Modeling Speech Sound Radiation With Different Degrees of Realism for Articulatory Synthesis

PETER BIRKHOLZ, (Member, IEEE), STEFFEN OSSMANN, RÉMI BLANDIN, ALEXANDER WILBRANDT, PAUL KONSTANTIN KRUG, AND MARIO FLEISCHER

1 Institute of Acoustics and Speech Communication, TU Dresden, 01062 Dresden, Germany
2 Department of Otorhinolaryngology, Head and Neck Surgery, Carl Gustav Carus Faculty of Medicine, TU Dresden, 01307 Dresden, Germany
3 Department of Audiology and Phoniatrics, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, 10117 Berlin, Germany

Corresponding author: Peter Birkholz (peter.birkholz@tu-dresden.de)

This work involved human subjects in its research. It was conducted according to the ethical principles based on the WMA Declaration of Helsinki.

ABSTRACT
Articulatory synthesis is based on modeling various physical phenomena of speech production, including sound radiation from the mouth. With regard to sound radiation, the most common approach is to approximate it in terms of a simple spherical source of strength equal to the mouth volume velocity. However, because this approximation is only valid at very low frequencies and does not account for the diffraction by the head and the torso, we simulated two alternative radiation characteristics that are potentially more realistic: the radiation from a vibrating piston in a spherical baffle, and the radiation from the mouth of a detailed model of the human head and torso. Using the articulatory speech synthesizer VocalTractLab, a corpus of 10 sentences was synthesized with the different radiation characteristics combined with three different phonation types. The synthesized sentences were acoustically compared with natural recordings of the same sentences in terms of their long-term average spectra (LTAS), and evaluated in terms of their naturalness and intelligibility. The intelligibility was not affected by the type of radiation characteristic. However, it was found that the more similar their LTAS was to real speech, the more natural the synthetic sentences were perceived to be. Hence, the naturalness was not directly determined by the realism of the radiation characteristic, but by the combined spectral effect of the radiation characteristic and the voice source. While the more realistic radiation models do not per se improve synthesis quality, they provide new insights in the study of speech production and articulatory synthesis.

INDEX TERMS
Radiation characteristic, articulatory synthesis.

I. INTRODUCTION
Along-standing issue in speech research is the creation of articulatory synthesizers, i.e. simulations of speech production that are capable of generating natural-sounding speech [1], [2], [3], [4], [5], [6]. These synthesizers have a number of different applications, including text-to-speech synthesis [7], [8], the simulation of speech acquisition [9], [10], [11], [12], and the study of paralinguistic effects [13], [14]. However, currently the quality of articulatory synthesis is not yet competitive with the best commercial unit-selection or neural speech synthesizers [15]. The key to improving the speech quality is the identification and improvement of potential shortcomings of the currently employed simulation models.

This study explored the effect of the radiation characteristic on the quality of the synthesized speech. Usually, acoustic models of the vocal tract generate the volume velocity at the lips and the nostrils (e.g. [16], [17]). Here, the radiating areas of the mouth and the nostrils are considered as vibrating...
surfaces, all parts of which move in phase. The radiation characteristic is the transfer function from this volume velocity to the sound pressure that is received by a microphone or human ear outside the vocal tract.

Most articulatory synthesizers model this sound radiation analogously to the radiation of a simple point source that radiates uniformly in all directions and whose strength corresponds to the volume velocity at the mouth (or nostrils) (e.g. [18], [19], [20]). This situation is illustrated in Figure 1a (where the radius of the sphere goes to zero). The radiation characteristic \( H_{\text{simple}}(\omega) \) for this case (henceforth “simple source”) has the closed-form expression [17], [21], [22]

\[
H_{\text{simple}}(\omega) = \frac{P_{\text{rad}}(\omega)}{U_0(\omega)} = \frac{j \omega \varrho_0}{4 \pi r} e^{-j \omega r/c},
\]

where \( P_{\text{rad}}(\omega) \) is the sound pressure at a distance \( r \) from the source, \( U_0(\omega) \) is the volume velocity generated by the source, \( \omega \) is the angular frequency, \( \varrho_0 \) is the ambient density of air, \( c \) is the sound velocity, and \( j \) is the imaginary unit. Hence, for a constant \( U_0 \) and \( r \), the source generates a sound pressure that has an amplitude proportional to \( \omega \), which corresponds to a slope of 6 dB/oct of the radiation characteristic. In the time domain, the factor \( j \omega \) on the right-hand side corresponds to the first derivative with respect to time, and the exponential function corresponds to a time delay of the signal by \( r/c \), so that

\[
P_{\text{rad}}(t) = \frac{\varrho_0}{4 \pi r} \frac{\partial U_0(t - r/c)}{\partial t}.
\]

Since the constant time delay usually plays no role for the synthesized speech signal, acoustic simulations of the vocal tract simply take the first time derivative of the volume velocity through the lips and nostrils to obtain the radiated sound pressure.

The simple model of a spherical source is an approximation that applies only to a small radiation area (mouth opening) and to low frequencies, where the wavelength is long compared to the head diameter. Although this is acknowledged in textbooks on speech acoustics [17], [22], the potential effects of this constraint on the naturalness of synthesized speech have not been studied. For a more precise radiation characteristic, Stevens [17] proposes to use that of a vibrating piston (corresponding to the mouth opening) in a spherical baffle (corresponding to the head), as illustrated in Figure 1b. Although Stevens did not provide the mathematical formulation of this model, he showed that high frequencies are emphasized by up to 5 dB compared to the simple radiation model on the axis of the piston. Using this radiation model instead of the simple source model would hence change the spectral balance between high and low frequency components of the synthesized speech.

However, because the model of a piston in a sphere is still a gross simplification (e.g., the precise lip shape and the torso are not considered), it is unclear how well it really describes sound radiation from the mouth. Measurements of speech radiation by Flanagan [23] using a life-size mannequin with a calibrated volume velocity source at its lips indicate that the slope of the radiation characteristic is about 7 dB/oct for a microphone 30 cm in front of the mouth (instead of 6 dB/oct for the simple source), which is similar to the radiation from a piston in a sphere. Unfortunately, these measurements were performed for only nine discrete frequency points spaced in 6-semitone steps between 250 Hz and 4000 Hz. There are a number of other studies on the radiation of speech (e.g. [24], [25]), but they have focused on the relative changes of sound pressure levels with varying microphone positions, i.e., on speech directivity patterns. To our knowledge, there are no published transfer functions from volume velocity at the lips to sound pressure in front of the mouth for realistic head and torso geometries with high frequency resolution and up to high frequencies.

The present study had two goals. The first goal was to determine detailed radiation characteristics for the piston-insphere model and for a realistic head-torso geometry, and to derive finite-impulse-response filters for their convenient implementation in articulatory synthesizers. These two radiation characteristics represent two alternatives to the simple source model with different degrees of realism.

The second goal was to explore the effect of the different radiation characteristics on the perceived naturalness of synthesized speech. Since the radiation characteristic affects the long-term spectral pattern of the speech signal in a similar way as the glottal source signal, the types of radiation characteristics were assessed in combination with three different phonation types. Therefore, a set of ten sentences was synthesized for each combination of three radiation characteristics (simple source model and the two alternatives) and three phonation types (breathy, modal and pressed phonation) using the articulatory synthesizer VocalTractLab (www.vocaltractlab.de), and assessed in a perception experiment. Finally, the different variants of the synthesized sentences were acoustically compared with their naturally spoken counterparts in terms of long-term average spectra, and evaluated with respect to their recognition by automatic speech recognizers (as a proxy for their intelligibility).

II. METHODS
This section starts with modeling sound radiation from a vibrating piston in a sphere, and from the mouth of a realistic head-torso model. The two obtained transfer functions are
then approximated by digital finite-impulse response filters. Then we describe the recording and analysis of a small corpus of sentences spoken by multiple speakers, and the articulatory re-synthesis of this corpus in nine variants, i.e. with three different methods for the sound radiation times three different phonation types. Finally we present the perceptual assessment of the naturalness of the nine variants of synthesized sentences, and the assessment of their intelligibility in terms of the word error rates of automatic speech recognizers.

A. MODELING THE RADIATION FROM A CIRCULAR PISTON IN A SPHERICAL BAFFLE

The radiation from a circular piston in a spherical baffle is illustrated in Figure 1b, where the piston is centered on the z-axis. According to Williams [26, Eq. 6.134], the sound pressure at the angle $\theta$ (relative to the piston axis) and the distance $r$ from the center of the sphere (point $Q$ in Figure 1b) is then given by

$$P_{\text{rad}}(\omega) = -\frac{j\varrho c \tilde{W}}{2} \sum_{n=0}^{\infty} [P_{n-1}(\cos \alpha) - P_{n+1}(\cos \alpha)]$$

$$\times \left( \frac{h_n(kr)}{h'_n(ka)} P_n(\cos \theta) \right),$$

where $\tilde{W}$ is the velocity of the piston, $a$ is the radius of the head, $\alpha$ is the angle between the piston axis and the circular edge of the piston on the surface of the sphere, $k = \omega/c$ is the wave number, $P_n(\cdot)$ is the Legendre function of the first kind of order $n$, $h_n(\cdot)$ is the spherical Hankel function of the 2nd kind of order $n$, and $h'_n(\cdot) = dh_n(\cdot)/dx$ is its first derivative. The Legendre polynomials of order $n$ can be effectively calculated with the recursion formula

$$P_n(x) = \frac{2n - 1}{n} x P_{n-1}(x) - \frac{n - 1}{n} P_{n-2}(x), \quad n \geq 2$$

$$P_1(x) = x, \quad P_0(x) = 1.$$  

For the special case of $n = 0$ in Eq. (3), $P_{n-1}(\cdot) = P_{-1}(\cdot) = 1$. The spherical Hankel function is given by [26, Eq. 6.57]

$$h_n(x) = \sqrt{\frac{\pi}{2x}} (J_{n+1/2}(x) - jY_{n+1/2}(x)),$$

where $J_n(x)$ is the Bessel function of the first kind and $Y_n(x)$ is the Bessel function of the 2nd kind. The derivative $h'_n(x)$ can be calculated as

$$h'_n(x) = \frac{n}{x} h_n(x) - h_{n+1}(x).$$

In the present study we assume that the microphone or the ear of the listener is on the piston axis, so that $\theta = 0$. In this case, the factor $P_n(\cos \theta)$ in Eq. (3) becomes 1 (because $P_n(1) = 1$ for all $n$). Furthermore, when the mouth is small compared to the head, i.e. when $\alpha$ is small, then the mouth area is $A_{\text{mouth}} \approx \pi a^2 \alpha^2$, where $a$ is the radius of the sphere. This can be used to write the piston velocity $\tilde{W}$ in terms of the volume velocity $U_0$ generated by the piston, i.e. as $\tilde{W} = U_0/A_{\text{mouth}}$, which gives the radiation characteristic as

$$H_{\text{piston}}(\omega) \approx \frac{P_{\text{rad}}(\omega)}{U_0(\omega)}$$

$$= -\frac{j\varrho c}{2A_{\text{mouth}}} \sum_{n=0}^{\infty} [P_{n-1}(\cos \alpha) - P_{n+1}(\cos \alpha)]$$

$$\times \left( \frac{h_n(kr)}{h'_n(ka)} P_n(\cos \theta) \right).$$

We implemented this equation in Matlab R2021a (Mathworks, Inc.) to obtain the radiation characteristics for different distances between the “mouth” and the microphone (located on the axis of the piston), and for different areas of the piston. The ambient density of air and the sound velocity were set to $\varrho_0 = 1.225$ kg/m$^3$ and $c = 343.2$ m/s (sound velocity at 20°C), respectively. Following Flanagan [22], the radius of the sphere (head) was set to $a = 9$ cm. For the Bessel functions $J_n(x)$ and $Y_n(x)$, the Matlab functions $\text{besselj}$ and $\text{bessely}$ were used.

The upper limit for the sum in Eq. (6) was set to $n_{\text{max}} = 30$. With this value, the calculated solution was stable (without ringing artifacts) up to a frequency of 12 kHz. Figure 2a shows $|H_{\text{piston}}(\omega)|$ for the distances $d$ of 15 cm (black curve), 30 cm (gray curve), and 60 cm (light gray curve). In accordance with the measurements by Flanagan [23], the slope of $H_{\text{piston}}(\omega)$ is about 7 dB/oct for frequencies...
The closed surface of this domain was partitioned into four patches based on their different acoustic boundary conditions: the lip region from where the sound waves are radiated, the surface of the head and torso that partly absorb and partly reflect the sound waves, the non-reflecting outer boundary of the sphere, and the fully reflective midsagittal symmetry plane. The lip patch is shown in Figure 4c and has an area of about 3.4 cm², which is close to that of the radiating piston in the piston-in-sphere model.

The FE simulation was performed with the development version of the open-source software dolfinx (https://jorgensd.github.io/dolfinx-tutorial/). It requires as input a 3D volume mesh of the simulation domain and a definition of the boundary conditions. The 3D mesh was created with the software Gmsh [27] (www.gmsh.info) based on the boundary representation of the simulation domain, and the boundary conditions were defined by means of the patches. The acoustic wave propagation in the simulation domain was determined by solving the complex-valued Helmholtz equation

\[ -(k^2 + \nabla^2)P(x, \omega) = 0 \quad x \in \Omega, \]  

where \( k = \omega/c \) is the wave number, \( \nabla \) is the nabla operator, \( P \) is the acoustic pressure, \( x \) is the position, and \( c = 343.2 \text{ m/s} \) is the speed of sound (same as before).

At the lips (the radiating surface), we implemented a Neumann boundary condition

\[ \nabla P \cdot n = -j\omega \rho_0 v_0 \quad x \in \partial \Omega_{\text{Lips}}, \]  

where \( n \) is the outside-pointing surface normal vector, \( v_0 \) is a constant particle velocity, and \( \rho_0 = 1.225 \text{ kg/m}^3 \) (as before).

At the surface of the head and torso, we applied the Robin boundary condition

\[ \nabla P \cdot n = -jk \frac{\rho_0 c}{Z} P \quad x \in \partial \Omega_{\text{Body}} \]  

with the surface impedance \( Z = 500 \rho_0 c \). This impedance makes the body surface mostly reflective with small acoustic losses (absorption) comparable to the losses in the vocal tract [28], [29]. To make the surface of the sphere non-reflective, a 3rd-order infinite-impedance element according to [30] was applied. Finally, the symmetry plane was forced to be hard-walled.

The trial and test functions in the variational form were approximated by Laplacian polynomials of 2nd order. The resulting degree of freedom was on the order of 2.6 million for an element size of about 6 mm. The linear algebraic system was solved on a high-performance computer with 96 cores for 480 frequency points between 25 Hz and 12 kHz (increment of 25 Hz). The total computing time for all frequency points was about 6 hours.

For each frequency point, the pressure \( P_{\text{rad}}(\omega) \) was picked up at the mesh node closest to 30 cm in front of the lips (see Figure 4b). Since the simulated volume velocity at the lips was constant and frequency-independent, the obtained

![FIGURE 3. Deviation of the (unwrapped) phase of the piston-in-sphere radiation model (gray curve) and the head-torso radiation model (black line) from the simple radiation model.](image-url)
FIGURE 4. a) Original geometry of the head-torso model. b) The model was intersected with a sphere of 40 cm radius centered around the mouth opening, and the region inside the sphere and outside the body was transformed into an FEM mesh. c) The red region marks the radiating part of the model surface.

FIGURE 5. a) Radiation characteristics for the simple source (dashed line) and the finite-element head-torso model for a mouth-microphone distance of 30 cm. b) Deviation of the radiation characteristic of the head-torso model from the simple source model. Pressure values represent the radiation characteristic

\[ H_{\text{FEM}}(\omega) = P_{\text{rad}}(\omega)/U_{0}(\omega). \]  

(10)

This transfer function is shown in Figure 5a and has been scaled to match the magnitude of \( H_{\text{simple}}(\omega) \) at the frequency of 100 Hz. As stated above, the FEM calculations were performed only for frequencies up to 12 kHz. For higher frequencies, the wavelengths become too short compared to the size of the finite elements and would make the results increasingly inaccurate. For frequencies above 12 kHz, \( H_{\text{FEM}}(\omega) \) was extrapolated with a constant slope of 6 dB/oct and an offset of 2.2 dB above \( |H_{\text{simple}}(\omega)| \). The offset corresponds to the average difference between \( H_{\text{FEM}}(\omega) \) and \( H_{\text{simple}}(\omega) \) between 8 kHz and 12 kHz. In the 8–12 kHz band, the magnitude and phase of \( H_{\text{FEM}}(\omega) \) was blended between the calculated FEM values and the extrapolated section.

Figure 5b shows the deviation of \( H_{\text{FEM}}(\omega) \) from the simple source model. Like for the piston-sphere model, \( H_{\text{FEM}}(\omega) \) generally boosts the higher frequencies compared to the simple radiation model, but the spectrum is much more complicated and indicates a complex diffraction pattern around the realistic head and torso geometry. As is known from studies on speech directivity [31], [32], this pattern is also likely to change significantly when the position of the receiver changes, e.g. to a position to the side of the speaker. The complex diffraction pattern is also indicated by the phase of \( H_{\text{FEM}}(\omega) \) in Figure 3 (solid black line). However, overall the phase of \( H_{\text{FEM}}(\omega) \) is still similar to that of the simple model.

To check the general correctness of the FEM simulation, it was additionally performed for the geometric situation of a piston in a sphere, analogous to the case described in Sec. II-A. The obtained transfer function (up to 12 kHz) is shown as the black dotted line in Figure 2b and agrees very well with the explicit solution for \( H_{\text{piston}}(\omega) \).

C. FINITE IMPULSE RESPONSE FILTERS FOR THE RADIATION CHARACTERISTICS

To apply the more realistic radiation characteristics \( H_{\text{piston}}(\omega) \) and \( H_{\text{FEM}}(\omega) \) to articulatory speech synthesis, they must be converted into digital filters \( H_{\text{piston}}(z) \) and \( H_{\text{FEM}}(z) \), where \( z \) is the complex frequency of the \( z \)-domain. With regard to the filter type, infinite-impulse-response (IIR) filters can generally approximate a prescribed amplitude spectrum with a lower filter order than a corresponding finite-impulse-response (FIR) filter, but provide little or no control over
with the transfer function

\[
H(\omega) = j \omega \quad \text{(dashed line)}
\]

At high frequencies, its magnitude starts to deviate from its analog counterpart with the transfer function \(H(\omega) = j \omega\), showing the effect of frequency warping of the digital filter at high frequencies (bending of the curve). The solid black lines in Figures 6b and c show \(H'(\omega)\) for the piston-in-sphere model and for the head-torso model, respectively. Note that these spectra differ from those in Figures 2b and 5b in the high-frequency part due to the frequency warping mentioned above. With the following approximation of \(H'(\omega)\), this warping is automatically compensated for.

To approximate \(H'(\omega)\) by an FIR filter, we exploit the fact that an FIR filter

\[
H'(z) = a_0 + a_1 z^{-1} + \ldots + a_M z^{-M}
\]

with an even order \(M\) becomes a linear-phase filter when its impulse response is symmetric, i.e., \(a_0 = a_M, a_1 = a_{M-1}\), etc. In this case, the system function can be written as

\[
H'(z) = z^{-M/2} \sum_{m=1}^{M/2} a_{M/2-m} (z^m + z^{-m}).
\]

The phase response. Hence, we created FIR filters for a good approximation of both the amplitude and phase response.

Given that the phase responses of \(H_{\text{piston}}(\omega)\) and \(H_{\text{FEM}}(\omega)\) do not strongly deviate from the simple source model \(H_{\text{simple}}(\omega)\) (Figure 3), the FIR filters were generated with the phase response of the simple radiation characteristic. This was achieved by the concatenation of a first-order digital high-pass filter that represents the radiation characteristic of the simple source model, i.e.,

\[
H_{\text{simple}}(z) = z - 1,
\]

and of a zero-phase digital filter \(H'(z)\) that approximates the deviation of \(H_{\text{piston}}(\omega)\) (analogously \(H_{\text{FEM}}(\omega)\)) from \(H_{\text{simple}}(z)\). Since the frequency response of \(H_{\text{simple}}(z)\) is given by \(e^{j\omega \Delta t} - 1\) with the continuous frequency variable \(\omega\) and the time step \(\Delta t\), the filter to be approximated is given by

\[
H'(\omega) = H_{\text{piston}}(\omega)/(e^{j \omega \Delta t} - 1).
\]

The solid black line in Figure 6a illustrates the amplitude spectrum of the digital filter \(H_{\text{simple}}(z)\) for a sampling rate \(f_s = 1/\Delta t = 44100\) Hz. Compared to the corresponding “analog” first-order high-pass filter with the transfer function \(H(\omega) = j \omega\) (dashed line) with the constant slope of 6 dB/oct, we see the effect of frequency warping of the digital filter at high frequencies (bending of the curve). The solid black lines in Figures 6b and c show \(H'(\omega)\) for the piston-in-sphere model and for the head-torso model, respectively. Note that these spectra differ from those in Figures 2b and 5b in the high-frequency part due to the frequency warping mentioned above.

The head-torso radiation characteristic from the (digital) simple source model, i.e.,

\[
H(\omega)(\text{digital}) = e^{-j \omega \Delta t} H(\omega)(\text{simple source})
\]

starts to deviate from its analog counterpart with the transfer function \(H(\omega) = j \omega\) (dashed line). The approximation of both the amplitude and phase response. Hence, we created FIR filters for a good characteristic (black line) and its approximation by an FIR filter (gray line).
complicated) head-torso spectrum. For lower sampling rates, the required filter order is proportionally lower.

As stated above, the final filter \( H(z) \) for the radiation characteristic (corresponding to either \( H_{\text{piston}}(z) \) or \( H_{\text{FEM}}(z) \)) was obtained by the concatenation of \( H'(z) \) with the filter \((z - 1)\). For the final filter to be causal, we have to multiply with an additional \( z^{-1} \), so that

\[
H(z) = H'(z) \cdot (z - 1) \cdot z^{-1} = H'(z) \cdot (1 - z^{-1}) = a_0 + (a_1 - a_0)z^{-1} + \ldots + (a_M - a_{M-1})z^{-M} + (-a_M)z^{-(M+1)},
\]

with the impulse response

\[
h(k) = [a_0, a_1 - a_0, a_2 - a_1, \ldots, a_M - a_{M-1}, -a_M].
\]

This filter has the order \( M + 1 \) and the length \( M + 2 \).

The supplemental material contains the coefficients of the FIR filter approximations \( H_{\text{piston}}(z) \) and \( H_{\text{FEM}}(z) \) for the most common sampling rates, so that they can be readily employed in synthesis models.

D. TEXT CORPUS

For the subsequent speech recordings and their articulatory re-synthesis with the different radiation characteristics and phonation types, we created a corpus of ten German sentences that are given in Table 1. The sentences were designed to include mainly voiced phonemes, so that the long-term average spectrum [33], [34] of these sentences (when recorded or synthesized) essentially represent the combined effects of the voice source signal and the radiation characteristic.

E. SPEECH RECORDINGS AND ANALYSIS

As a basis for the articulatory re-synthesis and for the subsequent acoustic assessment of the synthesized sentences, we made high-quality acoustic recordings of the ten sentences, read aloud by each of eight male native German speakers (27–63 years) in a neutral speaking style. The recording conditions were as similar as possible to the conditions under which the radiation characteristics were simulated. The speakers stood in the center of a soundproofed audio studio and maintained a distance of 30 cm to a measurement microphone (MK 250 capsule with the MV 210 amplifier by Microtech Gefell GmbH) positioned on the radiation axis (with a pop shield in front of the microphone). The microphone was connected to a power supply (CCP Supply Type 12AL by G.R.A.S.), an audio interface (TASCAM UH-7000) and a laptop computer running the recording software Audacity 2.3.3 (https://www.audacity.de/). Both the measurement microphone and the amplifiers have a near-perfect flat frequency response up to 10 kHz, and a maximum deviation of 1 dB from the flat response between 10–12 kHz. The recordings were made with a sampling rate of 44100 Hz and 16 bit quantization.

As a baseline for the acoustics assessment of the synthesized sentences, one long-term average spectrum (LTAS) was calculated from the utterances of each speaker. To this end, all ten sentences of the same speaker were peak-normalized and connected to one long audio file. The silent pauses before and after the sentences were removed, because they would otherwise affect the overall LTAS. The LTAS were calculated as the average of the power spectra of overlapping frames of 23.2 ms (=1024 samples) with 50% overlap between frames using a Hamming window. The power spectrum of each frame was calculated as

\[
P(n) = \frac{1}{N} \left[ \sum_{k=0}^{N-1} w(k)x(k)e^{-j2\pi kn/N} \right]^2,
\]

where \( N = 1024 \) is the frame length in samples, \( n \) is the frequency index, \( k \) is the time index, \( w(k) \) is the Hamming window, and \( x(k) \) is the audio signal of the frame. Each LTAS was normalized in such a way that the maximum amplitude was 0 dB.

F. ARTICULATORY SYNTHESIS AND ANALYSIS OF STIMULI

To assess how the different models for the radiation characteristic affect the perceived naturalness of synthetic speech, the ten sentences in Table 1 were synthesized with different settings and then used in a perception experiment (Sec. II-G). The synthesis was performed with an extended version of the articulatory synthesizer VocalTractLab 2.3 (www.vocaltractlab.de) where the different radiation models have been integrated. The synthesizer is based on a 3D-model of the vocal tract of an adult male speaker [1], a geometric model of the vocal folds [35], an aero-acoustic simulation [36], [37], and a gestural control model [38]. Since version 2.3, the synthesizer allows creating gestural scores automatically from a sequence of phones and the corresponding phone durations. This feature was used here to re-create the 10 sentences of one of the real speakers with the phone durations from the natural recordings. Along with the phone durations, the pitch contours of the natural sentences were closely copied with automatically determined pitch targets [39]. Hence, the prosody of the synthetic utterances closely matched that of the natural utterances.

For each of the 10 sentences, 9 variants were synthesized, i.e. for all combinations of three phonation types (using breathy, modal, and pressed voice) and the three models for the radiation characteristic. This allowed us to study a potential interaction between the phonation type and the radiation characteristic, as both features affect the long-term spectral balance of the speech signal. For all variants of the same sentence, the same (copied) phone durations and pitch contours were used. The three phonation types were implemented in terms of the corresponding glottal shapes defined for the geometric vocal fold model in the gestural scores [35]. The glottal shapes mainly differ in terms of the degree of glottal abduction and the size of a constant posterior chink. For breathy, modal, and pressed phonation, the rest displacement

https://www.vocaltractlab.de/index.php?page=birkholz-supplements
of the vocal folds from the glottal midline at the posterior end was 0.2 (0.3) mm, 0.1 (0.2) mm, and 0.0 (0.1) mm at the lower (upper) vocal fold edges, respectively. The chink areas for breathy, modal, and pressed phonation were 10 mm², 5 mm², and 2.5 mm², respectively. To illustrate the effect of the different glottal settings, Figure 7 shows one period of the simulated glottal flow during the synthesis of /a/ for the three phonation types. Consistent with estimates of the chink areas, the closing quotient increases from pressed to breathy phonation. In addition, there is an increasing DC offset of the glottal flow with increasing glottal rest area, which leads to increasing aspiration noise from pressed to breathy phonation.

The audio signals of the 90 synthesized utterances had a sampling rate of 44100 Hz and were quantized with 16 bit per sample. Using an 8th-order Chebyshev filter, the synthetic speech signals were band-limited to 12 kHz, which corresponds to the upper frequency limit of both the piston-in-sphere and head-torso simulations of the radiation characteristic. With regard to the subsequent perception experiment, all synthesized utterances were loudness-normalized analogous to Krug et al. [15] using the Python library pyloudnorm (https://github.com/csteinmetz1/pyloudnorm), because loudness differences might otherwise impact the psychometric ratings [41]. All normalized synthetic stimuli are included in the supplemental material.

Analogous to the recorded utterances, one LTAS was calculated across all 10 sentences for each of the 9 synthesized variants. Like the LTAS of the natural recordings, they were normalized for a maximum value of 0 dB. To quantify the differences between the LTAS of the synthetic and natural utterances, for each of the 9 synthetic LTAS, the root-mean-square deviation from the average natural LTAS (using differences on the dB scale) was calculated across the frequencies from 0–12 kHz.

G. LISTENING EXPERIMENT

The listening experiment was performed to assess potential differences in the naturalness of the synthesized sentences due to the different radiation characteristics. Thirty native speakers of German (18–41 years old, mean: 28 years; 17 male and 13 female) participated in the experiment after giving their informed consent. The participants performed the experiment individually in an audio studio where the stimuli were presented over high-quality closed headphones (STAX SR-202 Basic) that were connected to a headphone amplifier (STAX SRM-212) and a laptop computer. Since the perceptual differences due to different radiation characteristics were assumed to be small, the naturalness was assessed in a forced-choice pairwise comparisons of the stimuli. For each of the 10 sentences in each of the 3 phonation types, 3 pairs were created to compare the radiation models (simple vs. piston-in-sphere model; simple vs. head-torso model; piston-in-sphere vs. head-torso model). Hence, there were 90 different pairs of stimuli in total. To assess the consistency of the responses, each pair was presented twice, requiring participants to compare 180 pairs. The pairs were presented in an individually randomized order for each participant. Between the two stimuli of a pair, a short pause of 500 ms was included. 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. Playback of each pair could be repeated once on demand. After a decision, the stimuli of the next pair were automatically played.

H. AUTOMATIC SPEECH RECOGNITION

To assess the intelligibility of the synthetic utterances depending on the settings for the phonation type and the radiation characteristic, we used automatic speech recognition in a similar way as in the study by Krug et al. [15]. Three different state-of-the-art speech-to-text web services (Google

### TABLE 1. List of the 10 German sentences used in the experiments.

| #  | German sentence                  | Transcription              | Translation to English          |
|----|----------------------------------|----------------------------|---------------------------------|
| 1  | Am Himmel sind nur wenige Wolken. | /am lumol zim nuv venega volkon/ | There are only a few clouds in the sky. |
| 2  | Er mag die Orangen.              | /e muck di ooragen/         | He likes the oranges.           |
| 3  | Warum ging die Sirene los?       | /vaurn gum di siiren los/    | Why did the siren go off?      |
| 4  | Wieso wurden alle müde?         | /vizzu vurden ala myida/     | Why did everyone get tired?     |
| 5  | Leon will nur sieben Mandarinen. | /leon vil nuv ziben mandariniun/ | Leon only wants seven mandarins. |
| 6  | Ina wird nie Samurai.            | /ina vri ni samurai/         | Ina will never become a samurai. |
| 7  | Wir gehen bei grünem Signal.     | /vire gean bi grunen signal/ | We go when the signal is green. |
| 8  | Der Wagen war in der braunen Garage. | /der varan var in der braunen garaga/ | The car was in the brown garage. |
| 9  | Alle mögen Nele gerne.           | /ala mogun nulo gerno/       | Everyone likes Nele.           |
| 10 | Sie haben nur drei Beweise.      | /zie habon nur dira huvazz/   | You have only three pieces of evidence. |
Web API, IBM Watson speech-to-text and Wit.ai) were accessed via the Python libraries SpeechRecognition and IBM-WATSON. The audio files sent to the services were preprocessed in the same way as described in [15]. The word error rates (WER) were calculated from the true text representations of the utterances and the ASR responses using the Python library Jiwer. Before the WER calculation, true and recognized sentences were normalized on the textual level as in [15]. If numerals were returned by the ASR systems as numbers, they were converted into the corresponding normalized text, i.e., “7” was converted into “sieben” (English: seven). For each setting of phonation type and radiation characteristic, the WER was hence determined on the basis of 10 sentences (49 words in total).

III. RESULTS AND DISCUSSION
A. ACOUSTIC SIMILARITY OF SYNTHESIZED AND NATURAL UTTERANCES

Figure 8 shows the LTAS for each of the 9 synthesis conditions (3 phonation types × 3 radiation characteristics) as the black, gray and dashed lines. The white lines surrounded by the gray areas indicate the average LTAS of the real speakers and the interval of ±1 standard deviation. For each phonation type, the differences between the 3 synthetic LTAS reflect exactly the spectral differences between the 3 radiation characteristics, i.e., the high frequencies have the lowest amplitude for the simple radiation model, the highest amplitude for the piston-in-sphere model, and an intermediate amplitude for the head-torso model. For the synthetic modal voice, the three LTAS curves are most similar to the LTAS of natural speech, and deviate more for the synthesis with breathy and pressed voice. This is plausible because the speakers were instructed to speak with their “normal” voice, which should correspond to a modal voice when we average across multiple speakers. With regard to the spectra of the synthetic breathy voice it is noteworthy that the magnitude above 4 kHz is relatively high despite the steep spectral slope at lower frequencies. This boost at high frequencies can be attributed to the aspiration noise, which becomes the dominant excitation (compared to the periodic voice source) in breathy phonation at higher frequencies [17].

To quantify the differences between the synthetic and natural utterances, the root-mean-square deviations (RMSD in dB) between the 9 synthetic LTAS and the mean LTAS of the human utterances were calculated across the frequencies from 0–12 kHz. The results in Figure 9 show that for all three phonation types, the LTAS of the synthesis with the simple radiation model has the smallest RMSD and is hence most similar to the natural LTAS. The RMSD is highest for the piston-in-sphere radiation model, and inbetween for the head-torso model. This is against the expectation that a more realistic radiation model would lead to a higher similarity of the synthetic speech to natural speech. However, it must be kept in mind that the LTAS is determined by both the voice source model and the radiation characteristic, according to the acoustic theory of speech production [17]. According to this theory, for voiced sounds, the short-time spectrum of the radiated speech is the product of the voice source spectrum, the vocal tract filter, and the radiation characteristic. Hence, when we assume that the LTAS “averages out” the effect of the vocal tract filter, it is governed by the product of the source spectrum and the radiation characteristic. Apparently, the voice source settings used for the synthesis compensated for the lower realism of the simple radiation model, i.e., the too small spectral slope of the simple radiation characteristic was compensated for by a voice source with too strong high-frequency components. Finally, as observed above, the synthesis with modal voice deviated least from natural speech across all radiation models.

B. PERCEPTION EXPERIMENT

The results of the perception experiment are shown in Figure 10, where the total height of the bars indicates how often the listeners preferred a certain synthesis condition over the others in the pair-wise comparison. For each phonation type, the stimuli synthesized with the simple radiation characteristic were most preferred, the stimuli synthesized with the piston-in-sphere model were least preferred, and the head-torso radiation model was intermediate. Based on Wilcoxon signed-rank tests, the responses were significantly different (p < 0.05) for all three pairs of radiation models within each phonation-type group. For the stimuli with breathy phonation, we obtained p = 0.0022, p = 0.0447, and p = 0.0031 for the model pairs simple vs. piston-in-sphere, simple vs. head-torso, and piston-in-sphere vs. head-torso, respectively. For the same pairs, we obtained p = 0.0001, p = 0.0101, and p = 0.0022 for modal phonation, and p < 0.0001, p = 0.0001, and p = 0.0042 for pressed phonation. Since the listeners did not directly compare stimuli synthesized with different phonation types, the height of the bars can only be compared within the same phonation-type group. For each bar in Fig. 10, the lower part indicates the consistent responses of the listeners. They show the same pattern as the total responses.

The most striking observation here is that a decreasing preference for a radiation model coincides with an increasing acoustic deviation of the corresponding synthetic stimuli from natural speech based on the LTAS (Figure 9). This suggests that the long-term spectral deviation of synthetic from natural speech is a good indicator for the perceived naturalness of synthetic speech.

C. AUTOMATIC SPEECH RECOGNITION

Figure 11 shows the word error rates (WER) of the automatic speech recognizers for the sentences synthesized under the 9 different conditions, which can be interpreted as a proxy for the intelligibility of the synthetic speech. For each condition, the box plot represents the WER distribution of 30 data points (the 10 synthetic sentences recognized by each of 3 ASR systems). The box plots suggest that the radiation characteristic has only a minor effect on the WER. In fact, when all obtained
FIGURE 8. Long-term average spectra of the synthesized speech with a) breathy voice, b) modal voice, and c) pressed voice using the simple radiation characteristic (dashed line), the piston-in-sphere characteristic (black line), and the head-torso characteristic (gray line). The gray area indicates the ±1σ confidence interval of the LTAS of human speech estimated from 8 male speakers. The average LTAS of the human speakers is shown as the white line.

FIGURE 9. Root-mean-square deviations of the LTAS of the nine variants of synthesized sentences (3 phonation types × 3 radiation models) from the average LTAS of the human recordings of the same sentences.

FIGURE 10. Results of the perception experiment. The total height of each bar shows how often a particular variant (9 variants for 3 phonation types × 3 radiation models) of the synthesized sentences was preferred over the others in the pairwise comparison task (higher is better). The lower parts of the bars indicate the number of consistent responses of the participants.

FIGURE 11. Results of the automatic speech recognition experiment. Each box plot represents the word error rates obtained by 3 different ASR systems for the 10 sentences synthesized with a specific setting for the voice quality (breathy versus modal versus pressed voice) and radiation characteristic (simple versus piston-in-sphere versus head-torso). The boxplot at the bottom represents the word error rates obtained by the 3 ASR systems for the 10 natural utterances that served as templates for the re-synthesis. The crosses indicate median values, and the filled triangles indicated mean values.

IV. GENERAL DISCUSSION

In contrast to our expectations, the synthetic speech with the simple radiation characteristic was more similar to the natural utterances in terms of the LTAS than the utterances synthesized with the two more realistic radiation models. This result can be explained by the fact that the LTAS is generally determined by both the radiation characteristic and the spectrum of the voice source. Since the vocal fold model settings used here were adjusted using the simple radiation model prior to this study [35], the voice source model was obviously adjusted to compensate for the inaccurate (simple) radiation characteristic to sound as human as possible. This means that a more realistic radiation model alone does not automatically lead to greater spectral similarity with natural speech, but that the combination of the voice source model and the radiation characteristic is crucial.

A greater long-term spectral similarity between a synthetic utterance and the natural reference was also associated with a higher perceived naturalness of the utterance. This finding

WER values are grouped with respect to the radiation model, there are no significant statistical differences between the models according to two-sided Mann-Whitney-U tests (α = 0.05). When the obtained WER values are grouped with respect to the phonation type (pooled across the radiation models) and compared in 3 pairs with two-sided Mann-Whitney-U tests, the only significant difference is between the modal and pressed phonation types (p = 0.0190). When Bonferroni correction is applied (due to the three individual pairs), also this difference becomes insignificant with regard to α = 0.05.
suggests that the long-term spectral deviation of synthetic speech from natural speech (for the same utterances of multiple speakers) might serve as a feature for the automatic assessment of the naturalness of synthetic speech [42].

Given the above observations, in future work, the voice source settings of an articulatory synthesizer should be optimized while using a realistic radiation characteristic. This optimization could be done either on a perceptual level similar to [35], or on an acoustic level by matching the LTAS of the synthesized sentences with natural sentences.

A further finding of this study was that the magnitude spectra of the piston-in-square model and the head-torso model differ quite significantly. While the piston-in-square characteristic has a monotonically raising magnitude with increasing frequency (very similar to the simple radiation model), the most-realistic head-torso model shows a complex non-monotonic spectral pattern with multiple peaks and troughs in the magnitude spectrum. This suggests that the piston-in-square model for the radiation characteristic is still a strong simplification of the real situation, because it neglects the torso and neck, which create the complex diffraction pattern.

In conclusion, the radiation characteristics presented in this paper are a step toward a better understanding of speech production and toward improved speech production models. However, if the main goal of the synthesis is not the highest possible realism but simply natural-sounding speech, the new radiation characteristics may not be strictly essential.

Finally, we want to emphasize a few limitations of the present study. With regard to the head-torso model, the variation of the lip shape with different speech sounds was not taken into account, and the particle velocity in the plane of the lips was assumed uniform. With the onset of acoustic cross-modes in the vocal tract above 5 kHz, the latter is not strictly the case anymore and may affect the radiation characteristic at higher frequencies. With regard to the evaluation of the naturalness, the influence of the room has not been simulated (e.g. in terms of the room impulse response). Since we are not used to listen to speech in an anechoic environment, the absence of room reflections probably played a role in the perception of naturalness. Furthermore, the same radiation characteristic was used for the radiation from the nostrils and from the mouth during the synthesis. In reality, there might be small differences between them.

ACKNOWLEDGMENT

The authors would like to thank Patrick Häsner for his help with the preparation of the digital head-torso model and also would like to thank Jørgen S. Dokken for help with dolfinx, and the Centre for Information Services, and High-Performance Computing TU Dresden for providing its facilities for high throughput calculations.

REFERENCES

[1] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” PLoS ONE, vol. 8, no. 4, Apr. 2013, Art. no. e60603.
[2] K. Iskarous, L. Goldstein, D. H. Whalen, M. Tiede, and P. Rubin, “CASy: The Haskins configurable articulatory synthesizer,” in Proc. Int. Congr. Phonetic Sci. (ICPHS), Barcelona, Spain, 2003, pp. 185–188.
[3] B. Elie and Y. Laprie, “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., vol. 82, pp. 85–96, Sep. 2016.
[4] S. Fels, J. E. Lloyd, K. Van Den Doel, F. Vogt, I. Stavness, and E. Vatikiotis-Bateson, “Developing physically-based, dynamic vocal tract models using ArtiSynth,” in Proc. 7th Int. Seminar Speech Prod. (ISSP), Utabuta, Brazil, 2006, pp. 419–426.
[5] B. H. Story, “Phrase-level speech simulation with an airway modulation model of speech production,” Comput. Speech Lang., vol. 27, no. 4, pp. 989–1010, Jun. 2013.
[6] A. J. S. Teixeira, R. Martinez, L. N. Silva, L. M. T. Jesus, J. C. Principe, and F. A. C. Vaz, “Simulation of human speech production applied to the study and synthesis of European Portuguese,” EURASIP J. Adv. Signal Process., vol. 2005, no. 9, pp. 1435–1448, Dec. 2005.
[7] S. Stone, A. Azgin, S. Manz, and P. Birkholz, “Prospects of articulatory text-to-speech synthesis (abstract only),” in Proc. 12th Int. Seminar Speech Prod. (ISSP), 2020, pp. 1–2.
[8] A. Teixeira, C. Oliveira, and P. Barbosa, “European Portuguese articulatory based text-to-speech: First results,” in Proc. Int. Conf. Comput. Process. Portuguese Lang., 2008, pp. 101–111.
[9] G. Bailly, “Learning to speak. Sensori-motor control of speech movements,” Speech Commun., vol. 22, nos. 2–3, pp. 251–267, Aug. 1997.
[10] I. S. Howard and P. Messum, “Modeling the development of pronunciation in infant speech acquisition,” Motor Control, vol. 15, no. 1, pp. 85–117, Jan. 2011.
[11] M. Murakami, B. Kroger, P. Birkholz, and J. Triesch, “Seeing [U] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing,” in Proc. Joint IEEE Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL-EpiRob), Aug. 2015, pp. 208–213.
[12] A. K. Philippsen, R. F. Reinhart, and B. Wrede, “Goal babbling of acoustic-articulatory models with adaptive exploration noise,” in Proc. Joint IEEE Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL-EpiRob), Sep. 2016, pp. 72–78.
[13] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, “The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study,” J. Acoust. Soc. Amer., vol. 137, no. 3, pp. 1503–1512, Mar. 2015.
[14] Y. Xue, M. Marxen, M. Akagi, and P. Birkholz, “Articulatory and acoustical analysis and synthesis of shouted vowels,” Comput. Speech Lang., vol. 66, Mar. 2021, Art. no. 101156.
[15] P. K. Keug, S. Stone, K. J. Udo, and P. Birkholz, “Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies,” in Proc. 11th ISCA Speech Synth. Workshop (SSW), Aug. 2021, pp. 102–107.
[16] S. Mauda, “A digital simulation method of the vocal-tract system,” Speech Commun., vol. 1, nos. 3–4, pp. 199–229, Dec. 1982.
[17] N. Kenneth Stevens, Acoustic Phonetics. Cambridge, MA, USA: MIT Press, 1998.
[18] P. Birkholz and S. Drechsel, “Effects of the piriform fossae, transversal acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis,” Speech Commun., vol. 132, pp. 96–105, Sep. 2021.
[19] M. Sonodi and J. Schroeter, “A hybrid time-frequency domain articulatory speech synthesizer,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-35, no. 7, pp. 955–967, Jul. 1987.
[20] K. V. D. Doel and U. M. Ascher, “Real-time numerical solution of Webster’s equation on a nonuniform grid,” IEEE Trans. Audio, Speech, Language Process., vol. 16, no. 6, pp. 1163–1172, Aug. 2008.
[21] L. Leo Beranek, Acoustics. New York, NY, USA: McGraw-Hill, 1954.
[22] L. James Flanagan, Speech Analysis, Synthesis and Perception. Berlin, Germany: Springer-Verlag, 1965.
[23] J. L. Flanagan, “Analysis and measurement of sound radiation from the mouth,” J. Acoust. Soc. Amer., vol. 32, no. 12, pp. 1613–1620, Dec. 1960.
[24] B. B. Monson, E. J. Hunter, and B. H. Story, “Horizontal directivity of low- and high-frequency energy in speech and singing,” J. Acoust. Soc. Amer., vol. 132, no. 1, pp. 433–441, Jul. 2012.
[25] C. Porschmann and J. M. Arend, “Investigating phoneme-dependencies of spherical voice directivity patterns,” J. Acoust. Soc. Amer., vol. 149, no. 6, pp. 4553–4564, Jun. 2021.
[26] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. New York, NY, USA: Academic, 1999.

[27] C. Geuzaine and J.-F. Remacle, “Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities,” *Int. J. Numer. Methods Eng.*, vol. 79, no. 11, pp. 1309–1331, Sep. 2009.

[28] P. Birkholz, S. Kürbis, M. Fleischer, S. Stone, P. Häusser, R. Blandin, and M. Fleischer, “Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties,” *Sci. Data*, vol. 7, no. 1, pp. 1–16, Dec. 2020.

[29] M. Fleischer, A. Mainka, S. Kürbis, and P. Birkholz, “How to precisely measure the volume velocity transfer function of physical vocal tract models by external excitation,” *PLoS ONE*, vol. 13, no. 3, pp. 1–16, Mar. 2018.

[30] L. F. Kallivokas, J. Bielak, and R. C. MacCamy, “A simple impedance-infinite element for the finite element solution of the three-dimensional wave equation in unbounded domains,” *Comput. Methods Appl. Mech. Eng.*, vol. 147, nos. 3–4, pp. 235–262, Aug. 1997.

[31] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi, “A pilot study on the influence of mouth configuration and torso on singing voice directivity,” *J. Acoust. Soc. Amer.*, vol. 148, no. 3, pp. 1169–1180, Sep. 2020.

[32] A. Marshall and J. Meyer, “The directivity and auditory impressions of singers,” *Acta Acustica United With Acustica*, vol. 58, no. 3, pp. 130–140, 1985.

[33] T. Leino, “Long-term average spectrum in screening of voice quality in speech: Untrained male university students,” *J. Voice*, vol. 23, no. 6, pp. 671–676, Nov. 2009.

[34] B. B. Monson, A. J. Lotto, and B. H. Story, “Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives,” *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1754–1764, Sep. 2012.

[35] P. Birkholz, S. Drechsel, and S. Stone, “Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis,” in *Proc. Interspeech*, Sep. 2019, pp. 3765–3769.

[36] P. Birkholz, *3D-Artikulatorische Sprachsynthese*. Berlin, Germany: Logos Verlag, 2005.

[37] P. Birkholz, “Enhanced area functions for noise source modeling in the vocal tract,” in *Proc. 10th Int. Seminar Speech Prod. (ISSP)*, Cologne, Germany, 2014, pp. 37–40.

[38] P. Birkholz, “Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets,” in *Proc. Interspeech*, Aug. 2007, pp. 2865–2868.

[39] P. Birkholz, P. Schmazer, and Y. Xu, “Estimation of pitch targets from speech signals by joint regularized optimization,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2075–2079.

[40] P. Alku and E. Vilkman, “A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers,” *Folia Phoniatr et Logopaedica*, vol. 48, no. 5, pp. 240–254, 1996.

[41] E. Vickers, “The loudness war: Background, speculation, and recommendations,” in *Audio Engineering Society Convention*, vol. 129, San Francisco, CA, USA, 2010.

[42] G. Mittag and S. Möller, “Deep learning based assessment of synthetic speech naturalness,” in *Proc. Interspeech*. Shanghai, China, 2020, pp. 1748–1752.

**PETER BIRKHOLZ** (Member, IEEE) was born in Rostock, Germany, in 1978. He received the Diploma degree in computer science and the Ph.D. degree (Hons.) in signal processing from the University of Rostock, Germany, in 2002 and 2005, respectively. He worked as a Research Associate at the University of Rostock, from 2005 to 2009, and the Department of Phoniatrics, Pedaudiology, and Communication Disorders, RWTH Aachen University, Germany, from 2009 to 2014. He became a Junior Professor in cognitive systems at TU Dresden, Germany, in 2014, where he is currently a Full Professor in speech technology and cognitive systems, in 2020. His research interests include speech production, articulatory speech synthesis, computational neuroscience, silent speech interfaces, and measurement techniques for speech research. He was awarded the Joachim-Jungius Prize, in 2006, by the University of Rostock for his dissertation on articulatory speech synthesis, and the Klaus-Tschira Award for Achievements in Public Understanding of Science, in 2006.

**STEFFEN OSSMANN** was born in Riesa, Germany, in 1989. He received the Diploma degree in mechanical engineering from Technische Universität Dresden, in 2014. Since 2014, he has been working as a Research Assistant with the Department of Otorhinolaryngology, Head and Neck Surgery, TU Dresden. His research interests include biomedical image analysis and modeling of mechano-acoustic systems.

**RÉMI BLANDIN** was born in Pau, France, in 1989. He received the master’s degree in acoustics from the University Pierre and Marie Curie of Paris, France, in 2013, and the Ph.D. degree in acoustics from the University of Grenoble, France, in 2016. Since 2019, he has been working as a Postdoctoral Researcher at TU Dresden. His research interests include waveguide acoustics, semi-analytical simulation methods, speech production, articulatory synthesis, and speech directivity.

**ALEXANDER WILBRANDT** was born in Kiel, Germany, in 1992. He received the master’s degree in engineering economics from Christian-Albrechts-Universität Kiel, in 2019. He is currently working as a Postdoctoral Researcher at Technische Universität Dresden, as the Chair of Speech Technology and Cognitive Systems under Prof. Birkholz and specializes in biofeedback and intra oral measurement devices. His research interests include articulatory speech synthesis and automatic speech recognition.

**PAUL KONSTANTIN KRUG** received the bachelor’s and master’s degrees in physics from Georg-August-Universität Göttingen, Germany, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with Technische Universität Dresden. He is also a Research Associate of the Chair of Speech Technology and Cognitive Systems, Technische Universität Dresden. His research interest includes the field of acoustic-to-articulatory-inversion.