<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *IEEE Transactions on Audio, Speech, and Language Processing*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Arnela, M., Dabbaghchian, S., Guasch, O., Engwall, O. (2019)

MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs

IEEE Transactions on Audio, Speech, and Language Processing, 27(12): 2173-2182

<https://doi.org/10.1109/TASLP.2019.2942439>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-259580>

MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs

Marc Arnella, Saeed Dabbaghchian, Oriol Guasch and Olov Engwall

Abstract—The synthesis of diphthongs in three-dimensions (3D) involves the simulation of acoustic waves propagating through a complex 3D vocal tract geometry that deforms over time. Accurate 3D vocal tract geometries can be extracted from Magnetic Resonance Imaging (MRI), but due to long acquisition times, only static sounds can be currently studied with an adequate spatial resolution. In this work, 3D dynamic vocal tract representations are built to generate diphthongs, based on a set of cross-sections extracted from MRI-based vocal tract geometries of static vowel sounds. A diphthong can then be easily generated by interpolating the location, orientation and shape of these cross-sections, thus avoiding the interpolation of full 3D geometries. Two options are explored to extract the cross-sections. The first one is based on an adaptive grid (AG), which extracts the cross-sections perpendicular to the vocal tract midline, whereas the second one resorts to a semi-polar grid (SPG) strategy, which fixes the cross-section orientations. The finite element method (FEM) has been used to solve the mixed wave equation and synthesize diphthongs [ai] and [au] in the dynamic 3D vocal tracts. The outputs from a 1D acoustic model based on the Transfer Matrix Method have also been included for comparison. The results show that the SPG and AG provide very close solutions in 3D, whereas significant differences are observed when using them in 1D. The SPG dynamic vocal tract representation is recommended for 3D simulations because it helps to prevent the collision of adjacent cross-sections.

Index Terms—vocal tract acoustics, Finite Element Method, diphthongs, semi-polar grid, adaptive grid, speech synthesis.

I. INTRODUCTION

Articulatory speech synthesis has the potential to generate high quality and natural voice acoustics. For historical reasons and thanks to their simplicity, one-dimensional (1D) approaches have become widely exploited [1]–[4]. These models approximate the complex three-dimensional (3D) vocal tract geometry as a set of concatenated tubes of constant areas, which correspond to the cross-sections perpendicular to the vocal tract midline (1D area functions, see e.g. [5]). Assuming plane wave propagation, the acoustic pressure can be computed at the exit of the vocal tract, resulting in the synthesized sound.

Marc Arnella and Oriol Guasch are with the GTM - Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, Barcelona, Catalonia, Spain (e-mail: marc.arnella@salle.url.edu; oriol.guasch@salle.url.edu), and Saeed Dabbaghchian and Olov Engwall with the Department of Speech, Music, and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: saeedd@kth.se; engwall@kth.se). This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors. This includes WAV-format audio files of the synthesized diphthongs in the manuscript. This material is 644 KB in size. Contact marc.arnella@salle.url.edu for further questions about this work.

These models can currently generate many types of sounds (i.e., those that may be represented by the midsagittal plane), either static or dynamic, at a very low computational cost, yet at the price of simplifying the acoustics of the vocal tract. In fact, 1D models are only accurate in the low frequency range, up to around 5 kHz, where, as said, the assumption of plane propagation holds. However, beyond this limit not only planar modes are produced but also higher order ones [6], [7], which strongly affect the high frequency energy content of the spectrum. This is known to be relevant for achieving more natural sounds [8], [9].

Three-dimensional approaches can overcome the limitations of 1D because they account for complex 3D shape of the vocal tract, resulting in a more comprehensive description that includes propagation of plane waves as well as higher-order modes (see e.g. [10] and [11], which show a very good matching between simulations and experiments performed in 3D-printed mechanical replicas). The price to be paid, however, is that of a significantly higher computational cost. In this sense, it is worthwhile mentioning that some works have tried to find a balance between 1D and 3D models, working in 2D. For instance, in [12] 2D diphthongs were generated in real time, while in [13] it was proposed to use 2D tuned vocal tracts to improve the accuracy of the synthesized diphthongs.

Many works can be found in literature dealing with the numerical production of static vowel sounds in 3D. Various numerical techniques have been employed, such as the finite element method (FEM) in the time domain [7], [11], [14]–[17], or in the frequency domain [18]–[21], finite difference schemes [10], multimodal techniques [6], [22], or digital waveguide approaches [23], [24]. Less attention has been paid, however, to the synthesis of vowel-vowel utterances, like diphthongs. To generate this type of sound one has to take into account not only the movement of the 3D vocal tract geometry, but also the 3D propagation of sound waves in a 3D moving domain. Such a problem was first addressed in [25] using FEM in the time domain, and a simplified vocal tract geometry consisting of a straight 3D tube with circular cross-sections, driven by 1D area functions. A dynamic 3D vocal tract geometry was then obtained by linearly interpolating the 1D area function from a starting vowel to an ending one, so as to produce a diphthong. Although 3D wave propagation was obviously recovered, the vocal tract radial symmetry prevented the onset of higher order modes [7], resulting in a poor representation of the high frequency energy content.

Dynamic 3D representations of a detailed vocal tract are

therefore essential to generate naturally sounding diphthongs. Unfortunately, the construction of such a dynamic representation is not obvious. One option could be that of resorting to medical imaging, typically to Magnetic Resonance Imaging (MRI), but at present only static vocal tracts can be obtained with a high spatial resolution [20]. Further research is still needed to obtain time-evolving MRI-based vocal tracts with a simultaneously adequate spatial and time resolution. Another option would consist in using a biomechanical model, which allows one to drive the different articulators of the vocal tract. The authors of this work have recently extracted accurate geometries from an adapted version of FRANK [26], a 3D biomechanical model developed in Artistry [27], and employed them to generate some static vowel sounds [28]. It is to be mentioned that ongoing work is under development to generate diphthongs following this approach [29].

Alternatively, in [30] it was proposed to generate diphthongs using a 3D dynamic digital waveguide model and MRI-based vocal tracts. In that work the computational domain consisted of a fixed cuboid volume discretized with a Cartesian 3D grid. The specific impedance at each point of the grid was adjusted according to an impedance map, which contained, for each static vowel configuration, the impedance of the vocal tract airway (set to that of air) and the impedance of the surrounding tissues (head). Diphthongs were then generated by interpolating the impedance maps for different vowel sounds, avoiding movement of the computational domain.

A very different methodology is suggested in this work. We propose to generate diphthongs using a realistic 3D dynamic representation of the vocal tract that deforms its walls over time. In particular, our aim is to extend the dynamic FEM acoustic model in [25], based on straight circular vocal tracts, to 3D vocal tracts reconstructed from MRI. The interpolation between such geometries to generate a time-varying vocal tract shape for diphthong generation is an intricate task that, to the best of the authors' knowledge, has not yet been addressed. If one wants to directly deal with the full complex MRI geometries the only option is to resort to remeshing strategies, but unfortunately those are computationally very expensive. A less demanding option is to generate an initial mesh and then distort it following a diffusion process, or solving an elasticity problem. However, this poses some other difficulties. One has to transition between smooth deformations of the vocal tract walls while preventing, for instance, the generation of self-intersections in the vocal tract boundaries during the interpolation. Moreover, the number of vertices in the initial and final static MRI-based vocal tract geometries is different, increasing the complexity of the interpolation process. To circumvent such issues, we have decided to deal with slightly simplified 3D vocal tracts with cross-sections extracted from static MRI-based vocal tract geometries.

For the latter purpose, the geometries from [20] are discretized into a set of cross-sections, as usually done to extract 1D area functions. However, in our case we will preserve the cross-sectional shapes, as well as their location and orientation in the vocal tract midline. These three parameters can be used to reconstruct any 3D vocal tract geometry [7] and their interpolation would allow us to generate dynamic vocal tracts. To

that goal, we only need to ensure that all vocal tract geometries have the same number of cross-sections, and that each cross-section contains the same number of vertices. The latter can be achieved by resampling the cross-sectional shapes. Note that, in this way, a complex 3D interpolation problem is reduced to an easier one in which only 2D (shape) and 1D (location and orientation) functions need to be interpolated.

There exist different methods to extract cross-sections from an MRI-based vocal tract geometry. A first option considers the cross-sections to be perpendicular to the vocal tract midline. That method, hereafter termed Adaptive Grid (AG) (in accordance to [7]), is typically used to extract 1D area functions, since the cross-sections are aligned with the planar wavefronts propagating through the vocal tract. However, the AG can present some difficulties for 3D modeling. Adjacent cross-sections may collide during the interpolation process because their orientations change over time. An alternative solution consists in using a semi-polar grid (SPG) approach to extract the cross-sections. The SPG grid is made of horizontal planes in the pharynx, polar planes in the velar region, and vertical planes in the oral cavity. The cross-sectional orientation is fixed over time, which prevents the collision of adjacent cross-sections. The price to be paid is that the extracted cross-sections will no longer coincide with the planar wavefronts, as happens with the AG. Therefore, 1D area functions should not be extracted from the SPG cross-sections in 1D models because that would modify the generated sound. In contrast, this will not pose any problems for a 3D simulation, since the cross-sections are simply used to reconstruct the 3D vocal tract geometry, where the acoustic waves freely propagate in any direction. In this work we will examine all these assumptions. Some preliminary results were presented in [31] and [32].

The paper is organized as follows. Section II introduces the dynamic 3D vocal tract representations based on the interpolation of cross-sections. The AG and SPG methods are considered to extract the cross-sections from the MRI-based vocal tracts. Section III describes the time-domain FEM that we have employed to synthesize vowels and diphthongs. Section IV shows the simulation results. First, the static vowel sounds [a], [i] and [u] are analyzed to evaluate the performance of the AG and SPG methods. For completeness, conventional 1D simulations (similar to those in [1]) are also conducted and compared against 3D, using the corresponding 1D area functions extracted from the two grids. Next, the diphthongs [ai] and [au], are generated with the proposed 3D model and also with a conventional 1D synthesizer. The outputs are again compared and discussed in terms of the AG versus the SPG, and 3D versus 1D. Section V finally concludes the paper.

II. THE DYNAMIC 3D VOCAL TRACT REPRESENTATIONS

A. Extraction of cross-sections from MRI-based vocal tracts

The dynamic representations are based on the 3D vocal tract geometries of vowels [a], [i] and [u] generated from MRI in [20] (dental casts were superimposed onto the MRI data to consider the teeth). First, and for simplicity, the lips and side-branches such as the sinus piriformi and the vallecula are excluded from the geometries (lips mainly shift down some

Fig. 1. Extraction of cross-sections using a semi-polar grid (SPG). (a) Vocal tract of [i] with some cutting planes. (b) SPG superimposed on the vocal tract profile, with (1) vertical planes in the oral cavity, (2) polar planes in the velar region, horizontal planes (3a) in the oropharynx and (3b) in the larynx.

of the formants [11], [17], whereas side-branches introduce some anti-resonances in the spectrum [10]). Each geometry is then intersected with a set of cutting planes, obtaining a total of 40 cross-sections. The shape of each cross-section is next resampled to have 48 vertices. The vocal tract midline is finally formed by connecting the center of all cross-sections, from which we obtain the location and orientation of each cross-section. This number of cross-sections and vertices was deemed accurate enough according to results in [7] for simplified 3D vocal tract geometries of static vowels. Moreover, it is in line with works in literature (see e.g. [5], which presents area functions between 40 and 46 cross-sections).

A first option to define the cutting planes consists in using a semi-polar grid (see Figure 1). This grid is typically divided in three regions with vertical, oblique and horizontal planes, respectively. The vertical planes are distributed equidistantly along the anterior-posterior axis in the oral cavity, with the mouth termination plane as the most anterior one. The oblique part covers the velar region with uniform angle distribution. The horizontal cutting planes are positioned between the glottal cross-section and the oropharynx. However, this configuration does not guarantee the same number of planes cutting the larynx tube, given their uniform distribution in the grid and the varying length of the vocal tract. This may result in a situation in which a cross-section belonging to the larynx has to be interpolated to another one in the oropharynx. That would result in some difficulties during the interpolation process because of the abrupt change of cross-sectional shape from the larynx to the pharynx. To circumvent that inconvenience, we have added a constraint to the horizontal part of the grid. We use the same number of cutting planes in the larynx tube for all geometries, whereas the rest of the horizontal portion is distributed equidistantly. In this way, abrupt changes in the cross-sectional shape during the interpolation are minimized.

The second option examined in this work is that based on an Adaptive Grid (AG) [7]. Cross-sections resulting from the SPG method are approximately perpendicular to the vocal tract midline. However, one can be more accurate and use this midline as a starting point to determine the cutting planes that are perpendicular to it. This results in the new set of cutting planes that constitute the AG.

Fig. 2. Dynamic vocal tract representations at four time instants based on cross-sections extracted with an adaptive grid (AG) and a semi-polar grid (SPG), for the generation of diphthongs [ai] and [au].

B. Construction of the dynamic vocal tract representations

The vocal tract representations are based on the cross-sections extracted from the static vowels [a], [i] and [u]. In contrast to 1D synthesizers based on area functions, the shape of each cross-section is preserved, together with its orientation and location through the vocal tract midline. A diphthong can be generated by interpolating these three vocal tract parameters, from the starting to the ending vowel. For simplicity, linear interpolation is adopted in this work. However, more realistic approaches could be considered instead, following for instance the non-linear behaviour of a biomechanical model of the vocal tract [29].

Figure 2 shows the dynamic vocal tract representations for diphthongs [ai] and [au] obtained from the AG and SPG cross-sections. The initial and final vowel vocal tracts are represented, and also two intermediate configurations resulting from linear interpolation. The dotted red line indicates the vocal tract midline. As observed, each cross-section is deformed over time to produce the desired diphthong. However, note that the SPG does not change the cross-section orientations, as opposed

to the AG. This obviously reduces the number of functions to interpolate, but above all helps to prevent the collision of adjacent cross-sections. Observe, for instance, that the cross-sections for [ai] are really close to each other in the mouth region. Although in the AG they do not come into contact with each other, an increase in the number of cross-sections would probably produce an overlap. In contrast, this would never occur in the oral region using the cross-sections from the SPG. As previously explained in Section II-A, these are not defined to be perpendicular to the vocal tract midline, but are aligned with a semi-polar grid, which has always vertical parallel segments in the mouth area regardless of the analysed vowel. In addition, and as a consequence, the cross-section orientation remains fixed over time.

The SPG and AG obviously provide different cross-sections, but as discussed in the Introduction, it is not clear whether they will affect the synthesis of diphthongs or not. To examine this issue, two dynamic vocal tract representations have been constructed using the two sets of cross-sections.

III. THE TIME-DOMAIN FINITE ELEMENT METHOD FOR VOWEL AND DIPHTHONG SYNTHESIS

A. Solving the wave equation in mixed form

The propagation of small acoustic perturbations in an homentropic flow can be described by means of the linearized continuity equation (mass conservation),

$$\frac{1}{\rho_0 c_0^2} \partial_t p + \nabla \cdot \mathbf{u} = 0, \quad (1)$$

and the linearized momentum conservation equation,

$$\rho_0 \partial_t \mathbf{u} + \nabla p = 0. \quad (2)$$

In the above expressions, the variables p and $\mathbf{u}$ respectively stand for acoustic pressure and the acoustic particle velocity, whereas the parameter ρ_0 designates the air density and c_0 the speed of sound. ∂_t denotes the partial derivative with respect to time.

Equations (1) and (2) are sometimes referred to as the wave equation in mixed form (see e.g., [33]). They can be easily combined to derive the standard irreducible wave equation for the acoustic pressure, namely $c_0^{-2} \partial_{tt}^2 p - \nabla^2 p = 0$, by subtracting the divergence of (2) from ρ_0 times the partial time derivative of (1). One can also infer analogous wave equations for the acoustic particle velocity and for the velocity potential. The irreducible wave equation is often used for the numerical generation of vowels in 3D vocal tract geometries [16]. However, that is no longer valid when one has to describe wave propagation in a time dependent domain, like a vocal tract distorting from the shape of one vowel to another, in the generation of a diphthong [25]. This is because the mass and momentum conservation equations (1) and (2) must be expressed in a frame of reference that moves with the dynamic domain, rather than in a fixed one.

Assume that each point in a dynamic vocal tract $\Omega(t)$ evolves according to a given velocity $\mathbf{u}_d(\mathbf{x}, t)$. One can then replace the time derivative ∂_t with $\partial_t - \mathbf{u}_d \cdot \nabla$ in (1) and (2), to account for the domain motion (this is known as the quasi-Eulerian Arbitrary Lagrangian-Eulerian (ALE) approximation,

see e.g., [34]). The wave equation in mixed form (1)-(2) then becomes transformed to

$$\frac{1}{\rho_0 c_0^2} [\partial_t p - \mathbf{u}_d \cdot \nabla p] + \nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega(t), \quad (3a)$$

$$\rho_0 [\partial_t \mathbf{u} - \mathbf{u}_d \cdot \nabla \mathbf{u}] + \nabla p = 0 \quad \text{in } \Omega(t), \quad (3b)$$

which is to be supplemented with appropriate boundary and initial conditions. For the simulations in this work we have considered the following ones,

$$\mathbf{u} \cdot \mathbf{n} = g(t) \quad \text{on } \Gamma_G(t), \quad t > 0, \quad (4a)$$

$$\mathbf{u} \cdot \mathbf{n} = p/Z_w \quad \text{on } \Gamma_W(t), \quad t > 0, \quad (4b)$$

$$p = 0 \quad \text{on } \Gamma_M(t), \quad t > 0, \quad (4c)$$

$$p = 0, \quad \mathbf{u} = 0 \quad \text{in } \Omega(t), \quad t = 0. \quad (4d)$$

Here, $\Gamma_G(t)$ represents the glottis cross-section where a glottal source inflow $g(t)$ is prescribed. Wall losses are considered by imposing an impedance Z_w at the vocal tract walls $\Gamma_W(t)$, while a zero pressure release condition is applied to the mouth exit, $\Gamma_M(t)$, to ease the computational cost when simulating diphthongs. Alternatively, radiation losses could be considered by extending the computational domain outside the mouth by means of a hemispherical volume [28], for example. A Perfectly Matched Layer (see e.g., [10], [16]) or a Sommerfeld non-reflection condition (see e.g., [28]) could then be applied to the hemisphere boundaries to absorb incident sound waves and avoid reflections.

The weak formulation of (3) can be found multiplying (3a) with a scalar test function q and (3b) with a vector test function $\mathbf{v}$, and then integrating over the computational domain $\Omega(t)$. This results in

$$\frac{1}{\rho_0 c_0^2} \left[\int_{\Omega} \partial_t p q \, d\Omega - \int_{\Omega} \mathbf{u}_d \cdot \nabla p q \, d\Omega \right] + \int_{\Omega} \nabla \cdot \mathbf{u} q \, d\Omega = 0, \quad (5a)$$

$$\rho_0 \left[\int_{\Omega} \partial_t \mathbf{u} \mathbf{v} \, d\Omega - \int_{\Omega} \mathbf{u}_d \cdot \nabla \mathbf{u} \mathbf{v} \, d\Omega \right] + \int_{\Omega} \nabla p \mathbf{v} \, d\Omega = 0, \quad (5b)$$

where we have not explicitly indicated the dependence of the computational domain with time to lighten the notation. The essential boundary conditions (4) for problem (5) are imposed strongly on the boundaries $\Gamma_G(t)$, $\Gamma_W(t)$ and $\Gamma_M(t)$ (see [25], [35] for alternatives).

The method of lines has been employed to solve the variational equations (5) so as to generate diphthongs and vowels. In the case of vowels we simply set $\mathbf{u}_d = 0$ and consider a static domain Ω . The method of lines consists in using an FEM approach for the spatial discretization of the variational equation and a finite difference scheme for the time discretization. In what concerns the former, it is well known that for mixed problems well-posedness (existence of a unique and stable solution) is not automatically transferred from the continuum formulation to the discrete one. As a consequence, not all finite element types are admissible, and to avoid implementing different order interpolation polynomials for the acoustic pressure and particle velocity it becomes necessary to resort to stabilized FEM approaches. Our option has been to apply the variational multiscale stabilization strategy in [25].

As for the time discretization of (5), a third order backward differentiation formula (BDF3) has been implemented and proved accurate enough. Numerical tests revealed the second order scheme BFD2 was too dissipative to properly cover the analyzed frequency range.

B. Motion of the vocal tract inner nodes

In addition to the numerical methodology described above, the solution of (5) requires knowing the evolution of the vocal tract $\Omega(t)$. This determines the velocity of every point in the computational domain, $\mathbf{u}_d(\mathbf{x}, t)$, appearing in (5). Note that the dynamic vocal tract representations in section II prescribes the movement of the boundary of $\Omega(t)$, say $\partial\Omega(t) := \Gamma_G(t) \cup \Gamma_W(t) \cup \Gamma_M(t)$, which, in turn, needs to be transmitted to the domain inner points (namely the nodes of the finite element mesh). Whenever the domain distortion is not very large, this can be done solving an elastic problem for the inner node displacements, $\mathbf{w}(\mathbf{x}, t)$, avoiding remeshing strategies (see e.g., [36]). If the transition from one vowel to another is smooth enough, an even simpler option exists: that of transmitting the motion by diffusion, solving the Laplacian equation $\nabla^2 \mathbf{w} = 0$.

Suppose that the simulation time interval $[0, T]$ gets discretized into N equally spaced time steps of constant length $\Delta t = t^{n+1} - t^n$, with t^n indicating the time instant at the n -th iteration and $\mathbf{w}^n$ denoting the displacement evaluated at t^n . To account for the domain motion, at each time step $t = t^{n+1}$ of the simulation we will also use FEM to solve, along with (5), the auxiliary weak problem,

$$\int_{\Omega} \nabla \mathbf{w}^{n+1} \cdot \nabla \mathbf{v} = 0, \quad (6a)$$

with boundary conditions

$$\mathbf{w}^{n+1} = \mathbf{x}_{\text{walls}}^{n+1} - \mathbf{x}_{\text{walls}}^n \quad \text{on } \Gamma_W(t^{n+1}), \quad (6b)$$

$$\mathbf{w}^{n+1} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_G(t^{n+1}), \quad (6c)$$

$$\mathbf{w}^{n+1} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_M(t^{n+1}). \quad (6d)$$

In (6a), $\mathbf{x}_{\text{walls}}^{n+1}$ and $\mathbf{x}_{\text{walls}}^n$ indicate the positions of the nodes located on the vocal tract walls at times t^{n+1} and t^n . The positions of the nodes at t^{n+1} can then be updated as

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \mathbf{w}^{n+1}, \quad (7)$$

and the velocity $\mathbf{u}_d$ at a node $\mathbf{x}_i$ in (3), computed from

$$\mathbf{u}_d^{n+1}(\mathbf{x}_i) = \frac{\mathbf{x}_i^{n+1} - \mathbf{x}_i^n}{\Delta t}. \quad (8)$$

C. Simulation details

The simulation of a diphthong requires an initial finite element mesh which gets deformed in time using the dynamic vocal tract geometries in Section II. This mesh could be obtained by discretizing the vocal tract geometry of the starting vowel of the diphthong into a set of tetrahedrons. Unfortunately, large element deformations can be produced when transitioning from the initial mesh to that of the final vowel. Some elements can even be inverted, resulting in tetrahedrons with negative volume. This problem could be prevented by

remeshing the vocal tract geometry every certain time steps, and interpolating the solution (acoustic pressure and velocity) from the old mesh to the new one. Nonetheless, this procedure is computationally expensive and difficult to implement. One can avoid by starting from a mesh that corresponds to an intermediate articulation between the first and second vowels of the diphthong.

That was the option taken in this work. Tetrahedral elements of size 0.003 m were assigned to the initial mesh, which was then distorted to reach the articulation of the first vowel sound, and subsequently evolved to the second one to complete the diphthong. To do so, the dynamic vocal tract representations described in Section II were used to derive the node coordinates of the vocal tract walls, $\mathbf{x}_{\text{walls}}$, at each time step (every 1 ms). The Laplacian equation (6) was then solved using FEM to obtain the corresponding inner node coordinates. The acoustic wave propagation was simulated solving, also with FEM, the ALE mixed wave equation (3) supplemented with the boundary conditions in (4). A train of glottal pulses constituted $g(t)$ in Eq. (4a). Those were generated using an LF model [37] enhanced with a pitch curve, a fade in/out, and some jitter and shimmer, as in [13]. The wall impedance Z_w in Eq. (4b) was set to $Z_w = 83666 \text{ kg/m}^2\text{s}$ [14], and the speed of sound and air density were $c_0 = 350 \text{ m/s}$ and $\rho_0 = 1.1644 \text{ kg/m}^3$. The resolution of Eq. (3) also requires a time discretization. A sampling frequency of $f_s = 250 \text{ kHz}$ was chosen, colored which ensured a good stability with the selected time scheme (BDF3), mentioned before. The acoustic pressure was tracked at a node close to the vocal tract exit (0.003 m from the mouth) to obtain the synthesized diphthong. Simulations lasted about 6 hours on average to generate a diphthong 0.5 s long. The resolution of the ALE mixed wave equation required the vast majority of resources (> 95%), much more than moving the inner nodes or computing the trajectory of the exterior ones. A serial desktop computing system with processor Intel(R) Core(TM) i7-6700 3.4 GHz was used for all computations.

In addition, vocal tract transfer functions of (static) vowels were also computed. In that case, the train of glottal pulses was replaced by the following Gaussian pulse

$$gp(t) = e^{-[(t-T_{gp})/(0.29T_{gp})]^2} [\text{m/s}], \quad (9)$$

with $T_{gp} = 0.646/f_0$ and $f_0 = 10 \text{ kHz}$. This pulse was low pass filtered at the maximum frequency of analysis, 10 kHz, to avoid the appearance of numerical errors above this frequency. Only the ALE mixed wave equation was solved in this configuration, setting $\mathbf{u}_d = 0$ in Eq. (3) given that the computational domain (vocal tract) remains static for a vowel.

The vocal tract transfer functions

$$H(f) = P_o(f)/Q_i(f) \quad (10)$$

were computed from 50 ms simulations, with $P_o(f)$ standing for the Fourier transform of the acoustic pressure at the vocal tract exit (0.003 m from the mouth), and $Q_i(f)$ for the Fourier transform of the input volume velocity inserted at the glottal cross-section (i.e., the Gaussian pulse $gp(t)$ multiplied by the glottal cross-sectional area).

Fig. 3. Vocal tract transfer functions $H(f)$ for vowels [a], [i] and [u] obtained from 3D FEM simulations and with a 1D acoustic model based on transfer matrices. An adaptive grid (AG) and a semi-polar grid (SPG) are used to resample the corresponding 3D MRI-based vocal tract geometries and to obtain 1D area functions.

IV. RESULTS

The production of vowels [a], [i] and [u] was first studied to evaluate, in a static configuration, the performance of the dynamic vocal tracts based on the adaptive grid (AG) and the semi-polar grid (SPG). The diphthongs [ai] and [au] were next synthesized as examples of application using the two representations. For completeness, results are also compared with a standard 1D acoustic model based on the transfer matrix method (TMM). Details on the implementation can be found in the Appendix A. This 1D model was also driven by the same two dynamic vocal tract representations, but only the area of each cross-section was maintained (i.e., vocal tract area function). The vocal tract area functions for the vowels [a], [i] and [u] obtained with the AG and SPG are reported in Table I of Appendix B. In what follows, the dynamic vocal tract representations based on the SPG and AG will simply be termed as SPG and AG representations, respectively.

A. Vowels

Figure 3 shows the vocal tract transfer functions $H(f)$ of vowels [a], [i] and [u] computed in 3D and 1D with the SPG and AG vocal tract representations.

Let us first compare the AG and SPG approaches in 3D. The vocal tract transfer functions do not present significant differences between the two geometry representations. In the low frequency range, below 5 kHz, all vowels exhibit a similar behavior regardless of the grid. Somewhat larger deviations can be appreciated above 5 kHz. For instance, the curves in vowel [i] slightly depart between 9 kHz and 10 kHz, while

Fig. 4. Snapshots of the FEM simulation of diphthongs [ai] and [au], respectively generated with the adaptive grid (AG) and semi-polar grid (SPG) for illustrative purposes. The voice signal is also represented, with red dots indicating the time instants of the snapshots (10, 65, 115 and 185 ms).

the antiresonance of [a] at 6 kHz (generated by a transverse mode, see [7]) is not as prominent for the SPG as for the AG. Nonetheless, the vocal tract transfer functions can be considered very similar in the examined frequency range. Therefore, either the AG or SPG approaches can be used with confidence to generate accurate vowels in 3D.

As opposed to 3D, noticeable differences are found between the two vocal tract representations in 1D. For instance, below 3 kHz, the second formant of vowels [a] and [i] slightly modifies its location. Between 3 kHz and 5 kHz, most formants display large shifts in their frequencies for all vowels, whereas, beyond 5 kHz, strong disparities occur, see e.g. the 7th formant between 6 kHz and 7 kHz of vowel [i].

Finally, it is worth mentioning that the results reported here confirm that 1D models can only correctly emulate the 3D ones in the low frequency range, where plane wave propagation dominates. For vowel [a] and below 3 kHz, the 3D and 1D solutions are quite similar, with only some small deviations in the formant levels and in their frequencies. However, beyond 3 kHz, higher order modes also propagate which strongly alter the 3D vocal tract transfer functions as compared to the 1D ones, especially above 5 kHz. For vowel [i], the vocal tract transfer functions are similar up to 5 kHz, whereas for vowel [u] the similitude extends up to 6 kHz. Only some slight shifts can be observed in the low frequency range for the second formant of vowel [i] and for the fourth of vowel [u]. There is a general agreement in literature that 1D results are accurate enough up to about 5 kHz, but as just exemplified above, this frequency limit depends on each vowel. As can be appreciated in Figure 2, vowel [a] has the largest oral cavity.

Fig. 5. Spectrograms of the diphthongs (a) [ai] and (b) [au] generated with the 3D FEM and the 1D TMM acoustic models using an adaptive grid (AG) and a semi-polar grid (SPG), and differences between the AG approach with respect to the SPG. Values are expressed in dB.

This reduces the cutoff frequency at which higher order modes begin to propagate, which in turn diminishes the accuracy of a 1D model, if compared to other vowel sounds.

B. Diphthongs

The diphthongs [ai] and [au] have been synthesized as examples of application of the AG and SPG dynamic vocal tract representations. To do so, they were configured to sustain the articulation of the starting vowel for 15 ms, then to linearly transition to the ending vowel in 150 ms, and finally to sustain the ending vowel for 35 ms, which results in a total duration of the diphthong of 200 ms. Figure 4 shows some snapshots captured during the FEM simulations of the two diphthongs. In each frame the acoustic pressure distribution and the finite element mesh can be observed on the vocal tract surface. The limiting values of the color scale were set to match the maximum and minimum pressure values at each time instant, so as to enhance the visualization of acoustic waves. The snapshots were acquired at $t = 10, 65, 115$ and 185 ms; the first time instant corresponding to the starting vowel, the next two to the transition, and the last one to the ending vowel. Since the transition takes place in the time interval $t = [15, 165]$ ms, we note that $t = 65$ ms and $t = 115$ ms correspond to two time instants equally distributed within the interval (increments of 50 ms). The starting and ending time instants were simply selected to better illustrate the pressure distribution pattern in the vocal tract. The resulting voice signal is also represented in the figure, with red dots indicating the above time instants.

We can observe in the figure how the vocal tract walls deform from the starting to the ending vowel to generate a diphthong. As explained in Section III, the AG and SPG representations drive the movement of the mesh nodes located on the walls, while the nodes within the vocal tract move according to a diffusion process. The acoustic wave propagation in the deformed 3D meshes is determined by solving the ALE mixed wave equation. Note that the pressure distribution pattern mainly changes through the propagation axis. This is

so because the energy of the excitation signal (glottal pulses) decreases with frequency, thus providing more energy to the planar modes than to the higher order ones. However, this does not prevent higher order modes from propagating.

Let us next analyze the spectrograms of the synthesized diphthongs. A pre-emphasis FIR filter with coefficients $[1 - 0.97]$ was applied in their computation to enhance the visualization of higher frequencies, as typically done in speech analysis. Figure 5 shows the spectrograms for the 3D FEM simulations and also for the 1D TMM approach, using the two dynamic vocal tract representations. Moreover, the figure also presents the difference spectrogram between AG and SPG, for each diphthong and acoustic model.

Focusing first on the 3D solutions, we observe that the starting and ending vowel sounds are very similar for the two vocal tract representations (look at $t < 15$ ms and $t > 165$ ms in Figure 5). Only some small differences appear for the high frequency range, as already perceived in the $H(f)$ of Figure 3. However, this is not the case for the transition ($t = [15, 165]$ ms). Some differences manifest themselves, especially for the diphthong [ai]. Hence, the dynamic representations preserve the starting and ending vocal tract geometry, whereas they do not exactly transition in the same way. Indeed, a linear interpolation of the area, location and orientation of each cross-section does not guarantee the same deformation of the vocal tract walls, since they are not obtained in the same way in the SPG and AG approaches. Despite this, one could consider the formant trajectories to be quite similar, and use any of the two vocal tract representations with confidence in 3D.

As regards the 1D approach, the largest deviations between AG and SPG occur during the transition stage of the diphthongs, as in 3D, and also mainly affect the high frequency content. They are also more important for [ai] than for [au]. However, note that the deviations are also significant in the low frequency region of the starting and ending vowels, especially for [ai]. As observed in the $H(f)$ of Figure 3, the 1D solution is more sensitive to changes in the vocal tract representation

than the 3D ones (see e.g. the second formant of [i]). On the other hand, and compared to 3D, the differences between SPG and AG are larger regardless of the analyzed segment of the diphthong. As hypothesized in the Introduction, this can be explained because acoustic waves can freely travel in 3D within the vocal tract, whereas 1D needs to use the area of each cross-section. For 1D, the selection of a proper dynamic vocal tract seems to be more important than for 3D.

Finally, it is also interesting to examine the differences obtained in 3D against 1D for the two dynamic vocal tracts (compare the first row with the second one in Figure 5). As expected from the analysis on static vowels (see Section IV-A), substantial differences appear, mainly in the high frequency range, because of the propagation of higher order modes. However, note that some small deviations also appear for low frequencies, where plane wave propagation dominates. Those can be attributed to the curvature of acoustic front waves when encountering a sudden expansion or contraction in the vocal tract. It is known that when this happens, 1D models fail in estimating the formant frequencies. Indeed, this phenomenon can be understood as an artificial lengthening of the differential duct where the discontinuity occurs (see e.g., [38]). 1D simulations can correct these deviations by introducing inner length corrections [39]. Note that this is not necessary in 3D because the phenomenon occurs naturally. On the other hand, no remarkable difference can be appreciated when comparing the 3D AG and SPG approaches with their 1D counterparts. The two vocal tract representations do not significantly contribute to the deviations introduced when moving from 3D to 1D.

V. CONCLUSIONS

In this work, two dynamic vocal tract representations based on cross-sections extracted from static MRI-based vocal tracts have been developed for the synthesis of vowels and diphthongs in 3D. These two options, namely the SPG and the AG, have successfully been connected to a 3D acoustic model based on FEM in the time domain. These representations have shown to accurately describe the 3D vocal tract acoustics of vowels [a], [i] and [u], but above all, they have allowed us to generate diphthongs by simply interpolating the location, orientation and shape of the cross-sections of initial and ending vowels. No significant differences were observed between the SPG and AG representations when synthesizing the diphthongs [ai] and [au]. However, the SPG dynamic representation is favored because it helps to prevent the collision of adjacent cross-sections as their orientation is fixed over time. A 1D acoustic model derived from a TMM approach was also employed for comparison. As opposed to 3D, the 1D solution proved to be more sensitive to the vocal tract shape, either for static and dynamic vowel sounds. To end with, we note that the presented study is currently limited to diphthongs containing the sounds [a], [i] and [u], but it could be easily extended to other vowels using the corresponding MRI-based vocal tract geometries and the methodology presented herein. In future work we will consider the use of remeshing strategies to deform the finite element meshes since they facilitate the

automatization of the simulations, though at the price of increasing the computational cost. Moreover, radiation losses will be taken into account to achieve more natural sounds.

APPENDIX A

THE TRANSFER MATRIX METHOD FOR DIPHTHONG SYNTHESIS IN 1D

A Transfer Matrix Method (TMM) similar to that presented in [1] has been used in this work to generate diphthongs in 1D. This is a frequency domain method that relates the input and output acoustic pressures P and volume velocities Q in a system by a transfer matrix $\mathbf{K}$,

$$\begin{pmatrix} P_{\text{out}} \\ Q_{\text{out}} \end{pmatrix} = \mathbf{K} \begin{pmatrix} P_{\text{in}} \\ Q_{\text{in}} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} P_{\text{in}} \\ Q_{\text{in}} \end{pmatrix}. \quad (11)$$

In the case of a vocal tract, the input is the glottis and the output the lips. Let us first discretize the vocal tract geometry into a set of N concatenated tubes of constant cross-section S_i and length l_i , with $i = 1, 2, \dots, N$ and $i = N$ denoting the last tube at the mouth. Assuming plane wave propagation, the output and input acoustic pressure and volume velocities in each tube can be related as

$$\begin{pmatrix} P_i \\ Q_i \end{pmatrix} = \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix} \begin{pmatrix} P_{i-1} \\ Q_{i-1} \end{pmatrix} = \mathbf{K}_i \begin{pmatrix} P_{i-1} \\ Q_{i-1} \end{pmatrix}, \quad (12)$$

with

$$\begin{aligned} A_i &= \cos(k_i l_i), & B_i &= -jZ_0/S_i \sin(k_i l_i), \\ C_i &= -jS_i/Z_0 \sin(k_i l_i), & D_i &= \cos(k_i l_i). \end{aligned} \quad (13)$$

In Equation (13), $Z_0 = \rho_0 c_0$ is the characteristic impedance of air, and k_i is the wave number within the i -th tube. For a lossless vocal tract, $k_i = k_0 \equiv 2\pi f/c_0$, with f being the frequency in Hz. In this case, however, we consider a complex wave number that introduces the same walls losses in 1D than the 3D acoustic model of Section III. Let us introduce the boundary admittance coefficient μ which is related to the wall impedance Z_w as $\mu = \rho_0 c_0/Z_w$. It can be shown that k_i can be computed from μ as [16]

$$k_i \simeq k_0 \sqrt{1 - j \frac{\mu \mathcal{P}_i}{k_0 S_i}} = k_0 \sqrt{1 - j \frac{2\mu}{k_0 a_i}}, \quad (14)$$

where $\mathcal{P}_i$ and a_i are the perimeter and radius of the i -th tube and $j = \sqrt{-1}$.

The matrix $\mathbf{K}$ can finally be obtained as the concatenation of all tubes,

$$\mathbf{K} = \prod_{i=N}^1 \mathbf{K}_i. \quad (15)$$

As said, $\mathbf{K}$ links the acoustic pressure and volume velocity from the glottis to the mouth. However, in 3D simulations the acoustic pressure was tracked within the vocal tract at a distance of 0.003 m from the mouth exit. To consider that, $\mathbf{K}$ is split in a first matrix $\mathbf{K}^a$ that relates the glottis with the tracking point, and a second matrix $\mathbf{K}^b$ for the tracking point to the mouth. The input impedance Z_{in}^b is first computed from the coefficients of $\mathbf{K}^b$ as

$$Z_{in}^b(f) = \frac{D^b Z_r^b - B^b}{A^b - C^b Z_r^b}, \quad (16)$$

TABLE I
VOCAL TRACT AREA FUNCTIONS $A(x)$ FOR VOWELS [a], [i] AND [u] EXTRACTED WITH THE ADAPTIVE GRID (AG) AND THE SEMI-POLAR GRID (SPG), WITH x DENOTING THE DISTANCE FROM THE GLOTTAL CROSS-SECTION. UNITS ARE IN cm AND cm^2 FOR x AND $A(x)$, RESPECTIVELY.

[a]				[i]				[u]			
AG		SPG		AG		SPG		AG		SPG	
x	$A(x)$	x	$A(x)$	x	$A(x)$	x	$A(x)$	x	$A(x)$	x	$A(x)$
0	0.2	0	0.2	0	0.3	0	0.3	0	0.15	0	0.15
0.63	0.5	0.59	0.55	0.62	0.53	0.62	0.56	0.67	0.22	0.65	0.26
1.29	0.3	1.19	0.32	1.26	0.28	1.23	0.28	1.36	0.05	1.27	0.06
1.92	0.36	1.85	0.41	1.89	0.38	1.88	0.42	2.02	0.14	1.94	0.14
2.62	2.67	2.68	2.66	2.72	2.87	2.72	3.02	2.72	1.65	2.77	1.83
3.09	2.96	3.11	2.83	3.07	3.55	3.09	3.49	3.19	2.25	3.17	2.16
3.53	2.73	3.54	2.67	3.52	3.88	3.46	3.86	3.68	2.28	3.57	2.27
3.95	2.38	3.98	2.25	3.94	3.98	3.82	3.99	4.15	2.1	3.97	2.21
4.38	1.95	4.41	1.81	4.36	3.93	4.2	3.93	4.63	1.71	4.37	1.86
4.81	1.59	4.84	1.65	4.79	3.73	4.57	3.73	5.1	1.75	4.77	1.62
5.24	2.12	5.27	2.16	5.21	3.66	4.94	3.56	5.67	4.06	5.19	2.17
5.65	2.11	5.7	2	5.77	6.53	5.51	6.43	6.14	4.51	5.75	4.68
6.07	1.8	6.14	1.62	6.18	6.64	5.88	6.49	6.61	3.95	6.15	4.46
6.5	1.47	6.57	1.37	6.6	6.82	6.24	6.64	7.08	3.19	6.56	4.02
6.92	1.28	6.99	1.14	7.02	6.97	6.61	6.81	7.56	2.62	6.96	3.39
7.35	1.04	7.4	1.08	7.44	7.27	7.14	7.08	8.04	2.18	7.44	2.8
7.78	1.17	7.82	1.25	7.86	7.51	7.68	7.48	8.55	1.7	7.94	2.38
8.22	1.37	8.26	1.5	8.3	7.46	8.23	7.52	9.08	1.54	8.49	1.86
8.68	1.65	8.71	1.9	8.75	7.5	8.83	7.75	9.63	1.8	9.08	1.65
9.15	2.32	9.21	2.72	9.24	7.6	9.43	7	10.19	1.82	9.7	1.85
9.64	2.76	9.68	2.53	9.71	6.46	10.03	5.28	10.77	1.56	10.3	1.81
10.13	2.53	10.15	2.65	10.19	5.31	10.63	4.68	11.36	1.15	10.91	1.4
10.6	2.89	10.62	2.52	10.66	4.62	11.22	3.35	11.97	1.31	11.52	1.26
11.09	2.31	11.1	2.53	11.11	4.22	11.8	1.67	12.54	0.77	12.14	1.2
11.57	2.77	11.59	3.04	11.56	2.78	12.38	0.64	13.1	0.56	12.75	0.66
12.05	3.42	12.07	3.69	12.01	1.59	12.98	0.33	13.65	0.91	13.34	0.64
12.51	4.14	12.47	4.47	12.44	0.76	13.28	0.35	14.16	1.77	13.83	1.32
13.03	7.17	13.04	8.14	12.86	0.38	13.56	0.35	14.66	2.94	14.6	3.33
13.46	9.19	13.45	10.11	13.28	0.32	13.83	0.36	15.19	5.13	15.1	5.15
13.88	11.81	13.84	11.89	13.72	0.35	14.11	0.39	15.66	6.22	15.62	7.51
14.3	12.63	14.24	13	14.17	0.37	14.39	0.47	16.14	7.25	16.17	9.11
14.73	12.42	14.66	13.22	14.63	0.47	14.68	0.56	16.62	7.44	16.66	8.77
15.16	13.09	15.07	13.78	15.08	0.47	15	0.59	17.11	7.37	17.13	7.62
15.58	12.71	15.56	15.11	15.52	0.54	15.37	0.72	17.6	6.13	17.57	5.71
16	12.25	16.05	15.72	15.96	0.51	15.9	1.13	18.1	3.85	18.02	4.05
16.53	13.73	16.46	14.39	16.39	0.6	16.44	1.58	18.56	2.33	18.47	2.57
16.99	12.65	16.86	12.37	16.83	1	16.99	1.99	19.03	1.09	18.96	1.22
17.49	9.61	17.26	9.49	17.25	1.29	17.4	2.32	19.49	0.42	19.42	0.44
17.94	6.52	17.68	6.4	17.67	2.4	17.69	2.66	19.96	0.15	19.88	0.17
18.38	4.81	18.09	4.81	18.08	3.06	17.96	3.06	20.42	0.31	20.38	0.31

with Z_r^b being the radiation impedance, which is set to $Z_r^b = 0$ to match again with the the 3D acoustic model ($p = 0$ at the mouth). However, radiation losses could be easily introduced by using the analytical expressions of the piston set on a sphere [40] or on a flat infinite baffle [41], [42]. One should simply bear in mind that these models typically provide the specific acoustic impedance in Pa s/m normalized with respect to Z_0 , namely Z_r' . The acoustic impedance Z_r (Pa s/m³) can be recovered as $Z_r = Z_0 Z_r' / S_N$, with S_N standing for the mouth cross-sectional area. The vocal tract transfer function is next computed from the coefficients of $\mathbf{K}^a$ as

$$H(f) = \frac{Z_r^a}{A^a - C^a Z_r^a}, \quad (17)$$

but in this case $Z_r^a = Z_{in}^b$, as the load now is the input impedance of the second part of the vocal tract.

In order to generate a diphthong, a set of $H(f)$ are first computed as the vocal tract moves. These are then inverse Fourier transformed to obtain the corresponding impulse responses. The acoustic pressure evolution $p(t)$ of a diphthong is finally calculated as the convolution between the train of glottal pulses

$g(t)$ (m³/s) and the time-varying impulse response of the vocal tract $h(\tau, t)$,

$$p(t) = \int g(t - \tau) h(\tau, t) d\tau. \quad (18)$$

APPENDIX B VOCAL TRACT AREA FUNCTIONS

Table I contains the area functions $A(x)$ extracted with the adaptive grid (AG) and the semi-polar grid (SPG) from the MRI-based vocal tract geometries of vowels [a], [i] and [u] in [20]. They have been used as input for the 1D synthesiser described in Appendix A. The distance from the glottis is denoted with x , so that $x = 0$ (first row) corresponds to the glottal cross-section and the last row of the table to the mouth. The values in x have been obtained by computing the Euclidian distance between the centers of each cross-section. Sagittal variations of the cross-section centers have not been considered in the computation of x to avoid an artificial lengthening of the vocal tract, as done in [5] and [7].

ACKNOWLEDGMENTS

The authors would like to acknowledge David Vila, who helped in the development of the 1D synthesiser during his BSc final project (supervised by Dr. Marc Arnela). This research has been supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The third author would like to thank l'Obra Social de la Caixa and the Universitat Ramon Llull for their support under grant 2018-URL-IR2nQ-031.

REFERENCES

- [1] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Audio Speech Lang. Process.*, vol. 35, no. 7, pp. 955–967, 1987.
- [2] K. v. d. Doel and U. Ascher, "Real-time numerical solution of Webster's equation on a nonuniform grid," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 6, pp. 1163–1172, 2008.
- [3] 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.
- [4] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603. doi:10.1371/journal.pone.0060603, 2013.
- [5] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, 1996.
- [6] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. Van Hirtum, and X. Labal, "Effects of higher order propagation modes in vocal tract like geometries," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 832–843, 2015.
- [7] 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," *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 1707–1718, 2016.
- [8] B. B. Monson, A. J. Lotto, and S. Ternström, "Detection of high-frequency energy changes in sustained vowels produced by singers," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 2263–2268, 2011.
- [9] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, p. 587, 2014.
- [10] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3724–3738, 2010.
- [11] 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," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2852–2859, 2016.
- [12] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 577–585, 2007.
- [13] M. Arnela and O. Guasch, "Finite element synthesis of diphthongs using tuned two-dimensional vocal tracts," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 2013–2023, 2017.
- [14] P. Švancara and J. Horáček, "Numerical modelling of effect of tonsillectomy on production of Czech vowels," *Acta Acust. united with Acustica*, vol. 92, no. 5, pp. 681–688, 2006.
- [15] 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.
- [16] M. Arnela and O. Guasch, "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method," *J. Acoust. Soc. Am.*, vol. 133, no. 6, pp. 4197–4209, 2013.
- [17] M. Arnela, O. Guasch, and F. Alías, "Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2946–2954, 2013.
- [18] K. Motoki, "Three-dimensional acoustic field in vocal-tract," *Acoust. Sci. & Tech.*, vol. 23, no. 4, pp. 207–212, 2002.
- [19] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/, *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4466–4481, 2008.
- [20] 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. T. Soukka, and M. Vainio, "Large scale data acquisition of simultaneous MRI and speech," *Appl. Acoust.*, vol. 83, pp. 64–75, 2014.
- [21] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomech. Model. Mechanobiol.*, pp. 1–15, 2014.
- [22] R. Blandin, A. Van Hirtum, X. Pelorson, and R. Laboissière, "Influence of higher order acoustical propagation modes on variable section waveguide directivity: Application to vowel [u]," *Acta Acust. united with Acustica*, vol. 102, no. 5, pp. 918–929, 2016.
- [23] M. Speed, D. T. Murphy, and D. M. Howard, "Three-dimensional digital waveguide mesh simulation of cylindrical vocal tract analogs," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 2, pp. 449–455, 2013.
- [24] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3d digital waveguide mesh," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 2, pp. 453–464, 2014.
- [25] 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 Acust. united with Acustica*, vol. 102, no. 1, pp. 94–106, 2016.
- [26] P. Anderson, S. Fels, N. M. Harandi, A. Ho, S. Moisik, C. A. Sánchez, I. Stavness, and K. Tang, "Chapter 20 - FRANK: A Hybrid 3D Biomechanical Model of the Head and Neck," in *Biomechanics of Living Organs*, Y. Payan and J. Ohayon, Eds. Oxford: Academic Press, 2017, pp. 413–447.
- [27] J. E. Lloyd, I. Stavness, and S. Fels, "Artisynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," in *Soft tissue biomechanical modeling for computer assisted surgery*. Springer, 2012, pp. 355–394.
- [28] S. Dabbaghchian, M. Arnela, O. Engwall, and O. Guasch, "Reconstruction of vocal tract geometries from biomechanical simulations," *Int. J. Numer. Method. Biomed. Eng.*, vol. 35, no. 2, p. e3159, 2019.
- [29] —, "Synthesis of VV utterances from muscle activation to sound with a 3D model," in *Proc. of INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 3497–3501.
- [30] A. J. Gully, I. Daffern, and D. T. Murphy, "Diphthong synthesis using the dynamic 3d digital waveguide mesh," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 2, pp. 243–255, 2018.
- [31] M. Arnela, S. Dabbaghchian, O. Guasch, and O. Engwall, "Finite element generation of vowel sounds using dynamic complex three-dimensional vocal tracts," in *Proc. of 23th International Congress on Sound and Vibration (ICSV23)*, Athens, Greece, July 2016.
- [32] —, "A semi-polar grid strategy for the three-dimensional finite element simulation of vowel-vowel sequences," in *Proc. of INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 3477–3481.
- [33] R. Codina, "Finite element approximation of the hyperbolic wave equation in mixed form," *Comput. Methods Appl. Mech. Engrg.*, vol. 197, no. 13–16, pp. 1305–1322, 2008.
- [34] J. Donea, A. Huerta, J.-P. Ponthot, and A. Rodríguez-Ferran, *Arbitrary Lagrangian Eulerian Methods*, in: *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Ltd, 2004.
- [35] S. Badia, R. Codina, and H. Espinoza, "Stability, convergence and accuracy of stabilized finite elements methods for the wave equation in mixed form," *SIAM J. Numer. Anal.*, vol. 52, no. 4, pp. 1729–1752, 2014.
- [36] G. Chiandussi, G. Bugeđa, and E. Oñate, "A simple method for automatic update of finite element meshes," *Commun. Numer. Meth. Engrg.*, vol. 16, no. 1, pp. 1–19, 2000.
- [37] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 77, no. 4, pp. 1–13, 1985.
- [38] Z. Kang and Z. Ji, "Acoustic length correction of duct extension into a cylindrical chamber," *J. Sound Vib.*, vol. 310, no. 4–5, pp. 782–791, 2008.
- [39] M. M. Sondhi, "An improved vocal tract model," in *Proc. of the 11th ICA*, Paris, France, 1983, pp. 167–170.
- [40] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. McGraw-Hill, New York, 1968.
- [41] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. Wiley, New York, Dec. 2000.
- [42] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 3rd ed. Springer-Verlag, New York, 2008.