IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [ɑ]

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías, Oriol Guasch

GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain

{marc.freixes,marc.arnela,joanclaudi.socoro,francesc.alias,oriol.guasch}@salle.url.edu

Abstract

One-dimensional articulatory speech models have long been used to generate synthetic voice. These models assume plane wave propagation within the vocal tract, which holds for frequencies up to ∼5 kHz. However, higher order modes also propagate beyond this limit, which may be relevant to produce a more natural voice. Such modes could be especially important for phonation types with significant high frequency energy (HFE) content. In this work, we study the influence of tense, modal and lax phonation on the synthesis of vowel [ɑ] through 3D finite element modelling (FEM). The three phonation types are reproduced with an LF (Liljencrants-Fant) model controlled by the R_d glottal shape parameter. The onset of the higher order modes essentially depends on the vocal tract geometry. Two of them are considered, a realistic vocal tract obtained from MRI and a simplified straight duct with varying circular cross-sections. Long-term average spectra are computed from the FEM synthesised [ɑ] vowels, extracting the overall sound pressure level and the HFE level in the 8 kHz octave band. Results indicate that higher order modes may be perceptually relevant for the tense and modal voice qualities, but not for the lax phonation.

Index Terms: voice production, higher order modes, high frequency energy, glottal source modelling, LF model, numerical simulation, finite element method

1. Introduction

For many years, works on articulatory speech synthesis have considered a simplified one-dimensional (1D) representation of the vocal tract. This is built from the so-called vocal tract area functions, which describe the cross-sectional area variations along the vocal tract center midline (see e.g., [1]). Voice is then synthesised by simulating the propagation of acoustic waves within this 1D representation of the vocal tract (see e.g., [2, 3, 4]). However, 1D approaches assume plane wave propagation, so they can only correctly approximate the acoustics of the vocal tract in the frequency range below 4-5 kHz. Beyond this limit, not only planar modes get excited but also higher order propagation modes appear, which strongly change the high frequency energy (HFE) content of the spectrum [5, 6] compared to that from a 1D model. Although the high frequency range has not received much attention in the literature, some recent studies point out that the HFE may be relevant for voice quality, speech localisation, speaker recognition and intelligibility (see [7] and references therein).

On the other hand, three-dimensional (3D) models do not need to assume plane wave propagation, since they can directly deal with 3D vocal tract geometries to emulate the complex acoustic field generated during voice production [8, 9, 10]. However, higher order modes do not always appear even if a 3D acoustic model is used. As shown in [6], a straightened vocal tract based on circular cross-sections prevents the onset of such modes due to radial symmetry, in contrast to what occurs for realistic vocal tract geometries based on MRI data. Other vocal tract geometries simplifications were studied in that work, all of them showing large variations in the HFE while keeping a similar behaviour for low frequencies. One can then assert that the vocal tract shape is determinant for the HFE content of the generated sound. However, the vocal tract shape is not the only factor affecting the HFE. The type of phonation can also modify it, as shown for instance in [11] for sustained vowels with loud and soft phonation.

In this work we study the effect of tense, modal and lax phonation on the synthesis of vowel [ɑ], paying special attention to the HFE content. These three phonation types are reproduced using an LF (Liljencrants-Fant) model [12]. Although this model cannot consider the interaction between the vocal tract and the vocal folds [13, 14], it has proved to be useful to explore the phonatory tense-lax continuum [15] by controlling the R_d glottal shape parameter [16]. Regarding the vocal tract, we consider an MRI-based realistic geometry, and its simplified counterpart considering circular cross-sections in a straightened midline [6]. This allows us to somewhat "activate" and "deactivate" the higher order modes. Different versions of vowel [ɑ] are generated by convolving the LF glottal source signals with the vocal tract impulses responses obtained using a 3D acoustic model based on the Finite Element Method (FEM) [17]. In order to analyse the relevance of the higher order propagation modes for the lax, modal and tense phonation, the long-term average spectra (LTAS) and the HFE levels of the synthesised vowels are computed and compared.

The paper is structured as follows. The methodology used to study the production of vowel [ɑ] with the three phonation types and the two vocal tract geometries is explained in Section 2. Next, the obtained results are discussed in Section 3. Finally, conclusions and future work are presented in Section 4.

2. Methodology

Figure 1 depicts the process followed to synthesise six versions of vowel [ɑ]. These were obtained by convolving three glottal source signals with the FEM impulse responses of two vocal tract geometries that produce this vowel. In particular, and as mentioned before, we used the realistic vocal tract and the simplified straightened simplification with circular cross-sections from [6] (see Section 2.1), and computed their impulse responses $h(t)$ using the FEM (see Section 2.2). The glottal source signals $u_g(t)$ were generated by means of an R_d controlled LF model. The values $R_d = 0.3, 1$ and 2.7 were selected from the R_d range $[0.3, 2.7]$ (see [16]) to reproduce a tense, a modal, and a lax phonation, respectively (see Section 2.3).

10.21437/IberSPEECH.2018-28

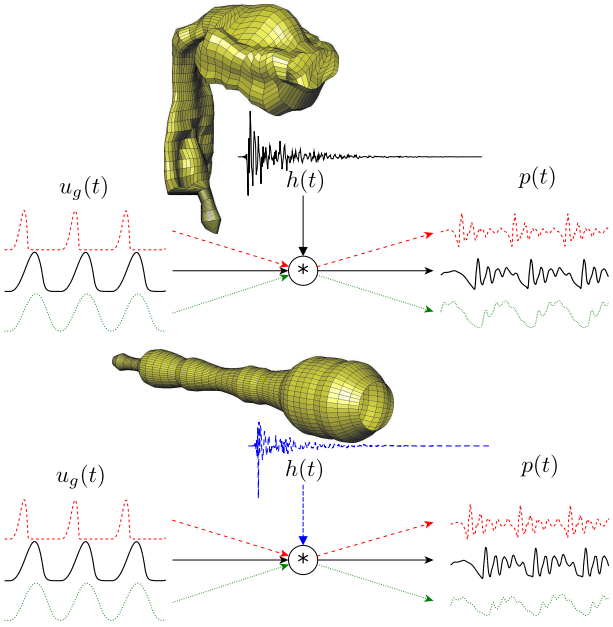


Figure 1: *Synthesis of vowel [a] with a realistic vocal tract geometry (above) and its simplified counterpart of circular cross-sections in a straightened midline (below). Three phonation types are considered to reproduce a tense (dashed red line), a modal (solid black line) and a lax (dotted green line) voice production. The output pressure signal $p(t)$ is computed as the convolution of the glottal source $u_g(t)$ with the vocal tract impulse response $h(t)$ obtained from 3D FEM simulations.*

For each vowel, the LTAS was computed as the Welch's power spectral density estimate, with a 15 ms hamming window, 50% overlap and a 2048-point FFT. The overall energy levels and the HFE levels in the 8 kHz octave band were also extracted as in [11]. The 16 kHz octave band was not considered, since HFE changes in this frequency range were found almost perceptually irrelevant in [11].

2.1. Vocal Tract Geometries

Two vocal tract geometry simplifications of vowel [ɑ] have been employed in this work, namely, the realistic configuration and the simplified straight vocal tract with circular shape (see Fig. 1). These geometries were obtained in [6] by simplifying the MRI-based vocal tract geometries in [18]. In a nutshell, the procedure consisted in the following. First, the subglottal tube, the face and the lips were removed from the original geometry (see [10] for a detailed analysis of the lips influence on simulations). Moreover, side branches such as the piriform fossae and valleculae were occluded (see e.g. [9, 19] for their acoustic effects). Cross-sections were next extracted as typically done to generate 1D area functions, but preserving their shapes and locations in the vocal tract midline. The realistic configuration was generated by linearly interpolating the resulting cross-sections. As shown in [6], this simplification provides very similar results to the original MRI-based vocal tract geometry without branches.

In the simplified straight vocal tract configuration, the cross-sectional shapes were modified to be that of a circle, preserving the same area. These circular cross-sections were located in a straightened version of the vocal tract midline and

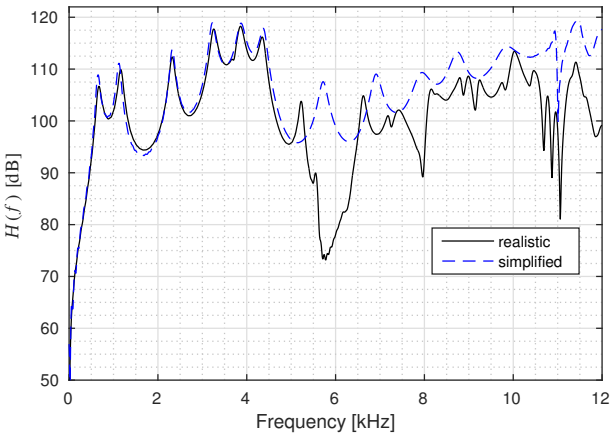


Figure 2: *Vocal tract transfer function $H(f)$ for the vowel [a] with the realistic and simplified vocal tract geometries.*

then linearly interpolated. The two configurations are hereafter referred as the realistic and the simplified vocal tracts.

2.2. Vocal Tract impulse response

The impulse response of each vocal tract geometry was computed using a custom finite element code that numerically solves the acoustic wave equation,

$$\partial_{tt}^2 p - c_0^2 \nabla^2 p = 0, \tag{1}$$

combined with a Perfectly Matched Layer (PML) to account for free-field propagation [17]. In Eq. (1) $p(\boldsymbol{x}, t)$ is the acoustic pressure, ∂_{tt}^2 stands for the second order time derivative, and c_0 is the speed of sound which is set to the usual value of 350 m/s. A Gaussian pulse was introduced on the glottal cross-sectional area as an input volume velocity $u_g(t)$. This pulse is of the type

$$u_g(t) = e^{-\left[(t-T_{gp})/0.29T_{gp}\right]^2} [\mathrm{m}^3/\mathrm{s}], \tag{2}$$

with $T_{gp} = 0.646/f_c$ and $f_c = 10$ kHz. Wall losses were considered by imposing a boundary admittance coefficient of $\mu = 0.005$ on the vocal tract walls. A 20 ms simulation was then performed capturing the acoustic pressure $p_0(t)$ at a node located outside of the vocal tract, 4 cm away from the mouth aperture center. The sampling frequency was set to $f_s = 8000$ kHz, which ensures a restrictive stability condition of the Courant-Friedrich-Levy type required by explicit numerical schemes (see [17] for details on the numerical scheme).

A vocal tract transfer function $H(f)$ was computed from each simulation to compensate for the slight energy decay in frequency of the Gaussian pulse. This is defined as

$$H(f) = \frac{P_o(f)}{U_g(f)}, \tag{3}$$

with $P_o(f)$ and $U_g(f)$ being the Fourier Transform of $p_o(t)$ and $u_g(t)$, respectively. $H(f)$ was computed up to 12 kHz, to allow the calculation of HFE level in the 8 kHz octave band [11]. The vocal tract transfer functions $H(f)$ for the realistic and the simplified geometries of vowel [ɑ] are shown in Fig. 2 (also reported in [6], but only up to 10 kHz). As can be observed, planar modes are mainly produced below 5 kHz giving place to the first vowel formants. Beyond this value, higher order modes can also propagate, resulting in the more complex spectrum of

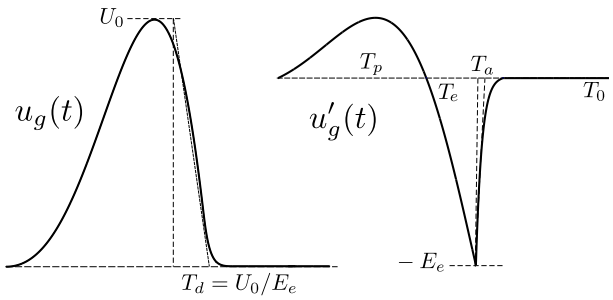


Figure 3: *Glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ according to the LF model [12].*

the realistic geometry. Note, however, that these modes do not appear in the spectrum of the simplified configuration. The radial symmetry of this geometry prevents their onset [5, 6].

Finally, the inverse Discrete Fourier Transform was applied to the vocal tract transfer functions $H(f)$ to obtain the vocal tract impulse responses $h(t)$ of the two geometries (see Fig. 1).

2.3. Voice Source Signal

An LF model [12] was used to produce the voice source signal. This model approximates the glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ in terms of four parameters (T_p, T_e, T_a, E_e) that describe its time-domain properties (see Fig. 3). The control of this model can be simplified with the single glottal shape parameter R_d [16]. This is defined as

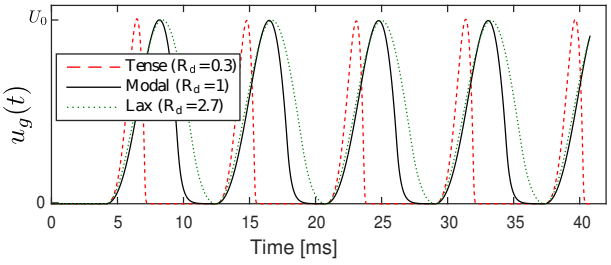
$$R_d = \frac{T_d}{T_0}\frac{1}{110} = \frac{U_0}{E_e}\frac{F_0}{110}, \tag{4}$$

where T_d is the declination time, T_0 the period, and F_0 the fundamental frequency. The declination time T_d corresponds to the quotient between the glottal flow peak U_0 and the negative amplitude of the differentiated glottal flow E_e.

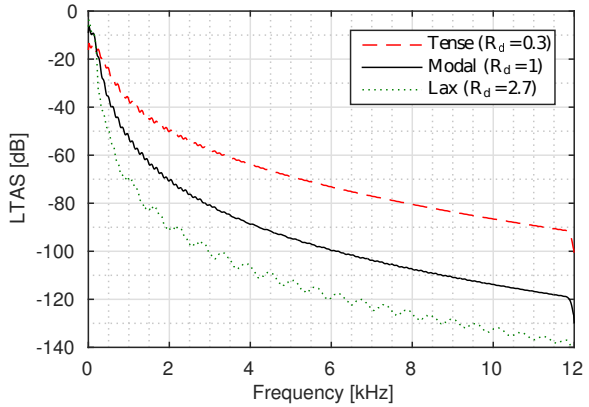
In this work, we used the Kawahara's implementation of the LF model [20], which generates a free-aliasing excitation source signal. We adapted this model to our purposes, modifying the sampling frequency from its original value of 44100 Hz to 24 kHz. Moreover, we introduced the R_d glottal shape parameter. This allows one to easily control the voice source with a single parameter, which runs from $R_d = 0.3$ for a very adducted phonation, to $R_d = 2.7$ for a very abducted phonation (see [16]). From the R_d range $[0.3, 2.7]$ two extreme values plus a middle one were chosen. We used $R_d = 0.3$ to generate a tense phonation, $R_d = 2.7$ for a lax production, and $R_d = 1$ for a normal (modal) voice quality. With regard to F0, a pitch curve was obtained from a real sustained vowel lasting 4.4 seconds. This pitch contour was placed around 120 Hz to generate all the source signals. Figure 4a shows four periods of the three simulated voice source waveforms. Moreover, the LTAS of the glottal source signals are represented in Fig. 4b. As observed, the phonation type obviously changes the glottal pulse shape, thus modifying the spectral energy distribution of the source signal.

3. Results

Six versions of vowel [ɑ] (see Fig. 1) have been generated using the three glottal source signals corresponding to a tense, a modal and a lax phonation, and the two impulse responses obtained



(a) *Waveform*



(b) *Long-term average spectrum*

Figure 4: *Glottal source for a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation.*

from the 3D FEM simulations of the realistic and simplified vocal tract geometries. The six synthesised vowels are normalised with the same scaling factor to obtain reasonable sound pressure levels. This factor has been selected so as to produce 70 $\mathrm{dB_{SPL}}$ in the realistic geometry with a modal phonation ($R_d = 1$). The LTAS have then been computed for each audio.

Figure 5 shows the obtained LTAS for the six generated vowels. As also appreciated in the vocal tract transfer functions (see Fig. 2), small differences between geometries are produced for frequencies below 5 kHz, whereas beyond this range higher order modes propagate in the realistic case, thus inducing larger deviations. This behaviour can be observed for the three phonation types. Essentially the glottal source modifies the overall energy level and also introduces an energy decay in frequency (compare Fig. 2 with Fig. 5). This decay, known as the spectral tilt, strongly depends on the phonation type. The laxer the phonation the larger the spectral tilt [16]. Furthermore, the voice source also affects the energy balance of the first harmonics (below $\sim$ 500 Hz). For instance, the lax phonation has the lowest overall energy values among all phonation types. However, one can see that the first harmonic (close to 120 Hz) has larger amplitude levels than the rest of the spectrum, in contrast to what occurs for the other phonations.

HFE levels have been computed by integrating the power spectral density in the 8 kHz octave band, as in [11]. In addition, the overall energy levels have been calculated following the same procedure but for the whole examined frequency range.

The obtained results are listed in Table 1. Note first that in the realistic case with a modal phonation ($R_d = 1$) the overall level is 70 $\mathrm{dB_{SPL}}$. Remember that this value was fixed to

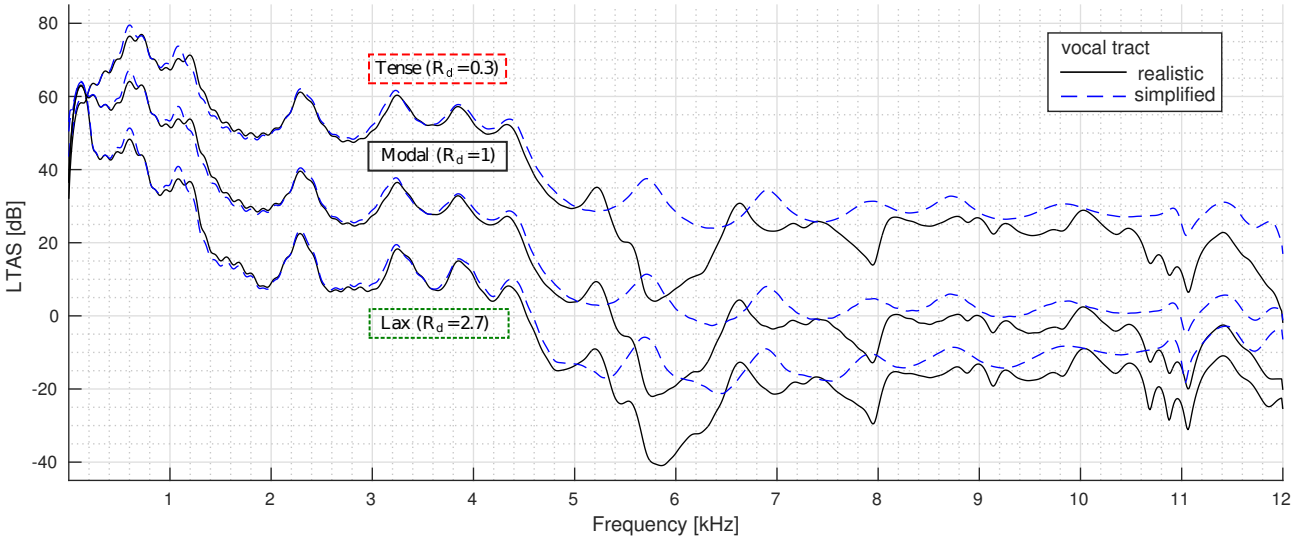


Figure 5: *Long-term average spectra (LTAS) of the FEM synthesised vowel [ɑ] using the realistic and simplified vocal tract geometries with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation.*

Table 1: *Overall and High-Frequency Energy (HFE) levels (in dB) obtained in the realistic and simplified vocal tract configurations of vowel [ɑ] with a tense ($R_d = 0.3$), a modal ($R_d = 1$), and a lax ($R_d = 2.7$) phonation. Δ denotes the difference between the two vocal tract geometries.*

R_d	Geometry	Overall	ΔOverall	HFE	ΔHFE
0.3	realistic	82.2	1.2	41.4	5.8
	simplified	83.4		47.3	
1	realistic	70.0	1.3	14.9	5.9
	simplified	71.3		20.8	
2.7	realistic	63.5	1.4	1.0	5.6
	simplified	64.9		6.6	

compute the scaling factor used to normalise the audio files. The overall level variations for the other configurations will thus correspond to modifications either introduced by the vocal tract geometry or by the glottal source. As expected, the larger the R_d value (laxer phonation) the smaller the overall levels.

Far more interesting is to compare the results between geometries. The HFE levels decay between 5.6 dB and 5.9 dB for the realistic vocal tract depending on the phonation type, which only manifests as an overall level difference of 1.2 dB and 1.4 dB. The higher order modes tend to reduce the levels in the HFE content. According to [11], minimum difference limen scores of about 1 dB are given for normal-hearing listeners in the 8 kHz octave band, so one may hypothesise that the higher order modes may be perceptually relevant. However, depending on the phonation type the HFE could be too small to notice any difference. This seems to be the case of the lax phonation ($R_d = 2.7$), which gives HFE levels of 1.0 dB and 6.6 dB, depending on the geometry. We may then conjecture, that for this phonation type no differences in the outputs from the two geometries will be perceived. In other words, we would not notice the influence of higher order modes.

4. Conclusions

In this work we have studied the influence of tense, modal and lax phonation on the 3D finite element synthesis of vowel [ɑ], considering a realistic and a simplified vocal tract geometry. The 3D simulations behave very similarly for both geometries below 5 kHz, but significant differences appear beyond this frequency because of the rising of higher order propagation modes. It is worth mentioning that these modes only appear when using the realistic vocal tract. They induce a reduction of the HFE levels at the 8 kHz octave band from 5.6 to 5.9 dB, depending on the phonation type. These differences may be perceptually relevant, according to previous works in the literature. Specifically, a realistic 3D vocal tract geometry would be required for an accurate synthesis of vowel [ɑ] through 3D FEM, when trying to simulate a modal and a tense voice production. Conversely, when a lax phonation is considered, the influence of higher order propagation may be imperceptible, since the HFE levels are very small. Therefore, a simpler 1D simulation would suffice in this case.

Future work will consider other R_d values and geometry simplifications as well as other vowels to complete the study. Finally, we also plan to include aspiration noise in the LF model to evaluate its impact on the HFE content of the numerical simulations.

5. Acknowledgements

The authors are grateful to Saeed Dabbaghchian for the design of the vocal tract geometry simplifications. This research has been supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The fourth author acknowledges the support from the Obra Social “La Caixa" under grant ref. 2018-URL-IR1rQ-021.

6. References

[1] B. H. Story, I. R. Titze, and E. A. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[2] 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, 2013.

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

[4] S. Stone, M. Marxen, and P. Birkholz, "Construction and evaluation of a parametric one-dimensional vocal tract model," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 8, pp. 1381–1392, 2018.

[5] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. V. Hirtum, and X. Laval, "Effects of higher order propagation modes in vocal tract like geometries," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 832–8, 2015.

[6] M. Arnela, S. Dabbaghchian, R. Blandin, O. Guasch, O. Engwall, A. Van Hirtum, and X. Pelorson, "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1707–1718, 2016.

[7] B. B. Monson, A. J. Lotto, and B. H. Story, "Gender and vocal production mode discrimination using the high frequencies for speech and singing," *Frontiers in Psychology*, vol. 5, p. 1239, 2014.

[8] T. Vampola, J. Horáček, and J. G. Švec, "FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels," *Acta Acust. united with Acustica*, vol. 94, no. 5, pp. 433–447, 2008.

[9] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3724–3738, 2010.

[10] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2852–2859, 2016.

[11] B. B. Monson, A. J. Lotto, and S. Ternström, "Detection of high-frequency energy changes in sustained vowels produced by singers," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2263–2268, 2011.

[12] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 26, no. 4, pp. 1–13, 1985.

[13] T. Murtola, P. Alku, J. Malinen, and A. Geneid, "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy," *Speech Communication*, vol. 96, pp. 67–80, 2018.

[14] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, "A review of lumped-element models of voiced speech," *Speech Communication*, pp. 667–690, 2013.

[15] A. Murphy, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "Rd as a Control Parameter to Explore Affective Correlates of the Tense-Lax Continuum," in *INTERSPEECH*, 2017, pp. 3916–3920.

[16] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 36, no. 2-3, pp. 119–156, 1995.

[17] M. Arnela and O. Guasch, "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 4197–4209, 2013.

[18] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, and M. Vainio, "Large scale data acquisition of simultaneous MRI and speech," *Applied Acoustics*, vol. 83, pp. 64–75, 2014.

[19] H. Takemoto, S. Adachi, P. Mokhtari, and T. Kitamura, "Acoustic interaction between the right and left piriform fossae in generating spectral dips," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2955–2964, 2013.

[20] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis," in *INTERSPEECH*, 2017, pp. 1358–1362.