27th International Congress
on Sound and Vibration
The annual congress of
the International Institute
of Acoustics and Vibration (IIAV)

11-16 July, 2021

ICSV27

Annual Congress of the International Institute of Acoustics and Vibration (IIAV)

COMBINING MULTIMODAL METHOD AND 2D FINITE ELEMENTS FOR THE EFFICIENT SIMULATION OF VOCAL TRACT ACOUSTICS

Rémi Blandin
Institute of Acoustics and Speech Communication , TU Dresden, Germany
e-mail: remi.blandin@tu-dresden.de
Simon Félix
Laboratoire d'Acoustique de l'Université du Mans, LAUM, UMR 6613, CNRS, Le Mans Université, France
email: simon.felix@univ-lemans.fr
Jean-Baptiste Doc
Conservatoire National des Arts et Metiers, Paris, France
email: jean-baptiste.doc@lecnam.net
Peter Birkholz
Institute of Acoustics and Speech Communication , TU Dresden, Germany
email: peter.birkholz@tu-dresden.de

The resonances of the vocal tract determine the different vowels of speech. But beyond that, they are a key factor in voice quality and singing techniques. Very simple and efficient acoustic models are extensively used to simulate vocal tract acoustics in the frequency range for which the plane wave assumption is valid (up to 4-5 kHz). However, outside of this range, more complex models must be used. A trade-off must be found between the accuracy of the description of the vocal tract geometry and the computational cost. Thus, as an example, 3D finite elements allow one to take into account all the details of the anatomy but require a lot of computational resources. On the other hand, the state-of-the-art multimodal methods can have a much lower computational cost, but require some simplification of the anatomy. The present work proposes to combine the advantages of both methods to simulate vocal tract acoustics with a reasonable computational cost and an accurate anatomical description. For this purpose, the vocal tract shape is sliced in successive planes, allowing one to simulate the multimodal propagation along a centerline connecting the frames. On each frame, the local transverse modes are computed with 2D finite elements, which allows one to take into account accurately the cross-sectional shapes.

Keywords: transverse modes, finite elements, multimodal method

1. Introduction

The vocal tract is defined as the air volume between the vocal folds and the lips. The resonances of this volume determine the different vowels of speech. Their frequency, amplitude and bandwidth vary with the shape and position of the articulators. In addition to encoding intelligible sounds, vocal tract resonances play a role in voice quality, and their control is a key element of singing techniques. One of the most well known illustration of resonance tuning is the so called singer's formant [1].

The plane wave model The acoustic properties of the vocal tract have been and are still nowadays extensively studied using the plane wave model [2, 3, 4]. This model relies on the assumption that, below the first cutoff frequency of the vocal tract, the acoustic waves can be approximated as uniform over transverse cross-sections of the vocal tract. Under this assumption, only the cross-sectional area along the vocal tract determines its acoustic properties. Thus, its 3D shape can be simplified using a very simple one dimensional description, often referred to as area function. The simplicity of this model enables very fast and efficient computations of the acoustic propagation inside the vocal tract, and thus, allows one to explore a wide variety of vocal tract configurations. The plane wave model reproduces accurately acoustic properties up to 4-5 kHz [5].

High frequencies in speech Even though, as the telephone bandwidth (300-3400 Hz) illustrates it, most of the relevant information in speech is below 5 kHz, interest in high frequencies recently increased. This is motivated by the development of speech-related technologies using frequencies beyond 5 kHz, such as video conference, wideband telephony, assistive technologies or virtual voice synthesis. This interest is motivated by the presence of perceptually relevant information beyond 5 kHz. Different studies have shown an influence of high frequencies on intelligibility, communication in noise, voice quality or localisation [6]. Thus, there are fundamental research questions regarding speech perception of high frequencies. As an example, a part of the answer to the so called cocktail party problem may be found in high frequencies [7]. More generally, the relationship between high-frequency acoustic signals and speech and singing perception is not well understood.

High-frequency simulation of vocal tract acoustics Simulating accurately acoustic propagation above 5 kHz in the vocal tract requires to take into account the three dimensional (3D) shape of the vocal tract.

Motivated by the increased interest in high frequencies in speech, or simply in order to investigate the effect of the precise 3D shape of the vocal tract regardless of the frequency, other simulation methods than the plane-wave model have been used. Thus, different numerical methods, such as finite elements (FE) [8], finite differences (FD) [9] or waveguide mesh [10] have been applied to the vocal tract. These simulation methods have a much higher computational cost than plane-wave simulations, the computation times being typically of the order of several hours for the computation of the transfer function of a single vocal tract geometry. Thus, much less vocal tract configurations can be explored. Nevertheless, works performed using these methods have highlighted how the side branches (pyriform sinus, nasal tract), the curvature and cross-sectional shape of the vocal tract substantially influence its acoustic properties from 4-5 kHz on.

Another approach, is to extend the plane wave theory to take into account the higher order propagation modes. These types of methods, known as multimodal methods, can be much more efficient than the methods discretizing the 3D vocal tract volume, especially if an analytical expression is used for the propagation modes. However, it requires simplification of the geometries, and only approximated vocal tract shapes can be studied. As an example, Motoki and Matsuzaki approximated the vocal tract with straight waveguides with rectangular cross-sections [11]. The lower computational cost of the multimodal method allowed them to perform numerous simulations which showed that the acoustic properties of the vocal tract above 4 kHz should be very sensitive to small pertubations of the vocal tract geometry. However, due to the absence of curvature and the substantial difference between actual cross-sectional shapes and rectangles, caution must be taken in generalizing this result to realistic vocal tract shapes.

Limitations of high-frequency simulations Even though valuable information have been obtained regarding the high-frequency acoustic properties of the vocal tract, it is difficult to generalize them since

these results rely on a limited number of geometries and/or simplified ones. Thus, more extensive exploration of the vocal tract high-frequency properties is necessary. On the other hand, the relationship between the observed properties and the high-frequency speech perception is not well known and investigated, because of the lack of high-frequency valid articulatory speech synthesis. Such a synthesis requires an accurate description of the high-frequency vocal tract acoustics at a reasonable computational cost. Thus, it appears as very beneficial for research concerning high frequencies in speech and singing to increase the efficiency of the acoustic simulations.

Proposed approach The objective of this work is to propose an efficient simulation method, relying on the multimodal and the two dimensional (2D) FE methods. In order to overcome the limitations of the multimodal method in term of cross-sectional shape, the transverse modes are computed using two dimensional FE. The curvature and change of the size of the cross-sections are taken into account using a geometrical transformation, as proposed by Maurel *et al* [12].

2. Geometrical transformation and transverse modes

Hereafter, bold letters and symbols represent vectors.

2.1 Simplification of the 3D vocal tract geometry

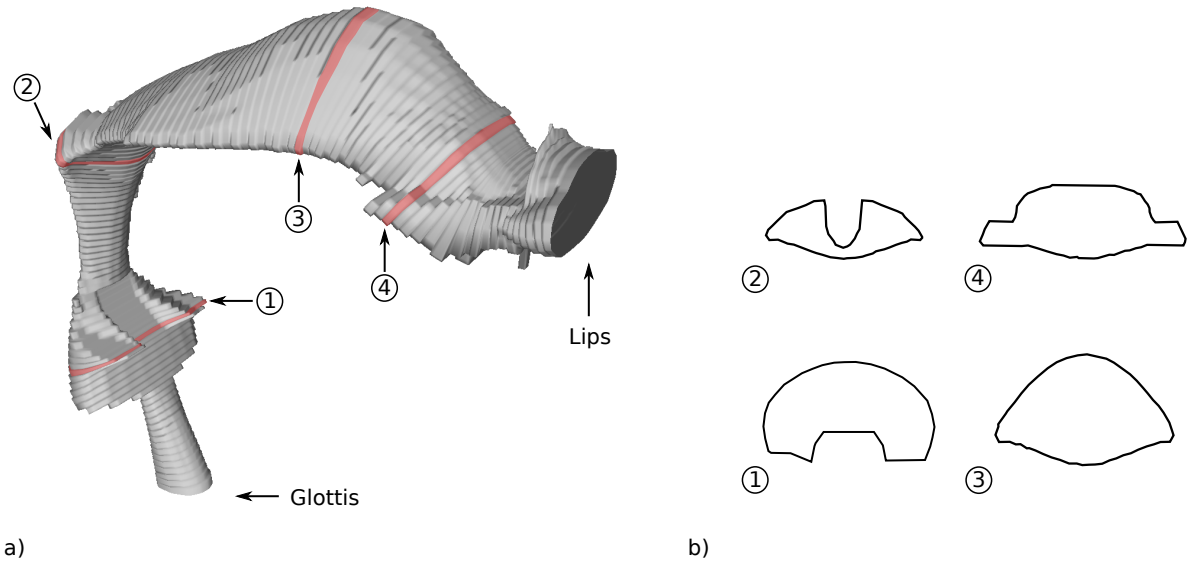

Figure 1: a) Vocal tract geometry corresponding to the vowel /a/ extracted from the articulatory model implemented in the speech synthesiser Vocaltractlab [13] sliced into 129 segments, b) cross-sectional shapes of different segments.

In order to apply the proposed method, the vocal tract geometries are sliced in segments with constant cross-sectional shape. Inside each segment the curvature and the cross-sectional area vary. The geometry is discontinuous at the junction between the segments.

The Fig. 1a presents an example of sliced 3D geometry corresponding to the vowel /a/ obtained with the articulatory synthesiser Vocaltractlab [13]. In Fig. 1b examples of cross-sectional shapes corresponding to segments highlighted in Fig. 1a are presented. One can see that the cross-sectional shapes are complex and can assume a variety of forms. This variability is further increased by the various positions

that the articulators can take. Thus, in order to simulate the acoustic properties of the vocal tract, one needs a generic method which allows one to take into account a wide variety of geometries. This method must solve two problems:

- linking different arbitrary cross-sectional shapes,
- taking into account changes of size and curvature.

2.2 Wave equation

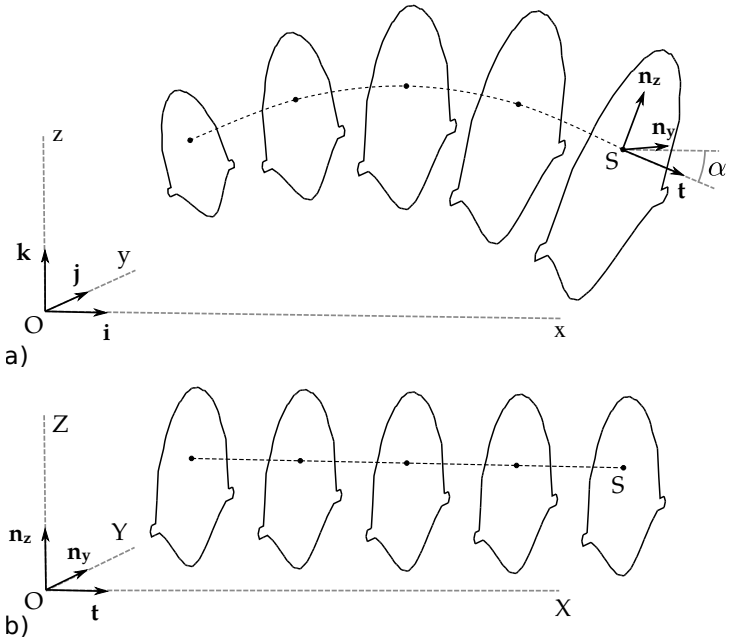

Figure 2: Constant cross-sectional shape waveguide segment represented (a) in Cartesian coordinates (x, y, z) and (b) in the transformed coordinates (X, Y, Z) removing curvature and cross-sectional area variations.

Each segment is transformed into a straight segment with constant cross-sectional area, using a coordinate transformation that is defined following Maurel *et al* [12]. The Fig. 2a shows an example of a segment represented in Cartesian coordinates (x, y, z), and in Fig. 2b, the same segment is shown in the transformed coordinates (X, Y, Z). The position of a point P can be described in both coordinate systems by the vector

$$\boldsymbol{OP} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} = \boldsymbol{OS}(X) + l(X)Y\boldsymbol{n_Y} + l(X)Z\boldsymbol{n_Z}, \tag{1}$$

where S is a point on the centerline and $l(X)$ a function that compensates the variations of cross-sectional area. Expressing $d\boldsymbol{OM}$ in both coordinate systems allows one to find the expression of the Jacobian matrix J of the transformation, which satisfies the relashionship

$$\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = J^{-1} \begin{pmatrix} dX \\ dY \\ dZ \end{pmatrix}, \text{ with } J^{-1} = \begin{pmatrix} f\cos(\alpha) - l'Z\sin(\alpha) & 0 & -l\sin(\alpha) \\ l'Y & l & 0 \\ f\sin(\alpha) + l'Z\cos(\alpha) & 0 & l\cos(\alpha) \end{pmatrix}, \tag{2}$$

$f = 1 - \kappa l Z$, where $\kappa(X)$ is the curvature, $l' = \frac{\partial l}{\partial X}$ and α is the angle between $\boldsymbol{i}$ and $\boldsymbol{t} = \frac{d\boldsymbol{OM}}{dX}$. Note that since the transformation does not change the cross-sectional shape, it does not end up in a separable geometry in the new coordinate system.

The proposed geometric transformation is applied to the wave equation

$$\left(\Delta + k^2\right) p = 0 \ , \tag{3}$$

in which k is the free field wavenumber, and p the acoustic pressure. For this purpose the Laplacian operator must be expressed in term of J, which transforms Eq. (3) in

$$\mathrm{div}\left(H\boldsymbol{\nabla} p\right) + \frac{k^2}{\det J} p = 0 \ , \text{ with } H = \frac{{}^{\mathrm{t}}JJ}{\det J}. \tag{4}$$

Introducing a new physical field, the axial velocity q related to the acoustic pressure p, Eq. (4) can be expressed as a system of first order differential equations

$$\frac{\partial}{\partial X}\begin{pmatrix} p \\ q \end{pmatrix} = M \begin{pmatrix} p \\ q \end{pmatrix} \ , \tag{5}$$

where M is a matrix whose terms depends on the expression of H and $\det J$. The geometrical complexity now lies in the wave equation and boundary condition.

2.3 Computation of the transverse modes

Since the cross-sectional shapes of the vocal tract are very different and relatively complex (see Fig. 1b), the simplest way to compute the transverse modes is to rely on a numerical method. One advantage of FE is that it allows a very accurate discretization of the cross-sectional surface compared to other methods. This insures a better accuracy in the computation of the eigen-frequencies of the transverse modes. The precision of the eigen-frequencies influences the accuracy of the propagation constants and hence the resonance frequencies. Transverse modes are computed for each constant cross-section, which thus, has its own modal basis.

The Fig. 3 presents the 5 first transverse modes and related eigen-frequencies corresponding to the cross-sectional shapes presented in Fig. 1b. The eigen-frequencies are not necessarily correlated with the cross-sectional area: the shape 2 has the lowest first eigen-frequency ($f_1 = 4.1kHz$) whereas it has the smallest cross-sectional area.

3. Multimodal method

3.1 Projection of the acoustic field

In each segment, the pressure field p and axial velocity q are developed as series over the local transverse modes $\varphi_n(Y,Z)$:

$$\begin{cases} p(X,Y,Z) &= \sum\limits_n p_n(X)\varphi_n(Y,Z), \\ \\ q(X,Y,Z) &= \sum\limits_n q_n(X)\varphi_n(Y,Z). \end{cases} \tag{6}$$

The amplitudes of the transverse modes $p_n(X)$ and $q_n(X)$ can be gathered in two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$. Eq. (5) is then projected onto these modes, giving a coupled matrix equation for the components of $\boldsymbol{p}$ and $\boldsymbol{q}$ in the basis

$$\frac{\partial}{\partial X}\begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix} = \mathcal{M} \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix}. \tag{7}$$

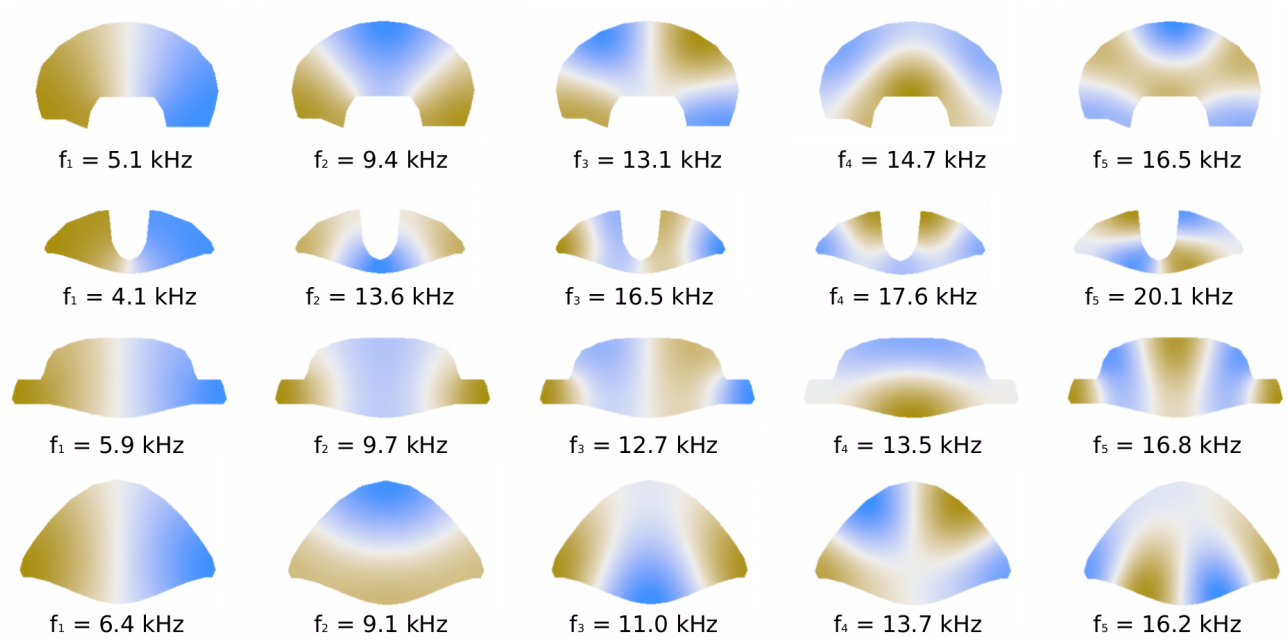

Figure 3: Mode shapes and eigen-frequencies of the 5 first transverse modes corresponding to the cross-sectional shapes presented in Fig. 1b computed with linear finite elements.

3.2 Junction between segments

In order to link two consecutive segments, a mode matching condition is written to satisfy the continuity of the fields. This condition is described by a coupling matrix F which allows one to relate the amplitudes of the transverse modes on both sides of the junction. Thus, for the acoustic pressure, the amplitudes $\boldsymbol{p^{(1)}}$ and $\boldsymbol{p^{(2)}}$ obey the relationship

$$\boldsymbol{p^{(1)}} = F\boldsymbol{p^{(2)}}, \tag{8}$$

with

$$F_{mn} = \int_S \bar{\varphi}_m^{(1)} \varphi_n^{(2)} dS, \tag{9}$$

in which $\varphi_m^{(1)}$ and $\varphi_n^{(2)}$ are the transverse modes of both cross-sectional shapes, the bar means conjugate and S is the surface intersection of both cross-sectional surfaces.

3.3 Computation of the acoustic field

To compute the amplitudes of the transverse modes everywhere in the vocal tract geometry, it is more convenient to solve the problem with an impedance matrix. Indeed, this allows one to avoid some numerical stability issues, and to describe the boundary condition at the lips as a radiation impedance matrix. This matrix can be calculated following the method proposed by Felix *et al* [14]. Then, it is used as an initial condition to compute the impedance matrix along the segments using a Magnus-Möbius scheme to integrate Eq. (7) as described by Pagneux [15]. At the junctions, the impedance matrix can be computed at the entrance of the following segment with the coupling matrix F (Eq. (9). This alternation of numerical integration along the segments and coupling at the junctions is repeated up to the sound source location (the glottis for vowel sounds). At this location, the axial velocity field of the sound source is projected over the local transverse modes. The corresponding projection for the acoustic

pressure is obtained from the impedance matrix, and the acoustic field can be computed in every segment of the vocal tract geometry using a procedure similar to the one used for the impedance matrix.

4. Summary and outlook

In this paper, a method that is valid at high frequencies is proposed for the simulation of vocal tract acoustics. It relies on a combination of 2D FE for the computation of the transverse modes, and a multimodal method for their propagation. Such a combination is expected to increase the efficiency of vocal tract acoustic simulations.

This increased efficiency is intended to be exploited for wideband articulatory speech synthesis, that is, speech synthesis relying on an acoustic model valid in the entire audible frequency range. This is intended to be implemented in the framework of the articulatory synthesiser Vocaltractlab [13]. Wideband speech synthesis can be used in the future to generate stimuli to investigate speech perception of high frequencies.

REFERENCES

1. Sundberg, J. Level and center frequency of the singer's formant, *Journal of voice*, **15** (2), 176–186, (2001).

2. Fant, G., *Acoustic theory of speech production*, no. 2, Walter de Gruyter (1970).

3. Stevens, K., *Acoustic phonetics*, vol. 30, MIT press (2000).

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

5. Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Van Hirtum, A. and Laval, X. Effects of higher order propagation modes in vocal tract like geometries, *The Journal of the Acoustical Society of America*, **137** (2), 832–843, (2015).

6. Monson, B., Hunter, E., Lotto, A. and Story, B. The perceptual significance of high-frequency energy in the human voice, *Frontiers in Psychology*, **5**, 587, (2014).

7. Monson, B., Rock, J., Schulz, A., Hoffman, E. and Buss, E. Ecological cocktail party listening reveals the utility of extended high-frequency hearing, *Hearing Research*, **381**, 107773, (2019).

8. Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Van Hirtum, A. and Pelorson, X. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds, *The Journal of the Acoustical Society of America*, **140** (3), 1707–1718, (2016).

9. Takemoto, H., Mokhtari, P. and Kitamura, T. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method, *The Journal of the Acoustical Society of America*, **128** (6), 3724–3738, (2010).

10. Gully, A., Yoshimura, T., Murphy, D., Hashimoto, K., Nankaku, Y. and Tokuda, K. Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network, *Proc. of the Interspeech*, Stockholm, Sweden, pp. 234–238, ICSLP, (2017).

11. Motoki, K. and Matsuzaki, H. Computation of the acoustic characteristics of vocal-tract models with geometrical perturbation, *Proc. of the Interspeech*, Jeju, Korea, pp. 521–524, ICSLP, (2004).

12. Maurel, A., Mercier, J. and Félix, S. 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*, **470** (2166), 20140008, (2014).

13. Birkholz, P. Modeling consonant-vowel coarticulation for articulatory speech synthesis, *PloS one*, **8**, e60603, (2013).

14. Félix, S., Doc, J. and Boucher, M. Modeling of the multimodal radiation from an open-ended waveguide, *The Journal of the Acoustical Society of America*, **143** (6), 3520–3528, (2018).

15. Pagneux, V. Multimodal admittance method in waveguides and singularity behavior at high frequencies, *Journal of computational and applied mathematics*, **234** (6), 1834–1841, (2010).