ARTICLE TYPE

Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model

Saeed Dabbaghchian¹ | Marc Arnela² | Olov Engwall¹ | Oriol Guasch²¹Department of Speech, Music, and Hearing,
KTH Royal Institute of Technology,
Stockholm, Sweden²GTM Grup de recerca en Tecnologies
Mèdia, La Salle - Universitat Ramon Llull,
Barcelona, Catalonia (Spain)**Correspondence**Marc Arnela, Email:
marc.arnela@salle.url.edu**Present Address**C/ Quatre Camins 30, 08022 Barcelona,
Catalonia (Spain).**Summary**

A link is established between biomechanical and acoustic 3D models for the numerical simulation of vowel-vowel utterances. The former rely on the activation and contraction of relevant muscles for voice production, which displace and distort speech organs. However, biomechanical models do not provide a closed computational domain of the 3D vocal tract airway where to simulate sound wave propagation. An algorithm is thus proposed to extract the vocal tract boundary from the surrounding anatomical structures at each time step of the transition between vowels. The resulting 3D geometries are fed into a 3D finite element acoustic model that solves the mixed wave equation for the acoustic pressure and particle velocity. An arbitrary Lagrangian-Eulerian framework is considered to account for the evolving vocal tract. Examples include six static vowels and three dynamic vowel-vowel utterances. Plausible muscle activation patterns are first determined for the static vowel sounds following an inverse method. Dynamic utterances are then generated by linearly interpolating the muscle activation of the static vowels. Results exhibit nonlinear trajectory of the vocal tract geometry, similar to that observed in electromagnetic midsagittal articulography. Clear differences are appreciated when comparing the generated sound with that obtained from direct linear interpolation of the vocal tract geometry. That is, interpolation between the starting and ending vocal tract geometries of an utterance, without resorting to any biomechanical model.

KEYWORDS:

Vowel-vowel utterances, biomechanical model, acoustic model, voice production, Finite Element Method

1 | INTRODUCTION

Voice is generated by the propagation of sound waves through a vocal tract that changes its shape over time. **This phenomenon can be simulated using a three-dimensional (3D) physics-based model. First, we need a 3D representation of the vocal tract airway. A widely used option consists in using Magnetic Resonance Imaging (MRI) to obtain detailed 3D geometries of the vocal**

tract. These can next be simplified as 1D area functions^{1,2}, which describe the vocal tract area variation along its center midline. However, new MRI sessions are required every time we want to modify the vocal tract shape. Moreover, long acquisition times are needed to acquire 3D geometries with a high spatial resolution, which limits the studies to static sounds. Interpolation of static geometries is usually applied in these cases to obtain dynamic vocal tracts. For instance, this strategy has been followed to interpolate 1D area functions³, 2D representations of the vocal tract^{4,5,6,7}, simplified 3D geometries⁸, realistic 3D geometries^{9,10,2}, and voxels of 3D images¹¹. Alternatively, we can use articulatory and biomechanical models to obtain all vocal tract geometries we like without performing MRI scans. Articulatory models are controlled by a set of high level phonetic descriptors, like the tongue height, and can only replicate the kinematics of the speech organs or articulators. Conversely, in a biomechanical model speech dynamics is reproduced implicitly by mechanical properties of the articulators. In addition to this advantage, the model is controlled by contraction of muscles, which makes it useful for neuromuscular studies^{12,13,14}. Early biomechanical models were developed in 2D, typically in the midsagittal plane^{15,16,17}. However, 3D models have also been developed more recently^{18,19,20,21}, providing a more truthful representation of the anatomy, especially for the tongue.

To generate a sound, the biomechanical model must be coupled with an acoustic model. This requires an explicit representation of the vocal tract, that is, a 3D vocal tract geometry or a 1D area function. 2D biomechanical models have typically used the $\alpha\beta$ method^{22,23} to approximate the area function from the midsagittal dimensions. In contrast, 3D models can be more accurate and provide a detailed representation of the vocal tract. However, it is common to use the 1D area function, determined from the 3D model, for acoustic simulations^{20,24}. The reason is that using the 3D vocal tract geometry instead of its 1D area function introduces several challenges. Firstly, the 3D vocal tract geometry must be reconstructed at each time step from the surrounding anatomical structures of the biomechanical model. Secondly, the acoustic model must use the above vocal tract geometries to simulate the propagation of sound waves in a 3D computational domain that has moving boundaries. The reward for such difficulties is that 3D modelling improves the accuracy of the produced sound, if compared to 1D. It has been shown that 1D plane wave modeling, commonly implemented with 1D area functions, not only affects the formant bandwidths and frequencies below 4-5 kHz¹, but neither accounts for the propagation of higher order modes beyond this frequency, which may be relevant for naturalness. In addition to evidence about the importance for high frequency bands²⁵, the simulation of wave propagation in a 3D domain provides a unified approach. That is, in traditional 1D strategies, the acoustic model needs to be adapted to deal with different situations such as large area discontinuities²⁶, or the inclusion of the piriform fossa²⁷ or the nasal tract²⁸. With 3D approaches, these phenomena are addressed naturally.

To the best of the authors' knowledge, the development of a coupled 3D dynamic biomechanical-acoustic model **that fully works in 3D** is unprecedented, despite previous separate efforts in 3D biomechanical and 3D acoustic modeling. **In literature, we can find 2D biomechanical models combined with 1D acoustic models¹⁶, 3D biomechanical models with 1D acoustic models^{20,24}, but not a 3D biomechanical model with a 3D acoustic model.** An exception, is the previous work by the authors of this study in²⁹, which was able to generate accurate vocal tract geometries from a biomechanical model. However, that model was limited to the simulation of static vowel sounds because, on the one hand, reconstructed vocal tract geometries were not suitable for dynamic acoustic simulations and, on the other hand, the acoustic model could not take into account a computational domain with moving boundaries. In this work, we propose a methodology to fully couple a 3D dynamic biomechanical model with a 3D dynamic acoustic model so as to deal with vowel-vowel utterances. In particular, we link an adapted version of the Artisynt FRANK biomechanical model²¹ with the 3D FEM-based acoustic model developed in⁸ for the simulation of diphthongs in the time domain. An algorithm is developed to extract, at each time step, 3D vocal tract geometries from the biomechanical model, which are next used by the 3D acoustic model to generate the corresponding sound. The performance of the proposed biomechanical-acoustic model is demonstrated by simulating six vowels [a, ə, ε, e, i, u] and three vowel-vowel utterances [ai, au, ui]. These vowels were selected to cover some of the most significant sounds that the biomechanical model can articulate without lip rounding, since the latter is not controlled in the current model.

The paper is organized as follows. Section 2 explains the methods used to develop the dynamic 3D biomechanical-acoustic model. It contains the description of the biomechanical model (Section 2.1), the derivation of muscle activation patterns (MAP) to drive this model (Section 2.2), the reconstruction of dynamic vocal tract geometries from the biomechanical model (Section 2.3), and the simulation of sound wave propagation with the acoustic model (Section 2.4). The results of the simulations are reported in Section 3 and compared with those obtained with a geometrical model, which generates a vowel-vowel utterance by linearly interpolating the starting and ending vowel vocal tract geometry². Section 4 finally concludes the paper. A preliminary version of this work was presented in³⁰.

FIGURE 1 Adapted biomechanical model developed in Artisynt showing the response of the tongue to activation of different muscles. *GG*: genioglossus, divided into three parts: posterior (*GGP*), middle (*GGM*), and anterior (*GGA*), *SG*: styloglossus, *HG*: hyoglossus, *MH*: mylohyoid, *GH*: geniohyoid, *V*: verticalis, *T*: transversus, *IL*: inferior longitudinal, *SL*: superior longitudinal.

2 | METHODS

2.1 | Biomechanical model

The biomechanical model used in this work is an adapted version of FRANK²¹, developed in ArtiSynth³¹. We have adapted this model as follows: 1) To speed up simulation and improve stability, the face mesh was removed, only keeping the geometry of the lips. As the facial muscles have been removed, lip rounding can no longer be controlled. Instead, the upper lip is attached to the maxilla and is static while the lower lip is attached to the mandible and follows its movement. 2) In order to generate adequate acoustic results from the model, the lower part of the pharynx geometry (in the laryngopharynx region) was modified to be more vertical and the larynx structure was translated 7 mm towards the anterior compared to the original FRANK model. This modification was found necessary, as the different anatomical structures in FRANK are partly modeled on different subjects. Adapting the entire model to an individual speaker is beyond the scope of this work, and we instead make smaller adjustments that are necessary for the acoustic simulations. 3) The epiglottis was removed from the model, since we lack data on how to control it systematically.

Figure 1 depicts the midsagittal view of the adapted model, which is composed of the rigid bodies maxilla, mandible, and hyoid bone, and the deformable bodies lips, tongue, soft palate (velum), pharynx wall, and larynx structure. For the simulations, the soft palate muscles were activated to close the velopharyngeal port. To speed up the simulation and improve stability, all structures are treated as rigid bodies, except the tongue. In addition, only the tongue, lower lip, mandible, and hyoid bone were treated as dynamic, while the remaining structures were static. The tongue is controlled by eleven muscles. Figure 1 shows how the tongue is deformed by activation of different muscles. Two muscle groups control the jaw (mandible) opening (JO) and closing (JC). For more details about the biomechanical model, we refer the reader to the previous publications on the ArtiSynth FRANK model²¹.

FIGURE 2 Inverse model for the estimation of muscle activation. (a) Location of four marker points on the tongue, (b) generation of the EMMA-like trajectory path (dashed line) between two points: P_{org} and P_{des} , (c) velocity profile of the trajectory.

2.2 | Muscle activation

2.2.1 | Inverse model for the estimation of muscle activation

Electromyography (EMG) data may be used to determine the activation of the muscles in the biomechanical model. However, collecting EMG data for a speech production task is extremely invasive, and normal speech production may be restricted by the measurement hardware. According to the authors' knowledge, there is only one study of measuring extrinsic tongue muscles, in which [əvpv] utterances were measured³². Moreover, as tongue muscles are interwoven in a complex manner, EMG measurements of an individual muscle risk having low confidence. As an alternative, we use an inverse model that determines muscle activations for a given spatial trajectory of points³³. This is a quadratic program that minimizes the velocity tracking error at each time step. To solve the muscle redundancy, it uses an ℓ^2 -norm regularization term.

To use the inverse model, four marker points are positioned in the model. One point is located on the lower incisor (LI), and the others on the tongue tip (TT), tongue middle (TM), and tongue back (TB), as depicted in Figure 2 a. For each utterance, the corresponding spatial trajectories of those markers are prescribed. In the first attempt, we tried to use trajectories from a previously collected electromagnetic midsagittal articulatory (EMMA) dataset³⁴, but although the subject in the EMMA dataset was the one for whom the jaw-palate-tongue model was created in ArtiSynth, it was not possible to use the data directly. The reason was firstly that the exact positions of the sensors in the EMMA dataset were not available, making it difficult to position the corresponding markers in the model. Secondly, the initial configuration of the subject and the model was different in rest position, leading to alignment problems. Thirdly, we found some differences in the palate shape created by EMMA tracing with the geometry of the palate in the ArtiSynth model.

Instead, we use observations of the EMMA data and general articulation knowledge, such as the fact that EMMA trajectories are usually curved rather than straight lines³⁵, to generate EMMA-like trajectories. The trajectory path between an origin (P_{org}) and a destination (P_{des}) is generated by fitting a Bézier curve using three points P_{org} , P_{mid} , P_{des} as shown by the dashed line in Figure 2 b, where P_{mid} controls the curvature of the trajectory and is determined as

$$\begin{aligned} \vec{P} &= \vec{P}_{des} - \vec{P}_{org}, \quad d = ||\vec{P}||, \quad \theta = \angle P - \pi/2, \\ \vec{P}_{mid} &= \vec{P}_{org} + 0.5 \times \vec{P} + 0.25 \times d \times [\cos\theta \quad \sin\theta]. \end{aligned} \quad (1)$$

The generated path of Figure 2 b is combined with the velocity profile shown in Figure 2 c to generate the trajectory.

2.2.2 | Sustained vowels

For a sustained vowel, artificial trajectories are generated to move the jaw and tongue from the rest position (with no active muscle force) to a target articulation of that vowel. The inverse model estimates the muscle activations for that given trajectory. Then, the estimated muscle activations are input to the forward model, and after reaching the equilibrium the area function and the corresponding formants are calculated using a 1D acoustic solver³⁶, as an intermediate check of the resulting sound. The procedure is repeated for a few iterations, in a supervised manner, by changing the final positions, P_{des} , of the markers. That is,

TABLE 1 Estimated activation of the tongue and jaw muscles to articulate the sustained vowels.

Vowel	GGP	GGM	GGA	SG	HG	MH	GH	V	T	IL	SL	JO	JC
[ɑ]	0	5	20	0	15	0	22	7	0	0	0	0.02	0
[ə]	0	0	0	0	0	0	0	0	0	0	0	0	0
[ɛ]	53	0	13	9	0	68	5	5	2	31	16	0	0.15
[e]	44	0	0	0	1	21	0	0	11	25	9	0	0.19
[i]	100	0	0	0	3	15	0	0	58	63	19	0	1.05
[ʊ]	26	0	1	25	0	52	0	9	1	7	0	0	0.77

Zero and hundred respectively represent no active and maximum active force generated by the muscle.

FIGURE 3 Trajectories of four markers (blue: desired, red: model output) when moving from rest position to sustained vowels. Tongue and jaw contours are shown at the rest (dashed) and at the target articulation (solid).

the initial P_{des} may not result in the typical articulation and/or formant frequencies of the target vowel. In some cases, it may also be impossible to reach the position of all four markers simultaneously because of the model constraints, e.g. if two marker's positions are too far from each other. The 1D acoustic solver is employed only in this step since 3D acoustic simulations are time consuming and hence not suitable for the iterative procedure. For the final simulations, when the muscle activations have been determined, 3D acoustic model is instead used.

Running the iterative procedure to estimate the muscle activations, we observed that in some cases, the inverse model activates a muscle that makes reaching the final positions difficult. If this happened, the activity of that muscle was forced to be zero, unless the muscle is known to be essential for articulating the vowel, from anatomical knowledge and data on tongue muscle activation from previous studies^{32,37,18,20,38}.

The final estimated muscle activations are reported in Table 1 . Using these muscle activations as input to the model, the tongue and jaw move from the rest position towards the target articulation. **The desired trajectories prescribed by Eq. (1) and the corresponding ones generated by the biomechanical model are depicted in Figure 3 .**

2.2.3 | Vowel-vowel utterances

Two ways of generating the muscle activation pattern for vowel-vowel utterances may be considered: either interpolation in the muscle activation space between the MAPs of the two vowels, or by designing a spatial trajectory and using the inverse method. These two alternatives have relevance in speech motor control theory and correspond to assuming that the units of speech are static or dynamic, respectively. As it was concluded in³⁹, there is not enough evidence that the input to the speech production

FIGURE 4 Muscle activation pattern in a V_1V_2 utterance generated by a four-phase interpolation scheme. The example contains the activation of three muscles.

system is dynamic, and hence linear interpolation in the muscle activation space is utilized here. That is, if the corresponding MAPs for vowels V_1 and V_2 are respectively M_1 and M_2 , the MAP for utterance V_1V_2 is generated using a four-phase interpolation scheme, as depicted in Figure 4 (only three muscles are shown). In the first phase $[t_0, t_1]$, the MAP is interpolated linearly from zero to M_1 , so that the biomechanical model moves from rest position to the articulation of the first vowel. Then, in the second phase $[t_1, t_2]$, the MAP stays constant at M_1 which sustains the first vowel. The transition between vowels is produced in the third phase $[t_2, t_3]$ by linearly interpolating the MAP from M_1 to M_2 . Finally, the second vowel is sustained in the fourth phase $[t_3, t_4]$ with the MAP M_2 . Note that the model always starts from the neutral position (zero activations), which produces a vocal tract shape like the neutral vowel $[\emptyset]$. Therefore, the generated MAP corresponds to an $[\emptyset V_1 V_2]$ utterance.

2.3 | Vocal tract geometry

An acoustic model requires a finite computational domain with specified boundary conditions. One may think of including all the structures of the biomechanical model in the acoustic computational domain, but that is far from trivial and would be computationally very costly. A practical and efficient alternative is to limit the computational domain to the vocal tract airway, which is what is relevant for the acoustics. The computational domain becomes delimited by the vocal tract boundary and the motion of the boundary is controlled through muscle activation in the biomechanical model. Note that this requires the geometry of the vocal tract boundary to be determined from the surrounding anatomical structures in the biomechanical model. One option to do so could be the growing circle method described in²⁹. That method provides a very complete description of the vocal tract which not only includes the main cavity but also all subbranches such as the interdental space, the sublingual cavity and the piriform fossae. Nonetheless, too many details will pose severe challenges to numerical simulations with evolving vocal tracts, given that subbranches can merge and unmerge with the main tract depending on the vowel²⁹. The computational domain would thus include regions (i.e., branches) that appear and disappear, an intricate feature of the model that is out of the scope of the current work and might be considered in the future. A simpler approach is suggested here so as to directly obtain the main cavity of the vocal tract. The general concept of our method to blend surrounding structures is described in Section 2.3.1 and then applied to the vocal tract in Section 2.3.2. In Section 2.3.3, we show how the acoustic computational domain is deformed using the moving boundaries of the vocal tract.

2.3.1 | Boundary of an enclosed region

When a region of interest (ROI) is enclosed by two or more objects (i.e. polygons), boundaries of these objects implicitly express the geometry of the ROI. To find an explicit representation, geometrical boolean (GB) operations (see, e.g.,⁴⁰) may be used. The problem with GB is that it requires the surrounding boundaries to be perfectly aligned. However, this requirement is almost never satisfied in a biomechanical model, where there are gaps and overlaps between the boundaries of surrounding structures. We therefore use an alternative approach, line-of-sight (LOS), which is illustrated by Figure 5. Assuming that the approximate

center of the ROI, p^c , is known, the boundary of the ROI, B_e , is determined as

$$P_j = R_j \cap B_i, \forall i \in (1, 2), \quad (2a)$$

$$P_j^e = \{p | p \in P_j \wedge (\forall x \in P_j : |p - p^c| < |x - p^c|)\}, \quad (2b)$$

$$B_e = \{P_j^e | \forall j \in (1, 2, \dots, N)\}. \quad (2c)$$

In (2a), R_j is a ray with origin at p^c , and P_j is a set of intersection points between R_j and all surrounding boundaries, B_i , and P_j^e in (2b) is the closest point to p^c among that set (see Figure 5 a). Equations (2a) and (2b) are repeated for N rays with origin at p^c and covering the range $[0, 2\pi]$, as shown in Figure 5 b. Using Equation (2c), the boundary of the ROI, B_e , is detected by ordering the points P_j^e and connecting them with linear segments, as shown by the red polygon in Figure 5 c. The accuracy of the ROI boundary, B_e , is determined by the number of rays (N). The situation in Figure 5 a-Figure 5 c is that of perfect alignment of the upper and lower polygons. However, as said, the LOS can deal with more complex scenarios in which gaps and overlaps between the boundaries appear. For instance, in Figure 5 d the LOS generates a closed and smoothed boundary, although there exist small tail-like branches between the two polygons which would be problematic for the 3D simulations. In contrast, it cannot reconstruct a bulge in the boundary as exemplified in Figure 5 e. However, this situation is rare when the vocal tract is a tube without subbranches. Alternatively, one may utilize the growing circle method which can deal with such situations and to include the subbranches²⁹.

FIGURE 5 Explanation of the LOS method to extract the boundary of the ROI enclosed by two polygons.

2.3.2 | Reconstruction of the vocal tract

The vocal tract is a cavity in 3D which is surrounded by several structures. In the previous section, the boundary detection problem was addressed in 2D. The complexity of the problem increases tremendously in 3D. Therefore, the 3D reconstruction of the vocal tract is simplified into a set of 2D problems using a set of slicing planes **having their centre of the ROI at the midsagittal cross-section of the plane, p_j^c** . Figure 6 a shows the midsagittal cross-section of the structures overlaid with a set of rays, each ray corresponds with one slicing plane. To calculate p_j^c , the surrounding structures of the vocal tract are grouped as $B_1 = \{\text{upper lip, maxilla, soft palate, pharynx, larynx}\}$, i.e. the structures that are either superior or posterior to the cavity, and $B_2 = \{\text{lower lip, mandible, tongue, hyoid bone, larynx}\}$, i.e. the structures that are either inferior or anterior (laryngeal part) to the cavity. As the larynx structure surrounds the cavity, it belongs to both groups. Then, each ray, noted as R_j^{ms} , is intersected with the midsagittal contours of the structures,

$$P_j^i = R_j^{ms} \cap C_i^{ms}, \forall i \in (1, 2), \quad (3a)$$

$$p_j^1 = \{p | p \in P_j^1 \wedge (\forall x \in P_j^1 : |p - p_j^o| < |x - p_j^o|)\}, \quad (3b)$$

$$p_j^2 = \{p | p \in P_j^2 \wedge (\forall x \in P_j^2 : |p - p_j^o| > |x - p_j^o|)\}, \quad (3c)$$

$$p_j^c = \text{mean}(p_j^i), \forall i \in (1, 2), \quad (3d)$$

$$C^{mid} = \{p | p \in p_j^c, \forall j \in (1, 2, \dots, M)\}. \quad (3e)$$

where C_i^{ms} represents the midsagittal contours of B_i , P_j^i is a set of intersection points between ray R_j^{ms} and C_i^{ms} . p_j^1 is the nearest point to p_j^o (origin of the ray) within set P_j^1 , and p_j^2 is the farthest points to p_j^o within set P_j^2 . The median point between them, p_j^c , is considered as a point on the centerline of the enclosed cavity. The centerline (C^{mid}) is formed by concatenating all p_j^c . Figure 6 b illustrates the calculations for one of the rays. In the next step, on each slicing plane, the LOS method detects the boundary of the ROI using the p_j^c as its center. This results in a set of boundaries, as shown in Figure 6 c, which also illustrates

FIGURE 6 Reconstructing the vocal tract: (a) midsagittal cross section and rays, (b) centerline calculation, (c) boundaries of ROI and triangulation between the consecutive cross sections.

how adjacent cross-sections are then connected to form a triangulated boundary mesh of the cavity. Applying the reconstruction method at each time step of the biomechanical simulations provides a deforming (i.e. time-varying) geometry.

In order to correct the centerline and the slicing planes, the procedure may be iterated. That is, once the boundary of the cavity is obtained, the centerline is recalculated using the geometry of the cavity, rather than the midpoint defined by (3), and the slicing planes are redefined accordingly (see⁴¹ for more details). This correction is not essential for 3D acoustic simulations¹⁰, but is relevant when the area function is considered.

A boundary detection method such as LOS is required to identify the cavity on each cross-section. Besides, omission of some anatomical parts in the model may result in wide gaps and ambiguous boundary of the ROI. In such a case, using LOS may result in anatomical artifacts such as the example of Figure 7, which depicts five consecutive cross-sections in the velar region. The blue boundaries in Figure 7 b and 7 c, are significantly different from the previous cross-section (Figure 7 a) and the next (Figure 7 d). To avoid such artifacts, boundary points P_j^e in (2) that are more distant from the center of ROI, p^c , than a threshold are discarded before forming the boundary. This results in the red boundaries shown in Figure 7. A global threshold of 3 cm is used for all cross-sections. This threshold was empirically determined based on anatomical knowledge of the oral cavity.

FIGURE 7 Five consecutive cross-sections in the velar region and their ROI boundaries detected by the LOS method without any distance thresholding (blue line), and with distance thresholding (red line).

2.3.3 | Computational domain

The proposed method in Section 2.3.2 provides the vocal tract geometry (i.e. 3D surface mesh) describing the boundaries of the cavity. Yet, sound waves also radiate outside from the mouth. To account for this effect, each vocal tract is set in the center of the flat circular surface of a hemisphere of radius 0.08 m as shown in Figure 8 a. The vocal tract geometry together with this radiation space constitute the computational domain Ω in which acoustic wave propagation has to be simulated. The boundaries of Ω (see Figure 8 a) consist of the vocal tract walls Γ_W , the glottal cross-section Γ_G , the circular baffle Γ_H in which the vocal tract is set, and a non-reflecting boundary Γ_∞ that absorbs sound waves so as to emulate free-field propagation. The next step

consists in generating a discrete representation of Ω to obtain a volumetric mesh. To do so, the air volume contained within the vocal tract geometry is meshed with unstructured tetrahedral elements of size $h \approx 0.003$ m, whereas within the hemisphere h is designed to smoothly change from $h \approx 0.003$ m in the immediate region close to the mouth to $h \approx 0.005$ m in the curved surface of the hemisphere. This gives about 140.000 elements in total. The above configuration is followed to generate the computational domains for the sustained vowels.

In the case of vowel-vowel utterances, the radiation space is removed to keep the computational cost reasonably low. Thus, Ω is terminated at the mouth exit Γ_M , as shown in Figure 8 b. As a vowel-vowel utterance requires the volume mesh to be deformed according to the movement of the vocal tract geometry Γ_W , we need to either perform remeshing of Ω at each time step, or move the vertices of an initial volume mesh so that it follows the deformation of Γ_W . Remeshing approaches keep the mesh quality very high. However, they are computationally expensive and difficult to implement, not only because of the remeshing itself, but also because the solution (acoustic pressure) of the previous time step has to be projected to the new mesh. Thus, the latter approach, deforming an initial mesh, is employed in this work, and the element distortion is minimized by choosing a proper initial mesh. That is, e.g., to generate the initial volume mesh for utterance [ai], the vocal tract geometry corresponding to an articulation between [a] and [i] is utilized as shown in Figure 8 b. Using tetrahedral elements of size $h \approx 0.004$ m results in volume meshes with ≈ 8.000 elements.

To deform the initial volume mesh, the displacements of the vertices located on Γ_W are provided by the reconstruction method of Section 2.3.2 at each time step. Since this displacement may not be smooth enough, a Bézier smoother is applied to remove small fluctuations over time. Inner vertices of the volume mesh have also to move according to the vocal tract walls, which is achieved by solving the Laplacian equation at each time step, using the Finite Element Method (FEM). Let us denote all vertex coordinates as $\mathbf{x}$ and the coordinates of those vertices on Γ_W as $\mathbf{x}_w$. The equation to solve at a time instant $t = t^{n+1}$ reads

$$\nabla^2 \mathbf{x}^{n+1} = 0 \quad \text{in } \Omega, \quad t = t^{n+1}, \quad (4a)$$

$$\mathbf{x}^{n+1} = \mathbf{x}_w^{n+1} \quad \text{on } \Gamma_W, \quad t = t^{n+1}, \quad (4b)$$

$$\mathbf{x}^{n+1} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_G, \quad t = t^{n+1}, \quad (4c)$$

$$\mathbf{x}^{n+1} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_M, \quad t = t^{n+1}. \quad (4d)$$

where the superscript n denotes the time step. Note that Eq. (4b) prescribes the displacement of the vertices on the vocal tract walls, while Eq. (4c) and Eq. (4d) impose the displacement to be zero in the normal direction $\mathbf{n}$ (pointing outwards), which avoids an artificial lengthening of the vocal tract. Solving this equation allows one to smoothly translate the movement of the boundary vertices to the inner ones. **Finally, the velocity of the domain is computed at a node $\mathbf{x}_i$ of the volum mesh as**

$$\mathbf{u}_{\text{dom}}^{n+1}(\mathbf{x}_i) = \frac{\mathbf{x}_i^{n+1} - \mathbf{x}_i^n}{\Delta t}, \quad (5)$$

with Δt being the time step. This magnitude will be needed by the acoustic model (see Section 2.4).

FIGURE 8 Examples of computational domains Ω used to simulate (a) the sustained vowels, and (b) the vowel-vowel utterances at the first time instant (initial Ω). The boundaries of Ω are denoted by Γ_G for the glottal cross-section, Γ_W for the vocal tract walls, Γ_H for the circular flat baffle in which the vocal tract is set, Γ_∞ for a non-reflecting boundary, and Γ_M for the mouth aperture.

2.4 | Acoustic model

The physics of acoustic waves propagating through a dynamic vocal tract can be described by the mixed wave equation for the acoustic pressure $p(\mathbf{x}, t)$ and acoustic particle velocity $\mathbf{u}(\mathbf{x}, t)$, expressed in an ALE (Arbitrary Lagrangian-Eulerian) frame of reference to account for the vocal tract movement⁸. This equation reads

$$\frac{1}{\rho_0 c_0^2} \partial_t p - \frac{1}{\rho_0 c_0^2} \mathbf{u}_{\text{dom}} \cdot \nabla p + \nabla \cdot \mathbf{u} = 0, \quad (6a)$$

$$\rho_0 \partial_t \mathbf{u} - \rho_0 \mathbf{u}_{\text{dom}} \cdot \nabla \mathbf{u} + \nabla p = 0, \quad (6b)$$

with c_0 standing for the speed of sound, ρ_0 for the air density, $\mathbf{u}_{\text{dom}}$ for the velocity of the domain (see Section 2.3.3), and ∂_t for the first partial time derivative. This equation has to be solved in the computational domains Ω described in Section 2.3.3, so it has to be supplemented with some boundary and initial conditions:

$$\mathbf{u} \cdot \mathbf{n} = g(t) \quad \text{on } \Gamma_G, \quad t > 0, \quad (7a)$$

$$\mathbf{u} \cdot \mathbf{n} = p/Z_w \quad \text{on } \Gamma_W, \quad t > 0, \quad (7b)$$

$$p = p/Z_0 \quad \text{on } \Gamma_\infty, \quad t > 0, \quad (7c)$$

$$p = 0 \quad \text{on } \Gamma_M, \quad t > 0, \quad (7d)$$

$$p = 0, \quad \mathbf{u} = 0 \quad \text{in } \Omega, \quad t = 0. \quad (7e)$$

Equation (7a) imposes a particle velocity $g(t)$ at the entrance of the vocal tract to introduce the sound waves generated by the vocal folds. In Eq. (7b) wall losses are considered using the impedance Z_w . Equation (7c) only applies to the simulation of vowels, in which the radiation from the mouth aperture is allowed, to absorb the sound waves that propagate in the radiation space (see Fig. 8 a). As mentioned in Section 2.3.3, radiation losses are not considered for vowel-vowel utterances (see Figure 8 b) to reduce the computational time. Thus, the boundary condition in (7c) is replaced by (7d), imposing zero acoustic pressure, to emulate an open-end boundary condition. The ALE mixed wave equation (6) with the corresponding boundary and initial conditions in (7) are solved numerically using the FEM. In particular, the subgrid scale strategy in⁸ is followed to avoid numerical instabilities.

For vowel sounds, the volume mesh is static, and thus it suffices to solve the ALE mixed wave equation (6) by setting $\mathbf{u}_{\text{dom}} = 0$. To obtain the vocal tract transfer function of vowels, the following Gaussian pulse is used as the excitation source,

$$u_g(t) = e^{-[(t-T_{gp})/0.29T_{gp}]^2} [\text{m/s}], \quad (8)$$

with $T_{gp} = 0.646/f_c$ and $f_c = 10$ kHz. This pulse is low-pass filtered at 10 kHz to avoid numerical errors above the maximum frequency of analysis, $f_{\text{max}} = 10$ kHz. Then a vocal tract transfer function is computed as

$$H(f) = \frac{P_o(f)}{Q(f)}, \quad (9)$$

with $P_o(f)$ and $Q(f)$ respectively standing for the Fourier transform of the acoustic pressure $p_o(t)$ captured at the exit of the vocal tract (red dot in Figure 8 a) and the input volume velocity $Q(t)$ at the glottis, i.e. $Q(t) = u_g(t)S_g$, with S_g being the glottal cross-sectional area. The corresponding sound is generated by convolving the inverse Fourier transform of $H(f)$ with a train of glottal pulses.

For vowel-vowel utterances, the volume mesh first deforms from an intermediate articulation to the first vowel of the utterance, as described in Section 2.3.3. During this time only the Laplacian equation (4) is solved to obtain the vertex coordinates. Once the articulation of the first vowel is reached, both Eq. (4) and (6) are solved to compute the displacement of the vertices, and also to simulate the propagation of sound waves. A train of glottal pulses is generated using a Rosenberg model⁴² to obtain the signal $g(t)$. These glottal pulses are enhanced by considering a pitch curve, a fade in/out to emulate the onset/offset of the vocal folds, and some shimmer and jitter (similar to⁷). Finally, the resulting signal is low pass filtered at 10 kHz to smooth the abrupt transitions that the Rosenberg model produces at the end of the closing phase.

In all simulations the speed of sound is set to $c_0 = 350$ m/s, and the wall impedance to $Z_w = 83666 \text{ kg/m}^2\text{s}^{43}$. Simulations are conducted with a constant time step of $\Delta t = 5 \mu\text{s}$ in the time interval $[0, T]$.

3 | RESULTS

3.1 | Sustained vowels

Feeding the estimated muscle activations, determined in Section 2.2 (Table 1), to the biomechanical model moves the jaw and defines the tongue shape. After reaching the equilibrium, the vocal tract geometry is reconstructed, and its volume mesh is generated as illustrated in Section 2.3. The 3D acoustical behavior of the resulting vocal tract configuration is then simulated, obtaining a vocal tract transfer function $H(f)$, and generating the corresponding sound, as described in Section 2.4. The six vowel sounds [ɑ], [ə], [ɛ], [e], [i] and [u] are simulated. The vocal tract reconstruction methodology is illustrated in Figure 9 ,

FIGURE 9 Reconstruction of vocal tract geometries for sustained vowels: (a) biomechanical model, cross-sections in (b) oropharynx, (c) velar , and (d) oral regions, (e) reconstructed vocal tract geometries.

TABLE 2 First five formant frequencies and bandwidths for vowels

Vowel	Formant frequencies (Hz)					Formant bandwidths (Hz)				
	F_1	F_2	F_3	F_4	F_5	BW_1	BW_2	BW_3	BW_4	BW_5
[ɑ]	664	1134	2810	3917	4446	123	98	243	219	1338
[ə]	521	1666	2738	3608	4465	93	106	163	175	346
[ɛ]	513	1763	2569	3475	4274	85	124	172	165	202
[e]	463	1916	2706	3533	4360	84	134	223	187	258
[i]	300	2116	2864	3506	4325	78	160	233	203	229
[ɯ]	386	1362	2602	3515	4115	108	127	161	183	203

which depicts the biomechanical model articulating each vowel (Figure 9 a) together with cross-sections in the oropharynx (Figure 9 b), velar (Figure 9 c), and oral (Figure 9 d) regions. The corresponding reconstructed vocal tract geometries are shown in Figure 9 e.

3.1.1 | Muscle activations and articulation

The tongue has a very complex muscular structure that plays an important role in shaping the vocal tract geometry. To gain some insight into the role of different tongue muscles, our observations from the biomechanical model for articulating different vowels are next described.

Low-back vowel [ɑ]: The most important properties are the constriction in the oropharynx (Figure 9 b) and a large opening in the oral region (Figure 9 d). This is achieved by pulling the tongue backwards and at the same time lowering the jaw. The oropharyngeal constriction is primarily created by *HG*, which lowers the tongue and pulls it backwards towards the pharynx wall. *Since this also raises the tongue tip and the tongue dorsum, and the large oral opening (Figure 9 d) requires the tongue tip and dorsum to be lowered, GGA and GGM need to be activated to, respectively, lower the tongue tip and dorsum.* The lateral sides of the tongue are lowered by activation of *V*, for which the fiber directions are similar to those of *GG*, but are running in the parasagittal plane. The jaw is lowered by *GH* and by *JO*, which is a group of muscles involved in opening the mouth.

High-front vowel [i]: Contrary to [ɑ], a large oropharyngeal cavity and a small oral constriction should be created. This is mainly achieved by activation of *GGP* which pulls the tongue forwards (Figure 9 b) and at the same time elevates it because of the hydrostatic nature of the tongue. Full activation of *GGP* is however not sufficient to achieve the oral constriction, and *MH* and *T* are therefore activated. *MH* elevates the whole tongue by stiffening the oral floor, whereas *T* does this by squeezing the lateral sides of the tongue. Cooperative activation of *IL* and *SL*, as synergy muscles, keeps the tongue apex behind the teeth and prevents it from touching the palate. This forms a proper constriction between the tongue and hard palate (Figure 9 d). Activation of *JC*, which is a group of muscles involved in closing the mouth, elevates the whole tongue.

High-back vowel [ɯ]: The constriction in the velar region (Figure 9 c) is mainly achieved through activation of *SG*, which pulls the tongue backwards and, to some less extent, up. In order to get the proper place of constriction in the velar region, rather than in the upper oropharynx, *GGP* and *MH* are activated to lift the tongue upwards, as is *JC* to slightly raise the jaw.

Central vowels [ɛ, e]: In general central vowels have more freedom and they can hence be produced with less precise control and more variations. It seems that *GGP*, *MH*, and *IL* are the three most important muscles for producing these vowels. As expected, the activation of muscles is small compared to the cardinal vowels.

In general the jaw muscles, *JO*, *JC*, have very small activations compared to tongue muscles. This may be explained by the fact that the jaw muscles are strong because of their functionality in chewing. In addition, bite-block experiments show that the jaw position is not vital for producing vowels.

3.1.2 | Vocal tract geometry

The reconstructed vocal tract geometries are shown in Figure 9 e, while Figures 9 b, 9 c, 9 d show three cross-sections, respectively in the oropharynx, velar, and oral regions. The extracted contours of the cavity show that the method is able to determine the airway boundary with adequate accuracy. The small overlaps that may exist between the boundaries of the airway and other structures is the result of boundary smoothing. In some cases, parts of the airway are excluded, primarily because they belong to a subbranch, such as the velopharyngeal port (c.f. Figure 9 c) or the sublingual cavity (c.f. Figure 9 d). This exclusion is intentional, in order to be able to perform the acoustic simulations of a time-changing vocal tract without remeshing of the computational domain. Note e.g., that the interdental cavity in vowel [i] would merge with the main cavity in vowel [a] and deforming the surface mesh for [ai] would hence be extremely difficult without the remeshing approach. It should however be noted that a more complex boundary detection method has been developed in the authors' previous work where the subbranches can be included in the geometry²⁹.

It also happens that regions belonging to the main cavity are excluded as a consequence of the implementation of the boundary detection method. The first cross-section of Figure 9 c (vowel [a]) shows an example where a region between the tongue and the pharynx has been excluded. **Similarly, Figure 9 d shows the exclusion of the interdental space, which is a consequence of the distance-to-center threshold, required because of the omission of some anatomical geometries in the biomechanical model.**

Determining the acoustic and possible perceptual consequences of these exclusions are out of the scope of the present study, and could be investigated in future research. According to previous studies, such subbranches generate dips in the acoustic transfer function of the vowels (see, e.g.,^{44,45}). In summary, the proposed method detects the boundaries of the main cavity with adequate accuracy in most of the cross-sections. The accuracy is indeed lower in a few ones, linked to imperfections of the biomechanical model and/or to the implementation of the boundary detection method. Despite the marginal shortcomings of the automatic boundary detection, it is a prerequisite for synthesis of sequences in which the vocal tract deforms, as e.g., a 300 ms simulation with a sampling period of 1 ms requires the generation of 300 geometries, and manual segmentation of 12,000 boundaries (assuming 40 slicing planes for each geometry) is not a sensible option. The importance of the computational efficiency and stability of the boundary detection method in our application is hence indisputable.

3.1.3 | Formants and sound

Figure 10 shows a snapshot of the acoustic pressure distribution of vowel [a] at time instant $t = 18.75$ ms, obtained by solving the wave equation. To enhance the visualization, two limiting values are used within the vocal tract (Figure 10 a), and in the

FIGURE 10 Snapshot of the acoustic pressure distribution of vowel [a] FEM simulation at time instant $t = 18.75$ ms. Two different limiting values are used to enhance the visualization of the acoustic waves (a) within the vocal tract and (b) in the free-field radiation space. The surface mesh of the vocal tract is shown in (b) instead of the acoustic pressure contour fill.

FIGURE 11 Generation of vowel sound [a]: train of glottal pulses $Q(t)$ and resulting acoustic pressure $p_o(t)$. (a) Total signal and (b) zoom on the first 100 ms.

free-field (Figure 10 b). For each vowel, the vocal tract transfer function $H(f)$ is computed as described in Section 2.4 to extract the first five formant frequencies and their bandwidths, reported in Table 2. Moreover, the corresponding vowel sounds are generated. In this case, a vowel sound is generated by convolving a train of glottal pulses of duration ~ 650 ms with the impulse response of each vocal tract. The latter is simply obtained after computing the inverse Fourier transform of $H(f)$. The pressure waveform $p_o(t)$ of vowel [a] is plotted in Figure 11 (the other vowels have similar waveforms) together with the acoustic volume velocity $Q(t)$ of the train of glottal pulses. Note the fade in/out applied to the glottal pulses used to emulate the onset/offset of the vocal folds, which is translated to the generated sound. The generated audio files correctly reproduce the simulated vowels (refer to a.wav, schwa.wav, e1.wav, e2.wav, i.wav and w.wav for [a], [ə], [ɛ], [e], [i] and [ʊ], respectively).

3.2 | Vowel-vowel utterances

The three vowel-vowel utterances [ai], [au], and [ui] are simulated. **We depart from the biomechanical model in a rest position, which produces a vocal tract shape similar to the neutral vowel [ə]. The muscle activations that produce the articulation of each vowel (see Table 1) are next linearly interpolated as shown in Figure 4. First, the MAP is zero for the first 10 ms ($t_0 = 10$ ms), that is, no muscles are activated. Then it linearly increases to the articulation of the first vowel in 100 ms ($t_1 = 110$ ms), and is sustained for 10 ms ($t_2 = 120$ ms). The transition time to the second vowel is also set to 100 ms ($t_3 = 220$ ms), and the second vowel is finally sustained for 80 ms ($t_4 = 300$ ms). This results in a 300 ms simulation, from which 190 ms corresponds to the vowel-vowel utterance to be generated (starting at $t_{1s} = 110$ ms). The boundary detection algorithm automatically generated 300 vocal tract geometries from the biomechanical model for each utterance, which corresponds to a time resolution of 1 ms. These geometries are next used by the volume mesh and acoustic modules to generate the corresponding sound, as described in Section 2.3.3 and Section 2.4, respectively. Simulations lasted about XX h in a serial desktop computing system with processor Intel(R) Core(TM) i7-6700 3.4 GHz, from which XX% of the time corresponds to the biomechanical model (XX min), XX% to the extraction of the vocal tract geometries from the biomechanical model (XX s), XX% to moving the volume mesh (4 min), and XX% to the 3D acoustic model (2 h 15 min).**

3.2.1 | Articulation

Applying the muscle activations to the biomechanical model results in the jaw and tongue movement shown in Figure 12 by the trajectory of the four markers and their smoothed velocity profiles. We observe the loop-like trajectories, similar to those which have been observed in EMMA measurements³⁵. Indeed, observing nonlinear trajectories while using linear activation signals is an indication of the fact that the mechanical properties cause these nonlinear trajectories rather than assuming complex activation patterns of the tongue. This is in agreement with a previous study using a 2D biomechanical model⁴⁶, where the authors concluded that this kind of trajectory is a result of biomechanical constraints rather than an active control mechanism. The first part ($t < 120$ ms) of the velocity profile corresponds to the movement from rest to V_1 while the second ($t > 120$ ms) corresponds to transition from V_1 to V_2 and sustained articulation of V_2 . The maximum velocity is the result of the mechanical properties and the timing of the activation signal in Figure 4.

FIGURE 12 Trajectories and smoothed velocity profiles of the jaw and the tongue markers.

3.2.2 | Sound and spectrogram

Figure 13 presents a sequence of four snapshots from the acoustics simulation of [ai] acquired at time instants $t = 120, 180, 220$ and 260 ms. As can be observed, the vocal tract shape deforms from a low-back articulation to a high-front. At $t = 120$ ms the first vowel is articulated ($t = [110, 120]$), $t = 180$ and 220 ms belong to the transition ($t = [120, 220]$), whereas at $t = 260$ ms the second vowel is being sustained ($t = [220, 300]$). The acoustic pressure on the vocal tract walls is presented together with the volume mesh. Different limiting values are chosen for the color scales to improve the visualization of acoustic waves in each frame. Besides, the acoustic pressure is captured on a node close to the mouth aperture to allow to listen to the generated sounds (see supplementary audio files).

FIGURE 13 Snapshots from the acoustics simulations of [ai] at time instants 120, 180, 220 and 260 ms.

The spectrograms of the synthesized [ai], [aʊ] and [ʊi] are shown in Figure 14. As usually done in speech research, a pre-emphasis FIR filter, with coefficients $[1 -0.97]$, is applied to the generated audio files to better visualize the high frequency components. It can be observed how the typical formants of the first vowel smoothly transit to those of the second vowel during the production of each utterance. However, despite it is difficult to observe in the spectrograms, it is worth mentioning that the formant values of sustained vowels and those of VV utterances differ. This is attributed to radiation losses (see e.g.⁴⁷ for their influence), which are not considered for vowel-vowel utterances to reduce the computational times, in contrast to sustained vowels.

FIGURE 14 Spectrograms of the synthesized vowel-vowel utterances. Values are in dB.

3.2.3 | Contribution of the biomechanical model

The vowel-vowel utterances simulated in this work are generated using the biomechanical model with linear interpolation of muscle activity. However, it would be also interesting to examine which differences are produced when the dynamic vocal tract geometries are directly obtained from the linear interpolation of the starting and ending vowel vocal tract geometries of the utterance. To this aim, the utterance [ai] is generated following the two approaches. For the geometric interpolation, we use the same time-sequence as for the simulations using the biomechanical model, but those geometries belonging to the transition between vowels (interval $t = [120, 220]$ ms) are replaced by geometries generated by linear interpolation between the starting and ending vowels (i.e., the geometries of time instants $t = 120$ ms and $t = 220$ ms).

FIGURE 15 Spectrogram of [ai] synthesized (a) by linear interpolation in the muscle space (biomechanical model) and (b) by linear interpolation of the vocal tract geometries. (c) Differences between the two spectrograms. Values are in dB.

The spectrograms of the two versions of [ai] and the difference spectrogram are depicted in Figure 15. As one may expect, the largest differences are produced between 120 ms and 220 ms when the transition takes place (the deviations at high frequencies at $t > 250$ ms are due to numerical errors caused by the very low pressure levels in this frequency region). The second formant F_2 is the most affected one, although the trajectory of the others also display some differences. It is to be noted that the slope of F_2 is an important perceptual cue for consonants in CV utterances^{48,49}, and observed differences in transition may hence have perceptual relevance. It should further be considered that geometrical interpolation cannot account for coarticulation. That is, changing the timing of the interpolation only scales the time axis, and not the target formants, whereas, in a biomechanical model, coarticulation is implicitly addressed because of the mechanical properties of the articulators. Changing the speech rate (timing of the interpolation in the muscle space) affects the target formants and transitions in a nonlinear way.

4 | CONCLUSIONS

In this work, a link is established between a 3D biomechanical model and a 3D acoustic model to simulate vowel-vowel utterances. We have demonstrated an automatic method to develop a deformable vocal tract mesh that follows the motion of the articulators in biomechanical simulations, and that is used as the computational domain to solve the mixed wave equation in an Arbitrary Lagrangian-Euler framework. This is important as it creates the possibility to study speech production from different perspectives, including physiology, articulation and acoustics, with a single, unified and realistic model. While 3D biomechanical models are essential to properly represent aspects like muscles fibers, tongue muscular hydrostat and other biomechanical constraints, 3D acoustic models improve the accuracy of the simulated sound, since they can deal with 3D propagation effects. A coupled biomechanical and acoustic 3D model may intrinsically account for speech dynamics, coarticulation, etc.

Two main simplifications are assumed in this work. First, the vocal tract geometry only includes the main cavity. Subbranches such as vallecula, piriform fossa, and sublingual cavity have been excluded, as they would pose significant difficulties for the acoustic simulations, only solvable with costly remeshing strategies. Second, we have limited the study to vowels and vowel-vowel sequences. Extending the work to all phonetic categories is not trivial, and requires notable efforts to improve all model parts. For example, to generate a fricative sound, the biomechanical model should be able to form a precise, narrow constriction, which demands a refined tongue mesh with very small elements, and a detailed deformation control. Also, the acoustic model needs to handle turbulent noise^{50,51,52}.

Regarding the muscle activation procedure, let us note that the utilized estimation for static vowels based on inversion of stylized articulatory trajectories is not the only option. Alternatives include e.g., the use of articulation data from EMMA or (static and/or real-time) MRI, or acoustic-to-articulatory inversion. However, such strategies involve problems which are beyond the scope of this work like the translation of articulation data from another subject to the biomechanical model; the many-to-one mapping between acoustics, articulation and muscle activation and the coarseness of the meshes in the biomechanical model compared to the resolution of imaging techniques such as MRI.

Concerning the vowel-vowel utterances, we have shown that linear interpolation in the muscle space results in spectrograms that are different from linear interpolation in the geometrical space. Such differences are less likely to change the category of the generated vowel, but they highlight the importance of the mechanical properties of the articulators in speech synthesis. The observed difference in the F_2 transition might be relevant for consonant-vowel sequences, but perceptual experiments with generated consonant-vowel utterances generated with linear interpolation in muscle space or geometrical space would be required to investigate this further.

Another interesting observation is that using linear interpolation in the muscle activation space creates articulatory trajectories similar to those observed in real data from EMMA, whereas linear interpolation in the geometric space leads to less realistic transitions. Based on the conclusions in³⁹, we argue that linear interpolation of muscle activation is an adequate speech motor control strategy. However, it would be of interest to investigate other alternatives, such as non-linear interpolation in the muscle activation space, inverse modeling from articulatory trajectories or optimization of the acoustic transition. It would further be of interest to investigate if differences in interpolation strategy, exclusion of subbranches, and the observed imperfections in the boundary extraction have any perceptual effects.

ACKNOWLEDGEMENTS

This research has partially been supported by EU-FET grant EUNISON 308874. The second and fourth authors also acknowledge the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P.

References

1. Arnela M, Dabbaghchian S, Blandin R, et al. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. *J. Acoust. Soc. Am.*. 2016;140(3):1707–1718.

2. Arnela Marc, Dabbaghchian Saeed, Guasch Oriol, Engwall Olov. MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs. *IEEE/ACM Trans. Audio Speech Lang. Process.*. 2019;27(12):2173–2182.
3. Story B H. TubeTalker: An airway modulation model of human sound production. In: First Int. Work. Performative Speech Sing. Synth.; 2011; Vancouver, Canada.
4. Mullen J, Howard D M, Murphy D T. Waveguide physical modeling of vocal tract acoustics: Flexible formant bandwidth control from increased model dimensionality. *IEEE Trans. Audio, Speech Lang. Process.*. 2006;14(3):964–971.
5. Mullen J, Howard D M, Murphy D T. Real-Time Dynamic Articulations in the 2-D Waveguide Mesh Vocal Tract Model. *IEEE Trans. Audio, Speech Lang. Process.*. 2007;15(2):577–585.
6. Arnela M, Guasch O. Two-dimensional vocal tracts with three-dimensional behavior in the numerical generation of vowels. *J. Acoust. Soc. Am.*. 2014;135(1):369–379.
7. Arnela M, Guasch O. Finite Element Synthesis of Diphthongs Using Tuned Two-Dimensional Vocal Tracts. *IEEE/ACM Trans. Audio Speech Lang. Process.*. 2017;25(10):2013–2023.
8. Guasch O, Arnela M, Codina R, Espinoza H. A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production. *Acta Acust. united with Acust.*. 2016;102(1):94–106.
9. Arnela M, Dabbaghchian S, Guasch O, Engwall O. Finite element generation of vowel sounds using dynamic complex three-dimensional vocal tracts. In: Proc. ICSV23 – 23th Int. Congr. Sound Vib.; 2016; Athens, Greece.
10. Arnela M, Dabbaghchian S, Guasch O, Engwall O. A semi-polar grid strategy for the three-dimensional finite element simulation of vowel-vowel sequences. In: Proc. Interspeech 2017; 2017; Stockholm.
11. Gully A J, Daffern H, Murphy D T. Diphthong synthesis using the dynamic 3D digital waveguide mesh. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*. 2018;26(2):243–255.
12. Perkell J S, Guenther F H, Lane H, et al. A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *J. Phon.*. 2000;28(3):233–272.
13. Zandipour M. A modeling investigation of vowel-to-vowel movement planning in acoustic and muscle spaces. PhD thesis, Boston University, 2007.
14. Dabbaghchian S, Engwall O. Does speech production require precise motor control?. In: 7th Int. Conf. Speech Mot. Control; 2017; Groningen, the Netherlands.
15. Perkell J S. A physiologically-oriented model of tongue activity in speech production. PhD thesis, Massachusetts Institute of Technology, 1974.
16. Payan Y, Perrier P. Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech Commun.*. 1997;22(2):185–205.
17. Sanguineti V, Laboissière R, Ostry D J.. A dynamic biomechanical model for neural control of speech production. *J. Acoust. Soc. Am.*. 1998;103(3):1615–1627.
18. Takemoto H. Morphological analyses of the human tongue musculature for three-dimensional modeling. *J. Speech Lang. Hear. Res.*. 2001;44(1):95–107.
19. Dang J, Honda K. Construction and control of a physiological articulatory model. *J. Acoust. Soc. Am.*. 2004;115(2):853–870.
20. Buchaillard S, Perrier P, Payan Y. A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning.. *J. Acoust. Soc. Am.*. 2009;126(4):2033–2051.
21. Anderson P, Fels S, Harandi N M, et al. Chapter 20 - FRANK: A hybrid 3D biomechanical model of the head and neck. In: Payan Yohan, Ohayon Jacques, eds. *Biomech. Living Organs*, Oxford: Academic Press 2017 (pp. 413–447).

22. Perrier P, Boë L-J, Sock R. Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast modeling the transition with two sets of coefficients. *J. Speech, Lang. Hear. Res.*. 1992;35(1):53–67.
23. Soquet A, Lecuit V, Metens T, Demolin D. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Commun.*. 2002;36(3):169–180.
24. Stavness I, Sánchez C A, Lloyd J, et al. Unified skinning of rigid and deformable models for anatomical simulations. In: SIGGRAPH Asia 2014 Tech. Briefs; 2014; Shenzhen, China.
25. Monson B B, Hunter E J., Lotto A J, Story B H. The perceptual significance of high-frequency energy in the human voice. *Front. Psychol.*. 2014;5:587.
26. Kang Z, Ji Z. Acoustic length correction of duct extension into a cylindrical chamber. *J. Sound Vib.*. 2008;310(4):782–791.
27. Dang J, Honda K. An improved vocal tract model of vowel production implementing piriform resonance and transvelar nasal coupling. In: Proc. ICSLP 96; 1996.
28. Dang J, Wei J, Honda K, Nakai T. A study on transvelar coupling for non-nasalized sounds. *J. Acoust. Soc. Am.*. 2016;139(1):441–454.
29. Dabbaghchian S, Arnela M, Engwall O, Guasch O. Reconstruction of vocal tract geometries from biomechanical simulations. *Int. J. Numer. Methods Biomed. Eng.*. 2019;35(2):e3159.
30. Dabbaghchian S, Arnela M, Engwall O, Guasch O. Synthesis of VV utterances from muscle activation to sound with a 3D model. In: Proc. Interspeech 2017; 2017; Stockholm, Sweden.
31. Lloyd J E, Stavness I, Fels S. ArtiSynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In: Payan Yohan, ed. *Soft tissue Biomech. Model. Comput. Assist. Surg.*, Berlin: Springer Berlin Heidelberg 2012 (pp. 355–394).
32. Baer T, Alfonso P J, Honda K. Electromyography of the tongue muscles during vowels in /əpvp/ environment. *Ann. Bull. RILP*. 1988;22:7–19.
33. Stavness I, Lloyd J E, Fels S. Automatic prediction of tongue muscle activations using a finite element model.. *J. Biomech.*. 2012;45(16):2841–2848.
34. Youssef A B. Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation. PhD thesis, Université de Grenoble, 2011.
35. Mooshammer C, Hoole P, Kühnert B. On loops. *J. Phon.*. 1995;23(1):3–21.
36. Espy-Wilson C, Zhou X, Tiede M, Boyce S. New features in VTAR: A Matlab-based computer program for vocal tract acoustic modeling. *J. Acoust. Soc. Am.*. 2007;121(5):3136–3136.
37. Honda K. Organization of tongue articulation for vowels. *J. Phon.*. 1996;24(1):39–52.
38. Wu X, Dang J, Stavness I. Iterative method to estimate muscle activation with a physiological articulatory model. *Acoust. Sci. Technol.*. 2014;35(4):201–212.
39. Lindblom B. The organization of speech movements: specification of units and modes of control. *From Sound to Sense 50+ Years Discov. Speech Commun.*. 2004;50.
40. Vatti B R. A generic solution to polygon clipping. *Commun. ACM*. 1992;35(7):56–63.
41. Dabbaghchian S, Arnela M, Engwall O. Simplification of vocal tract shapes with different levels of detail. In: Proc. 18th Int. Congr. Phonetic Sci.; 2015; Glasgow, Scotland, UK.
42. Rosenberg A E. Effect of glottal pulse shape on the quality of natural vowels.. *J. Acoust. Soc. Am.*. 1971;49(2B):583–590.

43. Švancara P, Horáček J. Numerical modelling of effect of tonsillectomy on production of Czech vowels. *Acta Acust. united with Acust.*. 2006;92(5):681–688.
44. Dang J, Honda K. Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.*. 1997;101(1):456–465.
45. Takemoto H, Mokhtari P, Kitamura T. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method.. *J. Acoust. Soc. Am.*. 2010;128(6):3724–3738.
46. Perrier P, Payan Y, Zandipour M, Perkell J. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *J. Acoust. Soc. Am.*. 2003;114(3):1582–1599.
47. Arnela M. Numerical production of vowels and diphthongs using finite element methods. PhD thesis, La Salle, Universitat Ramon Llull, 2015.
48. Sussman H M, Shore J. Locus equations as phonetic descriptors of consonantal place of articulation.. *Percept. Psychophys.*. 1996;58(6):936–46.
49. Lindblom B, Sussman H M. Dissecting coarticulation: How locus equations happen. *J. Phon.*. 2012;40(1):1–19.
50. Yoshinaga T, Nozaki K, Wada S. Experimental and numerical investigation of the sound generation mechanisms of sibilant fricatives using a simplified vocal tract model. *Phys. Fluids*. 2018;30(3):035104.
51. Pont A, Guasch O, Baiges J, Codina R, Van Hirtum A. Computational aeroacoustics to identify sound sources in the generation of sibilant /s/. *Int. J. Numer. Method. Biomed. Eng.*. 2018;;e3153.
52. Pont A, Guasch O, Arnela M. Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff vortices. *Int. J. Numer. Method. Biomed. Eng.*. 2020;36(2):e3302.