twenty-four native speakers of German (21–48 years old; mean: 33 years; 9 males and 15 females) participated in the experiment after providing informed consent. None of them reported speech or hearing problems. Six of them participated already in the first experiment. The experiment was conducted according to the ethical principles based on the WMA Declaration of Helsinki.

4.3. Procedure

Analogous to task 2 of the first experiment, the participants assessed the naturalness of the stimuli by forced-choice pairwise comparisons. For each of the 10 words, all possible stimuli pairs with different variants (feature combinations) were formed, i.e. $3 + 2 + 1 = 6$ pairs per word. Hence, for the 10 words, 60 stimuli pairs were obtained. Each pair was presented in both possible orders of the stimuli, so that there were 120 (ordered) stimuli pairs in total.

Except the number of pairs and the synthesizer settings, the rest of the experiment was conducted like task 2 of the first experiment. As before, the listener responses were analysed with Wilcoxon signed-rank tests.

4.4. Results and discussion

Fig. 10 shows how often the participants preferred each synthesis variant over the others. The preferences for three of the four variants were not significantly different ($p > 0.05$): the two variants without the piriform fossae, and the variant with the piriform fossae and stronger vocal fold adduction. The only synthesis variant that was clearly less preferred was the one with the piriform fossae and the wider rest displacement of the vocal folds (as used in experiment 1). Hence, when the vocal folds are sufficiently adducted so that the high-frequency components of the voice source are sufficiently intense, the presence or absence of the piriform fossae in the simulation does not matter for the perceived naturalness of the synthesis.

5. General discussion and conclusions

The goal of this study was to incorporate the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration into a transmission-line model of the vocal tract and to find out whether or not these effects would improve the perceived naturalness of synthesized speech. The first experiment showed that the acoustic changes caused by these features could be perceived, but they were *less* preferred by the listener (at least for the piriform fossae and transvelar acoustic coupling). The reason for this unexpected finding was that with the used voice source settings, the high-frequency components of the speech signals were rather low compared to real speech. In this case, listeners preferred settings of the acoustic vocal tract model that did not further reduce the high-frequency components compared to the low-frequency components. In the second experiment, it was shown that the acoustic effect of the piriform fossae did not affect the perceived naturalness anymore when the high-frequency components of the voice source were raised, making the synthetic voice more similar to natural voices.

These findings indicate that for the naturalness of articulatory speech synthesis, the examined acoustic features (at least in their current implementation) are far less relevant than the voice source characteristics. In other words, for highly natural synthesis, the voice source model plays the major role. Somewhat similar conclusions were drawn from another recent study by Freixes et al. (2019), where 3D acoustic simulations of realistic vocal tract geometries (including higher-order modes) were compared with simulations of simplified straight axis-symmetric vocal tract tubes, which prevent the onset of higher-order modes. It was found that the (more realistic) higher-order modes generally induced a reduction of the high-frequency energy that makes the higher frequencies acoustically less relevant. Also there, the glottal source excitation was found to play a more important role than some specific details that would seem significant a priori.

On the other hand, it is conceivable that the implemented models for the examined features need further improvement before they add to the naturalness of the synthesized speech. While the model for transvelar acoustic coupling is already strongly based on experimental data, the (identical) model for sound radiation from the skin of the neck is not based on dedicated experimental data yet. The model for the piriform fossae currently neglects a potential asymmetry between the left and right fossae (Takemoto et al., 2013; Zhang et al., 2016, 2019), which would cause a more complex spectral pole-zero pattern, and a potential change of its volume and length with the articulated phoneme (Delvaux and Howard, 2014; Zhang et al., 2019). The impact of such improvements remains to be explored in future studies.

In any case, even when the three examined features do not contribute to the naturalness of the synthesis in their current form, they do increase the realism of the simulation. Especially the acoustic effect of the piriform fossae in the real vocal tract has been convincingly shown in previous studies (e.g. Dang and Honda, 1997; Fujita and Honda, 2005; Takemoto et al., 2013) and cannot be reproduced in an articulatory synthesizer without an additional acoustic branch as in this study. Furthermore, the consideration of transvelar acoustic coupling and laryngeal wall vibration generated a voice bar in the synthetic stimuli (which would otherwise be absent), making them acoustically more realistic.

Finally, we have shown that the acoustic effects of the examined features are perceptible. Hence, especially the piriform fossae are likely to carry information about the individual speaker (Godoy et al., 2016). This is also suggested by the substantial spectral changes that result from rather small changes of their dimensions, as shown in Fig. 3. Therefore, the piriform fossae may be an important ingredient for articulatory speech synthesis with individual speaker characteristics.

CRediT authorship contribution statement

Peter Birkholz: Conceptualization, Methodology, Formal analysis, Supervision, Visualization, Writing. **Susanne Drechsel:** Methodology, Investigation, Writing - reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Simon Stone for valuable discussions about the interpretation of the results, Klaus Willmes for his help with the statistical analysis, and the anonymous reviewers for their valuable comments.

References

Alexander, R., Sorensen, T., Toutios, A., Narayanan, S., 2019. A modular architecture for articulatory synthesis from gestural specification. J. Acoust. Soc. Am. 146 (6), 4458–4471.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C., 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. J. Phonet. 30, 533–553.

Beautemps, D., Badin, P., Bailly, G., 2001. Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. J. Acoust. Soc. Am. 109 (5), 2165–2180.

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

Birkholz, P., 2007. Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In: Interspeech 2007 - Eurospeech, Antwerp, Belgium, pp. 2865–2868.

Birkholz, P., 2011. A survey of self-oscillating lumped-element models of the vocal folds. In: Kröger, B.J., Birkholz, P. (Eds.), Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011. TUDPress, Dresden, pp. 47–58.

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

Birkholz, P., 2014. Enhanced area functions for noise source modeling in the vocal tract. In: Proc. of the 10th International Seminar on Speech Production (ISSP 2014), Cologne, Germany, pp. 37–40.

Birkholz, P., Drechsel, S., Stone, S., 2019. Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis. In: Proc. of the Interspeech, pp. 3765–3769.

Birkholz, P., Jackèl, D., 2004. Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In: Interspeech 2004-ICSLP, Jeju, Korea, pp. 1125–1128.

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

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

Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C., 2011b. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. IEEE Trans. Audio, Speech Language Process. 19 (5), 1422–1433.

Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., Fleischer, M., 2020. Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. Scientific Data 7 (1), 1–16.

Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., Neuschaefer-Rube, C., 2017. Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. Comput. Speech Lang. 41, 116–127.

Birkholz, P., Pape, D., 2019. How modeling entrance loss and flow separation in a two-mass model affects the oscillation and synthesis quality. Speech Commun. 110, 108–116.

Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Hirtum, A.V., Laval, X., 2015. Effects of higher order propagation modes in vocal tract like geometries. J. Acoust. Soc. Am. 137 (2), 832–843.

Bouabana, S., Maeda, S., 1998. Multi-pulse LPC modeling of articulatory movements. Speech Commun. 24, 227–248.

Browman, C.P., Goldstein, L., 1992. Articulatory phonology: An overview. Phonetica 49, 155–180.

Cranen, B., Schroeter, J., 1996. Physiologically motivated modelling of the voice source in articulatory analysis/synthesis. Speech Commun. 19, 1–19.

Dang, J., Honda, K., 1996a. An improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling. In: Proceedings of the International Congress on Speech and Language Processing, pp. 965–968.

Dang, J., Honda, K., 1996b. Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. J. Acoust. Soc. Am. 100 (5), 3374–3383.

Dang, J., Honda, K., 1997. Acoustic characteristics of the piriform fossa in models and humans. J. Acoust. Soc. Am. 101 (1), 456–465.

Dang, J., Honda, K., 2004. Construction and control of a physiological articulatory model. J. Acoust. Soc. Am. 115 (2), 853–870.

Dang, J., Wei, J., Honda, K., 2016. A study on transvelar coupling for non-nasalized sounds. J. Acoust. Soc. Am. 139 (1), 441–454.

Delvaux, B., 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.

Deng, L., dynamic, A., 1998. Feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Commun. 24, 299–323.

van den Doel, K., Ascher, U.M., 2008. Real-time numerical solution of webster's equation on a nonuniform grid. IEEE Trans. Audio, Speech, Language Process. 16 (6), 1163–1172.

Elie, B., Laprie, Y., 2016. Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. Speech Commun. 82, 85–96.

Erath, B.D., Zañartu, M., Stewart, K.C., Plesniak, M.W., Sommer, D.E., Peterson, S.D., 2013. A review of lumped-element models of voiced speech. Speech Commun. 55 (5), 667–690.

Fant, G., Nord, L., Branderud, P., 1976. A note on the vocal tract wall impedance. STL-QPSR 4, 13–20.

Flanagan, J.L., Ishizaka, K., Shipley, K.L., 1975. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell Syst. Tech. J. 54 (3), 485–506.

Fleischer, M., Pinkert, S., Mattheus, W., Mainka, A., Mürbe, D., 2015. Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. Biomech. Model. Mechanobiol. 14 (4), 719–733.

Freixes, M., Arnela, M., Socoró, J.C., Alías, F., Guasch, O., 2019. Glottal source contribution to higher order modes in the finite element synthesis of vowels. Appl. Sci. 9 (21), 4535.

Fujita, S., Honda, K., 2005. An experimental study of acoustic characteristics of hypopharyngeal cavities using vocal tract solid models. Acoust. Sci. Technol. 26 (4), 353–357.

Godoy, E., Dumas, A., Melot, J., Malyska, N., Quatieri, T.F., 2016. Relating estimated cyclic spectral peak frequency to measured epilarynx length using magnetic resonance imaging. In: Proc. of the Interspeech, pp. 948–952.

Ishizaka, K., Flanagan, J.L., 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Syst. Tech. J. 51 (6), 1233–1268.

Iskarous, K., Goldstein, L., Whalen, D.H., Tiede, M., Rubin, P., 2003. CASY: The haskins configurable articulatory synthesizer. In: International Congress of Phonetic Sciences. pp. 185–188.

Kitamura, T., Honda, K., Takemoto, H., 2005. Individual variation of the hypopharyngeal cavities and its acoustic effects. Acoust. Sci. Technol. 26 (1), 16–26.

Kröger, B.J., 1993. A gestural production model and its application to reduction in German. Phonetica 50, 213–233.

Kröger, B.J., 1998. Ein Phonetisches Modell der Sprachproduktion. Niemeyer, Tübingen.

Kröger, B.J., Hoole, P., Sader, R., Geng, C., Pompino-Marschall, B., Neuschaefer-Rube, C., 2004. MRT-sequenzen als datenbasis eines visuellen artikulationsmodells. HNO 52, 837–843.

Liu, L.-J., Ding, C., Jiang, Y., Zhou, M., Wei, S., 2017. The IFLYTEK system for blizzard challenge 2017. In: Blizzard Challenge Workshop.

Maeda, S., 1982. A digital simulation method of the vocal-tract system. Speech Commun. 1, 199–229.

Maeda, S., 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Boston, pp. 131–149.

Meltzner, G.S., Kobler, J.B., Hillman, R.E., 2003. Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices. J. Acoust. Soc. Am. 114 (2), 1035–1047.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am. 53 (4), 1070–1082.

Monson, B.B., Hunter, E.J., Lotto, A.J., Story, B.H., 2014. The perceptual significance of high-frequency energy in the human voice. Front. Psychol. 5, 587.

Murphy, D.T., Jani, M., Ternström, S., 2015. Articulatory vocal tract synthesis in supercollider. In: 18th Int. Conference on Digital Audio Effects (DAFx-15). pp. 1–7.

Okadome, T., Honda, M., 2001. Generation of articulatory movements by using a kinematic triphone model. J. Acoust. Soc. Am. 110 (1), 453–463.

Payan, Y., Perrier, P., 1997. Synthesis of V-V sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis. Speech Commun. 22, 185–205.

Piepho, H.-P., 2018. Letters in mean comparisons: what they do and don't mean. Agron. J. 110 (2), 431–434.

Pont, A., Guasch, O., Arnela, M., 2020. Finite element generation of sibilants /s/ and /z/ using random distributions of kirchhoff vortices. Int. J. Numer. Methods Biomed. Eng. 36 (2), e3302.

Saltzman, E.L., Munhall, K.G., 1989. A dynamic approach to gestural patterning in speech production. Ecol. Psychol. 1, 333–382.

Shadle, C.H., Damper, R.I., 2001. Prospects for articulatory synthesis: A position paper. In: Fourth ISCA Tutorial and Research Workshop on Speech Synthesis, Pitlochry, Scotland, pp. 121–126.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783.

Sondhi, M.M., Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. Acoust. Speech Signal Process. 35 (7), 955–967.

Stavness, I., Lloyd, J.E., Payan, Y., Fels, S., 2011. Coupled hard–soft tissue simulation with contact and constraints applied to jaw–tongue–hyoid dynamics. Int. J. Numer. Methods Biomed. Eng. 27 (3), 367–390.

Stevens, K.N., 1998. Acoustic Phonetics. The MIT Press.

Stone, S., Marxen, M., Birkholz, P., 2018. Construction and evaluation of a parametric one-dimensional vocal tract model. IEEE/ACM Trans. Audio Speech Language Process. 26 (8), 1381–1392.

Story, B.H., Bunton, K., 2019. A model of speech production based on the acoustic relativity of the vocal tract. J. Acoust. Soc. Am. 146 (4), 2522–2528.

Story, B.H., Vorperian, H.K., Bunton, K., Durtschi, R.B., 2018. An age-dependent vocal tract model for males and females based on anatomic measurements. J. Acoust. Soc. Am. 143 (5), 3079–3102.

Suzuki, H., Nakai, T., Dang, J., Lu, C., 1990. Speech production model involving subglottal structure and oral-nasal coupling through closed velum. In: First International Conference on Spoken Language Processing (ICSLP 1990). pp. 437–440.

Takemoto, H., Adachi, S., Mokhtari, P., Kitamura, T., 2013. Acoustic interaction between the right and left piriform fossae in generating spectral dips. J. Acoust. Soc. Am. 134 (4), 2955–2964.

Takemoto, H., Mokhtari, P., Kitamura, T., 2010. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. J. Acoust. Soc. Am. 128 (6), 3724–3738.

Teixeira, A.J.S., Martinez, R., Silva, L.N., Jesus, L.M.T., Principe, J.C., Vaz, F.A.C., 2005. Simulation of human speech production applied to the study and synthesis of European portuguese. EURASIP J. Appl. Signal Process. 9, 1435–1448.

Titze, I.R., 1989. A four-parameter model of the glottis and vocal fold contact area. Speech Commun. 8, 191–201.

Toutios, A., Ouni, S., Laprie, Y., 2011. Estimating the control parameters of an articulatory model from electromagnetic articulograph data. J. Acoust. Soc. Am. 129 (5), 3245–3257.

Vampola, T., Horáček, J., Radolf, V., Švec, J.G., Laukkanen, A.-M., 2020. Influence of nasal cavities on voice quality: Computer simulations and experiments. J. Acoust. Soc. Am. 148 (5), 3218–3231.

Vampola, T., Horáček, J., Švec, J.G., 2015. Modeling the influence of piriform sinuses and valleculae on the vocal tract resonances and antiresonances. Acta Acust. United Acust. 101 (3), 594–602.

Švec, J.G., Titze, I.R., Popolo, P.S., 2005. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. J. Acoust. Soc. Am. 117 (3), 1386–1394.

Wu, L., Xiao, K., Dong, J., Wang, S., Wan, M., 2014. Measurement of the sound transmission characteristics of normal neck tissue using a reflectionless uniform tube. J. Acoust. Soc. Am. 136 (1), 350–356.

Xu, Y., Liu, F., alignment, Tonal., 2006. Tonal alignment syllable structure and coarticulation: Toward an integrated model. Italian J. Linguist. 18, 125–159.

Zhang, J., Honda, K., Wei, J., Kitamura, T., 2019. Morphological characteristics of male and female hypopharynx: A magnetic resonance imaging-based study. J. Acoust. Soc. Am. 145 (2), 734–748.

Zhang, C., Honda, K., Zhang, J., Wei, J., 2016. Contributions of the piriform fossa of female speakers to vowel spectra. In: 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 1–5.