INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Comparison of the finite element method, the multimodal method and the transmission-line model for the computation of vocal tract transfer functions

Rémi Blandin[1], *Marc Arnela*[2], *Simon Félix*[3], *Jean-Baptiste Doc*[4], *Peter Birkholz*[1]

[1]Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01062, Germany
[2]GTM - Grup de Recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull C/Quatre Camins 30, 08022 Barcelona, Catalonia Spain
[3]Laboratoire d'Acoustique de l'Université du Mans (LAUM), UMR 6613, Institut d'Acoustique Graduate School (IA-GS), CNRS, Le Mans Université, Le Mans 72085, France
[4] Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers, Paris 75003, France

remi.blandin@tu-dresden.de, marc.arnela@salle.url.edu, simon.felix@univ-lemans.fr, jean-baptiste.doc@lecnam.net, peter.birkholz@tu-dresden.de

Abstract

The acoustic properties of vocal tract are usually characterized by its transfer function from the input acoustic volume flow at the glottis to the radiated acoustic pressure. These transfer functions can be computed with acoustic models. Three-dimensional acoustic simulation are used to take into account accurately the three-dimensional vocal tract shape and to generate valid results even at high frequency. Finite element models, finite difference methods, three-dimensional waveguide meshes, or the multimodal method have been used for this purpose. However, these methods require much more computation time than simple one-dimensional models. Among these methods, the multimodal method can achieve the shortest computation times. However, all the previous implementations had limitations regarding the geometrical shapes and the losses. In this work, we evaluate a new implementation that intends to overcome these limitations. Vowel transfer functions obtained with this new implementation are compared with a transmission-line model and a proven, robust and highly accurate method: the finite element method. While the finite element method remains the most reliable, the multimodal method generates similar transfer functions in much less time. The transmission line model gives valid results for the four first resonances.

Index Terms: vocal tract, speech acoustics, vowel transfer functions

1. Introduction

Acoustic simulations of the vocal tract transfer function are a key element to understand the relationship between the vocal tract geometry and its acoustic properties. This is traditionally done using one-dimensional (1D) approaches, such as transmission-line models (TLM), which rely on a simplified description of the vocal tract accounting only for the variation of cross-sectional area. However, 1D models need to assume the propagation of plane waves, valid up to about 5 kHz.

In the context of articulatory synthesis, using a three-dimensional (3D) model of the vocal tract allows for the accurate estimation of the cross-sectional area function [1, 2, 3, 4]. On the other hand, using 3D acoustic simulation methods such as finite elements (FEM) ensures an accurate computation of the transfer function, even above 5 kHz [5, 6]. Thus, combining articulatory synthesis with precise 3D acoustic simulations seems promising for providing high-quality speech synthesis.

As a matter of fact, vowels and diphthongs have already been synthesized using the Artisynth FRANK biomechanical model [4] and FEM simulations [7, 8]. However, this type of combination requires a lot of computation time for both the biomechanical and acoustic simulations. Synthesising words or sentences would take so much time that only a few samples could be generated. Yet, generating connected speech is very important to study speech production. As an example, evaluating intelligibility or studying the effect of physical constraints on phonetics requires to synthesize speech with a time-varying vocal tract. Thus the objective of this work is to evaluate alternative methods which can potentially provide a better compromise between accuracy and computation time.

Using a geometric articulatory model, such as the one implemented in VocalTractLab [9], can reduce the computation time in comparison to a biomechanical model. On the other hand, the multimodal method (MM) [5] can reduce the simulation time compared to other 3D acoustic simulation methods. However, previous applications of the MM to the vocal tract had limitations in term of geometrical accuracy and modeling of losses. In this work, we used an improved implementation which addresses both of these limitations. In particular, the curvature of the tube sections is modeled using a geometrical transformation, whereas losses are introduced at the vocal tract walls through a surface impedance.

The theoretical complexity of the MM requires a careful validation. This is done by comparing MM with FEM, which is a robust, proven and highly accurate 3D simulation method [10, 11]. On the other hand, the coupling of FEM and a geometrical articulatory model offers already more flexibility than using a biomechanical model, and constitutes by itself an interesting approach. Finally, it was evaluated to what extent the 1D acoustic simulations differ from the 3D simulations to determine their benefit and in which frequency range they bring a real advantage. For this purpose, a TLM implemented in VocalTractLab was compared with FEM and MM.

2. Method

2.1. Vocal tract geometries

The vocal tract geometries were extracted from the articulatory synthesizer VocalTractLab 2.3 (`www.vocaltractlab.de`).

http://dx.doi.org/10.21437/Interspeech.2021-975

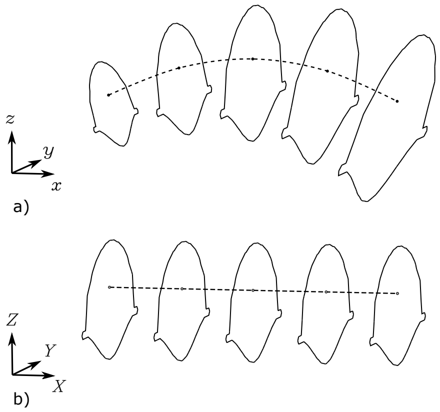

Figure 1: *Vocal tract segment in (a) Cartesian coordinates (x, y, z) and (b) in the transformed coordinates (X, Y, Z) after a geometric transformation straightening the segment and compensating cross-sectional area variations.*

Here we used the geometries of the four vowels, /a/, /i/, /u/ and /ə/, generated with two different configurations of the articulatory model corresponding respectively to a female [12] and a male speaker [9]. These configurations were obtained by tuning the anatomical and geometrical parameters of the articulatory model on magnetic resonance images obtained for multiple phonemes of the two subjects.

These geometries were segmented and the coordinates of the points of the contours of the obtained segments were exported in text files. Matlab and GID [1] were next used to generate the 3D meshes necessary for the FEM simulations. Only 65 segments of the 129 provided by VocalTractLab were used by the 3D FEM model to comply with the selected element size (about 3 mm in the inner domain). Note that this process interpolates linearly the successive contours without abrupt cross-sectional changes, unlike the geometry simulated with the MM. Therefore, in order to obtain the most accurate description of the geometries, the MM was applied to all the 129 segments generated by the segmentation process. For the TLM, only the area function was used. Also for this implementation the area function was down-sampled resulting in 80 tubes. The exported 3D tube shapes are available in the supplementary material and at `https://vocaltractlab.de/index.php?page=birkholz-supplements`.

2.2. Multimodal method

In this document the hat symbol ˆ represents vectors, and the bold capital letters represent matrices. In order to apply the multimodal method, the vocal tract geometries are sliced in segments with constant cross-sectional shape characterized by a reference point, a normal vector, a curvature radius and a scaling factor varying along the curvilinear abscissa X. An example segment is shown in Fig. 1a.

To take into account the curvature and the cross-sectional area variations of the vocal tract, a geometric transformation to a straight segment with a constant cross-sectional area is used, as illustrated in Fig. 1b. Such a transformation is defined following Maurel *et al.* [13].

The acoustic pressure p in a vocal tract segment can be expressed as the summation of the local transverse modes φ

$$p(X, Y, Z) = \sum_{n}^{\infty} p_n(X)\varphi_n(Y, Z), \qquad (1)$$

where $p_n(X)$ is the amplitude of the transverse mode $\varphi_n(Y, Z)$ along the longitudinal dimension of the segment, and (Y, Z) is the transverse plane.

The geometrical transformation is applied to the wave equation and the wall boundary condition

$$\begin{aligned} (\Delta + k^2)p &= 0, & (2)\\ \hat{\nabla} p \cdot \hat{n} &= -jk\mu p, & (3) \end{aligned}$$

where k is the wavenumber, $\hat{n}$ the outward pointing normal and $\mu = \frac{1}{\rho c Z_w} = 0.005$, ρ the volumetric mass, c the sound speed and Z_w the wall surface impedance. Using the Jacobian $\boldsymbol{J}$ of the geometric transformation, one obtain an expression of Eqs. (2) and (3) in the transformed coordinates

$$\begin{aligned} \mathrm{div}\boldsymbol{H}\hat{\nabla} p + \frac{k^2}{\det\boldsymbol{J}} p &= 0, & (4)\\ \boldsymbol{H}\hat{\nabla} p \cdot \hat{n} &= -jk\mu \frac{p}{\det\boldsymbol{J}}, & (5) \end{aligned}$$

where $\boldsymbol{H} = \frac{{}^t\boldsymbol{J}\boldsymbol{J}}{\det\boldsymbol{J}}$, and superscipt ^{t} meaning transpose.

A new physical field q related to the acoustic pressure is introduced so that equations (4) and (5) can be expressed as a system of first order differential equations

$$\frac{\partial}{\partial X}\begin{pmatrix} p \\ q \end{pmatrix} = \boldsymbol{\mathcal{M}}\begin{pmatrix} p \\ q \end{pmatrix}, \qquad (6)$$

where $\boldsymbol{\mathcal{M}}$ is a matrix whose terms depend on the expression of $\boldsymbol{H}$ and $\boldsymbol{J}$.

Eq. (6) is then expressed as a function of the transverse modes φ_n. It is solved by setting a radiation impedance boundary condition at the mouth end, enforcing continuity of the acoustic field at the segment interfaces, and implementing a source input volume velocity at the vocal folds end. The radiation impedance matrix is computed following the method presented in [14]. The field continuity is expressed using a mode matching matrix, as defined in [5].

Since the vocal tract cross-sectional shapes are quite different from simple shapes such as ellipses or rectangles, the transverse eigenproblem giving, in each cross-section, the modes φ_n and the corresponding propagation constants, is solved using 2D-FEM.

This method has been implemented directly in an extended version of the articulatory synthesizer VocalTractLab 2.3. A detailed presentation of this method is in the process of publication in a journal article.

2.3. Finite elements

The FEM approach in [15, 16] was used to numerically solve the mixed wave equation for the acoustic pressure $p(\hat{x}, t)$ and particle velocity $\hat{u}(\hat{x}, t)$. Losses at the vocal tract walls were introduced using the same Z_w as in the MM. However, radiation losses were simulated by allowing sound waves propagate from the mouth exit into a semi-spherical computational domain, as in [8]. A Sommerfeld boundary condition was applied at its outer boundary to absorb incoming waves.

The vocal tract airway was meshed using tetrahedra of size $h \sim 3$ mm, whereas h ranged between $[4, 5]$ mm in the outer

[1] https://www.gidhome.com/

computational domain. A 50 ms time-domain simulation was run introducing a Gaussian pulse $q_i(t)$ [16] at the vocal tract entrance (glottis) as excitation. A time step of 2e-6 s was used for the numerical scheme. The acoustic pressure $p_o(t)$ was captured outside the vocal tract at a distance of 3 cm. The vocal tract transfer function was finally computed as $H(f) = P_o(f)/Q_i(f)$, with $P_o(f)$ and $Q_i(f)$ being the Fourier transforms of $p_o(t)$ and $q_i(t)$, respectively.

2.4. Transmission-line model

The articulatory synthesizer VocalTractLab provides two implementations of the TLM, in the time and frequency domains. The frequency domain implementation corresponding more closely to MM and FEM was used here. The radiation condition was set to a baffled piston, and all the side branches, the static pressure drop, the lumped element approximation and the inner length correction were deactivated. Boundary layer resistance, heat condition losses and Hagen-Poiseuille resistance were taken into account, but the soft wall option was desactivated because it was not taken into account in the FEM and MM simulations. Note that in the TLM implementation the boundary layer losses are frequency dependent, in contrast to the MM and FEM, which consider frequency-independent losses. Because of this differences, the bandwidths of the resonances are not comparable between the 1D simulation method and the 3D simulations and not further discussed. The theoretical basis of the implementation associated to this configuration is described in [17, 1].

3. Result and discussion

The transfer functions obtained with FEM, MM and TLM are presented in Fig. 2. They exhibit resonances in the frequency ranges expected for the simulated vowels. The male and female configurations have different resonance frequencies, which tend to be higher for the female configuration as the vocal tract is shorter for women.

The three simulation methods give reasonably similar resonance frequencies for the first four resonances: the average relative difference to the FEM is about 2% for TLM (maximum 6%) and 1.2% for the MM (maximum 5.1%). The first four resonance frequencies of the TLM are generally higher than those of the FEM (with 3 exceptions for /u/ male and female and /i/ male that have 2 lower frequencies). This could be due to 3D acoustic effects not accounted for at the interfaces between the segments. A length correction model proposed to compensate for this phenomenon [18] has been tested (the option ”Inner length correction” in VocalTractLab) but induced significantly more severe deviations from the FEM reference in the opposite direction. Thus, a more accurate model of this phenomenon would be necessary to further improve the accuracy of the TLM.

The FEM and the MM have globally very similar transfer functions, generating in most of the cases the same resonances and anti-resonances. The largest global differences are observed for the male /i/ and the /u/ for both genders (Figs. 2f, 2g and 2h).

The differences between FEM and MM in the range 2–4 kHz are a bit surprising in their extent and difficult to explain. See in particular for the vowel /u/ (Figs. 2g and 2h) the two resonances which are much closer to each other for MM than for FEM, yielding a relative difference of about 5%. Understanding the origin of this difference is of great interest since this frequency range includes the maximum of hearing sensitivity, and because singing techniques imply the control of resonances in this range [19, 20].

Above 4–5 kHz the transfer functions obtained with FEM and MM have a different aspect than at lower frequency. There are more variations in the number, amplitude, and bandwidth of the resonances between the different vowels and antiresonances can be seen (except in Fig. 2a). This can be understood as the effect of higher order modes [5].

The TLM gives substantially different transfer functions than the 3D methods above 4–5 kHz, in particular the vowel /u/ (Figs. 2g and 2h). This can be easily understood as the consequence of the fact that the 3D aspect of the acoustic field cannot be accounted for by this simulation method. However, some differences could be potentially reduced using a more accurate approximation of the baffled piston radiation model: a low frequency approximation was used in this implementation.

Generally, the -3 dB bandwidth of the resonances is smaller for MM than FEM: of 58 resonances analyzed, 46 have a smaller bandwidth. The total average relative difference is of 18.1%. The difference is more pronounced for the female configuration: it is on average 19.9% and 16.4% for female and male resonances, respectively. It is also more pronounced for closed vowels: 21.6% for /i/ and even more for /u/, 25.7%, whereas it is 13.9% and 11.3% for /a/ and /ə/ respectively.

The smaller damping of the MM may be related to the model of junction between the segments. In fact, it assumes that the portion of vocal tract wall on the interface between two segments is perfectly reflective (assuming a 0 axial velocity) [5]. Given the small area implied, this is not expected to have a large impact on the simulation, but can potentially reduce the bandwidths of the resonances and shift the resonance frequencies. However, it could have a stronger impact for strong area discontinuities (e.g. a doubling of cross-sectional area). This may also explain the stronger difference in the range 2-4 kHz if it appears that the resonance in this frequency range imply reflection on strong discontinuities. On the other hand, the greater bandwidth difference for closed vowels may be due to the fact that more internal reflections are involved for these vowels. However, this needs to be confirmed with a proper modelling of the wall absorption at the discontinuities.

Another source of difference maybe the artificial stiffening introduced by the finite element discretization [21]. This can shift resonances and anti-resonances to higher frequencies. As an example, the anti-resonance, which occurs above 8 kHz in /i/ of the female configuration, occurs at a higher frequency for FEM (see Fig. 2e). This anti-resonance is the signature of a transverse resonance due to higher-order modes. The artificial stiffening would induce an overestimation of the cut-off frequency of the the higher order mode implied, and thus, of the anti-resonance that it generates. This effect diminishes when the element size is reduced. Thus, the difference between FEM and MM could be explained by the difference of element sizes used for FEM and for the computation of the propagation modes for MM (about 3 mm for FEM and 1.2 mm on average for MM).

In the MM the radiation is described by a radiation impedance matrix whereas FEM naturally accounts for radiation losses by allowing sound waves to emanate from the mouth. Given the relatively good agreement between both methods, it can be assumed that radiation is reasonably well modeled by both approaches. However, performing simulations with the same 0 pressure condition at the mouth surface, or extracting the radiation impedance matrix would more accurately confirm this.

Finally, the small geometric differences may also contribute to the differences between FEM and MM.

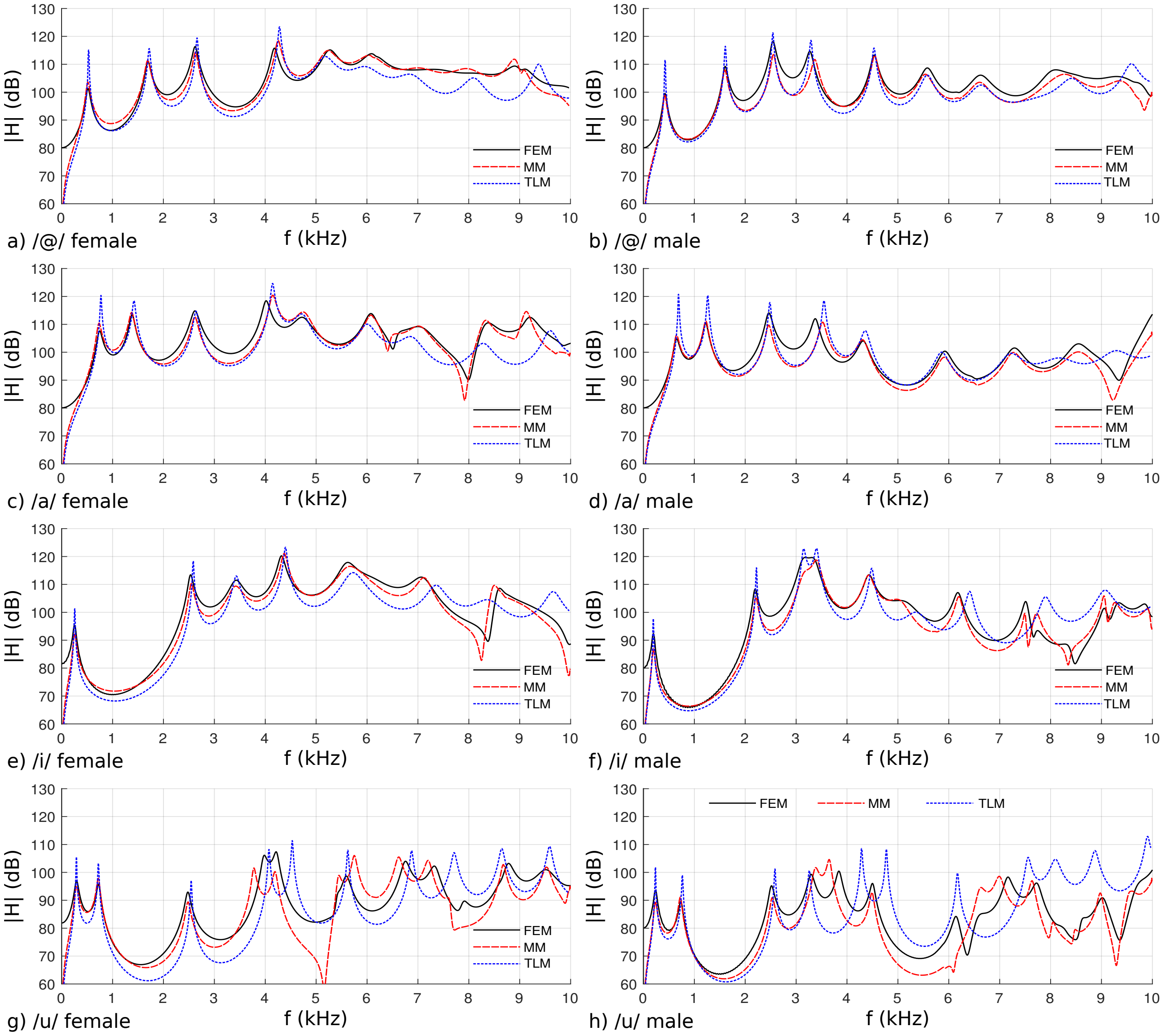

Figure 2: *Vocal tract transfer functions of different vowels computed with finite elements (FEM), the multimodal method (MM) and the transmission-line model (TLM) implemented in VocalTractLab for female and male vocal tract geometries.*

The TLM simulations run at real time. The computation time of the FEM is about 30 hours to simulate 50 ms using an Intel® Core^{TM} i7-6700 processor, which allows us to obtain a transfer function up to 10 kHz with 1 Hz of resolution. Lower computational times could be achieved truncating the computational domain at the lip opening and imposing a radiation impedance on it, as in [11], which needs about 10 hours in a single processor to compute, in the frequency domain, 1000 frequencies. The computation time of MM is on average 40 minutes for 1000 frequencies with an Intel® Xeon® W-2145 processor with 3.7 GHz.

4. Conclusions

The MM generates transfer functions very similar to the ones obtained with FEM with a significantly shorter computation time. Thus, the implementation of the evaluated MM seems a promising method to reduce the computation time of 3D vocal tract acoustic simulations. However, significant differences remain to be understood, and thus the FEM remains for now the most reliable method for 3D acoustic simulations.

The TLM simulation generates valid results at low frequencies (for the first four resonances) in comparison with the 3D methods. However, improvements of the segmentation algorithm or the radiation condition modelling may further improve the accuracy and extend its frequency range of validity.

5. Acknowledgements

The first and last authors acknowledge the support of the German Research Foundation (DFG) with the grant no. BI 1639/7-1. The second author acknowledges the support of grant 2020-URL-Proj-056 from the Generalitat de Catalunya and the Universitat Ramon Llull.

6. References

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

[2] O. Engwall, "Vocal tract modeling in 3D," *TMH-QPSR*, vol. 1, pp. 1–8, 1999.

[3] P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[4] P. Anderson, S. Fels, N. M. Harandi, A. Ho, S. Moisik, C. A. Sánchez, I. Stavness, and K. Tang, "Frank: A hybrid 3D biomechanical model of the head and neck," in *Biomechanics of Living Organs*. Elsevier, 2017, pp. 413–447.

[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. 832–843, 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] S. Dabbaghchian, M. Arnela, O. Engwall, and O. Guasch, "Reconstruction of vocal tract geometries from biomechanical simulations," *International journal for numerical methods in biomedical engineering*, vol. 35, no. 2, p. e3159, 2019.

[8] S. Dabbaghchian, M. Arnela, O. Engwall, and O. Guasch, "Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 37, no. 1, p. e3407, 2021.

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

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

[12] S. Drechsel, Y. Gao, J. Frahm, and P. Birkholz, "Modell einer frauenstimme für die artikulatorische sprachsynthese mit vocaltractlab," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 239–246, 2019.

[13] A. Maurel, J. Mercier, and S. Félix, "Propagation in waveguides with varying cross section and curvature: a new light on the role of supplementary modes in multi-modal methods," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 470, no. 2166, p. 20140008, 2014.

[14] R. Blandin, A. Van Hirtum, X. Pelorson, and R. Laboissière, "Multimodal radiation impedance of a waveguide with arbitrary cross-sectional shape terminated in an infinite baffle," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2561–2564, 2019.

[15] O. Guasch, M. Arnela, R. Codina, and H. Espinoza, "A stabilized finite element method for the mixed wave equation in an ale framework with application to diphthong production," *Acta Acustica united with Acustica*, vol. 102, no. 1, pp. 94–106, 2016.

[16] M. Arnela, S. Dabbaghchian, O. Guasch, and O. Engwall, "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2173–2182, 2019.

[17] M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 955–967, 1987.

[18] M. Sondhi, "An improved vocal tract model," *Proceedings of the 11th ICA. Paris, France*, pp. 167–170, 1983.

[19] J. Sundberg, "Articulatory interpretation of the "singing formant"," *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 838–844, 1974.

[20] J. Sundberg, F. M. Lã, and B. P. Gill, "Formant tuning strategies in professional male opera singers," *Journal of Voice*, vol. 27, no. 3, pp. 278–288, 2013.

[21] J. O. Dow and D. E. Byrd, "The identification and elimination of artificial stiffening errors in finite elements," *International Journal for Numerical Methods in Engineering*, vol. 26, no. 3, pp. 743–762, 1988.