Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds

Marc Arnela[a)]
GTM–Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, C/Quatre Camins 30, Barcelona, E-08022, Catalonia, Spain

Saeed Dabbaghchian
Department of Speech, Music and Hearing, School of Computer Science & Communication, KTH Royal Institute of Technology, Stockholm, Sweden

Rémi Blandin
GIPSA-lab, Unité Mixte de Recherche au Centre National de la Recherche Scientifique 5216, Grenoble Campus, St. Martin d'Heres, F-38402, France

Oriol Guasch
GTM–Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, C/Quatre Camins 30, Barcelona, E-08022, Catalonia, Spain

Olov Engwall
Department of Speech, Music and Hearing, School of Computer Science & Communication, Kungliga Tekniska högskolan Royal Institute of Technology, Stockholm, Sweden

Annemie Van Hirtum and Xavier Pelorson
GIPSA-lab, Unité Mixte de Recherche au Centre National de la Recherche Scientifique 5216, Grenoble Campus, St. Martin d'Heres, F-38402, France

(Received 17 December 2015; revised 29 July 2016; accepted 16 August 2016; published online 15 September 2016)

For many years, the vocal tract shape has been approximated by one-dimensional (1D) area functions to study the production of voice. More recently, 3D approaches allow one to deal with the complex 3D vocal tract, although area-based 3D geometries of circular cross-section are still in use. However, little is known about the influence of performing such a simplification, and some alternatives may exist between these two extreme options. To this aim, several vocal tract geometry simplifications for vowels [ɑ], [i], and [u] are investigated in this work. Six cases are considered, consisting of realistic, elliptical, and circular cross-sections interpolated through a bent or straight midline. For frequencies below 4–5 kHz, the influence of bending and cross-sectional shape has been found weak, while above these values simplified bent vocal tracts with realistic cross-sections are necessary to correctly emulate higher-order mode propagation. To perform this study, the finite element method (FEM) has been used. FEM results have also been compared to a 3D multimodal method and to a classical 1D frequency domain model. © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4962488]

[ZZ] Pages: 1707–1718

I. INTRODUCTION

The vocal tract geometry has a very complex three-dimensional (3D) shape. Its volume representation has been captured in many works by using, for instance, magnetic resonance imaging (MRI) (see, e.g., Rokkaku *et al.*, 1986; Baer *et al.*, 1991; Story *et al.*, 1996; Engwall and Badin, 1999) or computed tomography (CT) (see, e.g., Sundberg *et al.*, 1987). For years, this volume has subsequently been simplified to generate a vocal tract area function, which describes the variations of the cross-sectional area along its center midline. This has allowed 1D approaches to generate voice with a fairly good quality and also with a large flexibility thanks to the low dimensional representation of the vocal tract geometry (see, e.g., Kelly and Lochbaum, 1962; Fant, 1970; Sondhi and Schroeter, 1987; Story, 2005; Doel and Ascher, 2008; Story, 2013; Birkholz, 2013). However, it is well known that this classical approach can only approximate vocal tract acoustics in the low frequency range (below 4–5 kHz) where the plane wave assumption is satisfied. For higher frequencies, higher order modes are also excited (see, e.g., Blandin *et al.*, 2015) that cannot be captured by 1D methods. Beyond this, little is known on the information that is lost when simplifying the complex 3D vocal geometry to a 1D area function. It is the main purpose of this work to shed some light on this topic.

Current 3D approaches have allowed one to directly resort to 3D vocal tract geometries and deal with the induced complex 3D acoustic field. Very detailed MRI-based vocal tract geometries have previously been used in earlier studies

a)Electronic mail: marnela@salle.url.edu

to analyze the vocal tract acoustics (see, e.g., Švancara and Horáček, 2006; Takemoto *et al.*, 2010; Arnela *et al.*, 2016). Alternatively, very rough simplified geometries consisting of 3D circular tubes have also been generated from 1D vocal tract area functions (see, e.g., Vampola *et al.*, 2008; Speed *et al.*, 2013; Arnela and Guasch, 2014). The first geometries give a very detailed representation of the vocal tract and therefore of its acoustics, while the second ones are easy to generate and manipulate, which makes them especially appealing for the production of dynamic sounds (see, e.g., Arnela *et al.*, 2014; Guasch *et al.*, 2015, 2016, where interpolation between static geometries was used to generate diphthong sounds). Yet, a large variety of options exist between these two configurations, which may help us to understand which are the effects of simplifying the vocal tract from a very detailed 3D MRI-based geometry to a 1D area function. These simplifications may provide a better balance between voice quality and flexibility.

In this study we will focus on the effects of using simplifications of the main conduct of the vocal tract for vowel sounds. The lips and side-branches such as the piriform fossae or valleculae will not be considered, as their acoustics effects were previously explored, for instance, in Takemoto *et al.* (2010), Takemoto *et al.* (2013), Arnela *et al.* (2013), Vampola *et al.* (2015), and Arnela *et al.* (2016). MRI-based vocal tract geometries for vowels [ɑ], [i], and [u] adapted from a 3D vocal tract database (Aalto *et al.*, 2014) will be used as a reference to generate simplified vocal tract shapes with different levels of detail. To do so, the cross-sectional shape and the vocal tract midline will be extracted. These parameters will be used to generate different simplified vocal tract shapes with the same area function. Three configurations for the cross-sectional shape will be examined, namely, realistic, elliptical, and circular. In the first one, the original shape of each cross-section will be preserved, while in the second and third they will be approximated by an ellipse and a circle of equivalent area. The obtained cross-sections will be combined with the original vocal tract midline and with a straightened centerline so as to produce bent and straight vocal tracts (see Fig. 1).

The finite element method (FEM) will be used to examine the vocal tract acoustics of each configuration. The time-domain wave equation for the acoustic pressure combined with a perfectly matched layer (PML) to account for free-field radiation will be numerically solved (see, e.g., Arnela and Guasch, 2013, for details). For completeness, the results obtained with FEM will be contrasted to those computed using a 3D multimodal method (Blandin *et al.*, 2015) and to a simple 1D frequency domain model based on transfer matrices (similar to that of Sondhi and Schroeter, 1987). Both approaches need to use simplifications of the vocal tract, such as those examined in this work. The former can work with any cross-sectional shape, but it is limited to straight configurations, while the latter needs 1D vocal tract area functions, which are somehow analogous to a straight 3D vocal tract with circular cross-sections. Despite these limitations, they have been shown to provide very low computational costs compared to FEM, and can be viewed as reasonable alternatives depending on the particular application.[1]

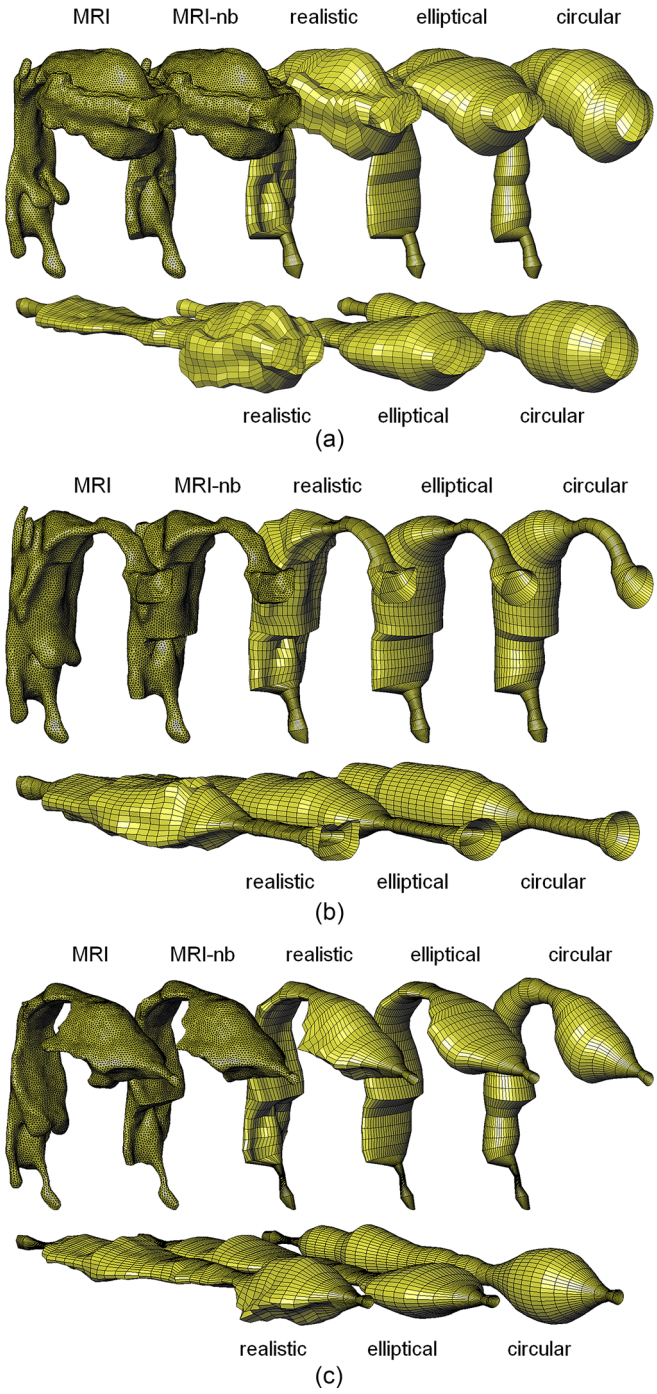

FIG. 1. (Color online) Vocal tract models.

The paper is structured as follows. The methodology used to generate each one of the vocal tract geometry simplifications is presented in Sec. II, as well as the configuration of the finite element simulations used to examine their acoustic behavior. The obtained results are next presented in Sec. III. First, the variations produced when a vocal tract geometry is discretized by a finite set of realistic cross-sections are analyzed. Second, the influence of the number of cross-sections is examined. Third, the effects of different cross-sectional shapes in a bent configuration are studied. Fourth, the effects produced when a vocal tract with different cross-sectional

shapes is straightened are investigated. Finally, the results obtained for a 3D multimodal method and a 1D model are compared with those of FEM. Conclusions close the paper in Sec. IV.

II. METHODOLOGY

A. Vocal tract models

The three-dimensional vocal tract geometries for vowels [ɑ], [i], and [u] generated from MRI by Aalto *et al.* (2014) were adapted for this work. The subglottal tube, part of the face and neck were first removed. The lips were also removed from the vocal tract at the mouth termination plane, which was defined as the last front-plane that produces a closed outline when it intersects with the vocal tract (see Arnela *et al.*, 2016, for their influence). The resulting geometry constitutes the reference case and the one that will be successively simplified. Hereafter it will be termed as the MRI case or MRI geometry (see Fig. 1, top leftmost configuration for each vowel). On the other hand, since in this work the effects of the piriform fossae and valleculae will not be considered (see, e.g., Takemoto *et al.*, 2010, and Takemoto *et al.*, 2013, for some related works), additional reference cases without these side branches have also been generated ("MRI no branches" or "MRI-nb," see Fig. 1, top second configurations for each vowel).

The MRI geometries were next simplified with different degrees of detail. This requires one to first extract the centerline and the shape of every MRI geometry at different cross-sections. The methodology to do so will be described in Sec. II B. Three configurations for the shape of each cross-section were defined, which either preserve the original outline shape (termed realistic in Fig. 1) or approximate it as an ellipse or a circle. Then, each one of the obtained cross-sections were linearly interpolated to produce a bent vocal tract by using the original vocal tract midline, or with a straightened centerline that keeps the overall vocal tract length. Some examples for vowels [ɑ], [i], and [u] with 40 cross-sections can be observed in Fig. 1.

B. Procedure for simplifying vocal tract geometries

The original, intricate geometry reconstructed from MR images (see MRI in Fig. 1) is simplified through resampling and reconstruction to create simplified vocal tract meshes with specified cross-sectional shape. The governing criterion for the simplification procedure is that the simplified vocal tract geometries should have the same area function as the original shape.

Figure 2 gives a flow-chart overview of the two-step simplification procedure divided into analysis and synthesis. The analysis part determines the vocal tract centerline and the corresponding perpendicular cross-sections (PCS), while the synthesis part generates simplified vocal tract geometries using the centerline and the PCS. Six different types of geometries are generated by combining either the original curved centerline or a straightened centerline with the three cross-sectional shapes.

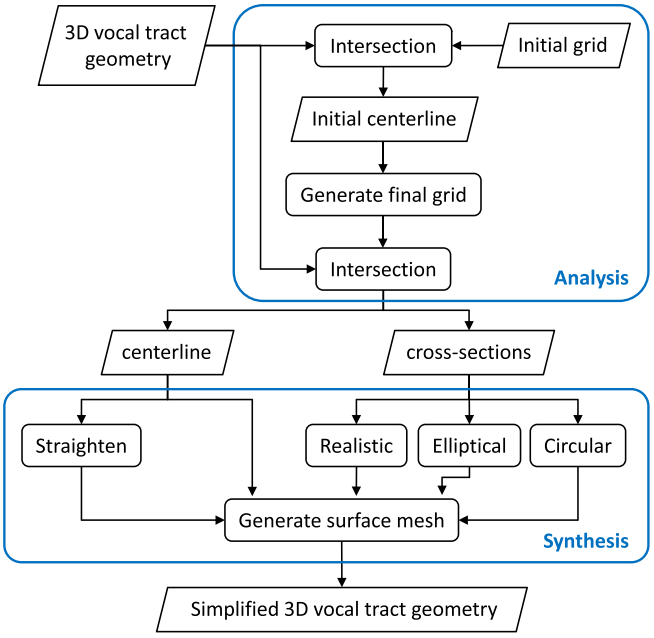

FIG. 2. (Color online) Flow chart of the two-step simplification procedure of 3D vocal tract geometries. In the first step, the vocal tract centerline and its perpendicular cross-sections are extracted. In the second step, the cross-sectional shape is simplified to realistic, elliptical, or circular. The resulting cross-sections are then linearly interpolated through a bent or a straight centerline so as to generate simplified vocal tract geometries (see Fig. 1).

1. Analysis

The concept of defining the vocal tract centerline and the PCS relies on the classical assumption that the acoustic waves propagate along the vocal tract centerline, and that the PCS therefore approximates the wavefronts, which are perpendicular to the direction of propagation. This means, on the one hand, that the centerline is required in order to determine the PCS; and on the other hand, that the centerline is in fact the line through the centers of these cross-sections. In order to solve this circular reference, a standard methodology for 3D vocal tract shape reconstruction and area function calculation from MR images is followed (similar to Kröger *et al.*, 2000), but herein modified to be applicable to 3D vocal tract geometries (Dabbaghchian *et al.*, 2015).

The method takes the vocal tract geometry and a grid with a set of planes as input, as shown in Fig. 2. A semi-polar grid is employed as initial grid, with 30 horizontal planes in the pharynx, 20 polar planes in the velar region, and 30 vertical planes in the oral cavity [see Fig. 3(a)]. Each line in the grid represents a plane that is approximately perpendicular to the vocal tract [shown as the midsagittal contour in Fig. 3(a)]. Cross-sections are generated as the intersection between the planes and the vocal tract geometry. The algorithm forms an initial centerline by connecting the centers of these cross-sections.

The initial grid is next adapted by using the tangent of the centerline as the normal vector of the gridplanes. This means that the derivative of the centerline has to be calculated. In order to avoid abrupt changes, the initial midline is smoothed using Bézier splines and interpolated so as to obtain a large number of points. By sampling this smoothed centerline uniformly with the number of desired gridplanes

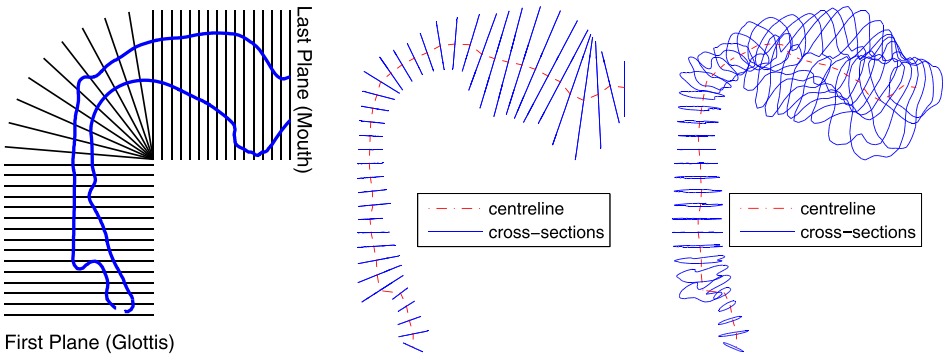

FIG. 3. (Color online) (a) Initial grid superimposed on the vocal tract midsagittal boundary of vowel [ɑ], (b) and (c) final centerline with 40 perpendicular cross-sections (2D and 3D view).

(e.g., 40 gridplanes to obtain 40 cross-sections), the normal vectors of the new gridplanes are defined. Manual intervention may be required if a large number of gridplanes is used or when the centerline tangent changes significantly, since neighboring gridplanes may then overlap. Should this happen, the gridplane normals are slightly adjusted to avoid the overlap. This hardly affects the results as seen from the close matching between the performance of the MRI-nb geometry and its realistic simplification, in Sec. III A. As far as the termination planes at the glottal end and the mouth opening are concerned, these are, respectively, defined as the first and last planes for which the intersection with the vocal tract is a closed contour. The normal vector of these two gridplanes is not altered by the grid adaptation.

Once the final grid is generated, the procedure of intersecting the vocal tract with the grid is repeated to find the PCS. The final centerline is next determined by connecting their centers. Note that, although the initial smoothed centerline has been uniformly sampled, the new centerline does not guarantee that the cross-sections are evenly distributed. However, this is not a necessary requirement for the correct development of this work. Figures 3(b) and 3(c) show the final centerline and cross-sections when 40 gridplanes are selected.

2. Synthesis

The first part of the algorithm results in a curved centerline and the actual PCS that perfectly match the vocal tract outline [see Figs. 3(b) and 3(c)]. The second part of the algorithm simplifies the cross-sections and optionally straightens the centerline (see Fig. 2).

The cross-sectional shape can be set to realistic, elliptical, or circular (see Fig. 4). For the realistic shape the outline of the original PCS is maintained, but it is resampled to 48 points which are distributed evenly. For the elliptic shape, the lateral dimension of the vocal tract specifies the ellipse's major axis length and the minor axis is calculated to preserve the area of the original PCS. Similarly, for the circular shape, the radius is set to preserve the PCS area. To generate a simplified vocal tract geometry with bending, the original centerline is used to locate in space each PCS. In the case of straightened vocal tracts, the cross-sections are placed with the same consecutive distances as on the curved centerline. This distance is defined as the Euclidean distance between the center of these cross-sections. However, sagittal variations of the cross-section centers are not considered in this computation, since they can artificially lengthen the resulting straight vocal tract geometry (Story *et al.*, 1996). Finally, the contour points of neighboring cross-sections are connected to form a quad-faced surface mesh (see Fig. 1, where 40 cross-sections are selected and 48 points are used to discretize each cross-section).

C. Time domain finite element simulations

A custom finite element code for acoustic wave propagation was used to simulate the acoustics of each vocal tract simplification. This program numerically solves the time-domain wave equation for the acoustic pressure,

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

combined with a PML formulation to emulate free-field propagation. In Eq. (1), $p(\boldsymbol{x}, t)$ denotes the acoustic pressure, c_0 is the speed of sound and ∂_{tt}^2 stands for the second order time derivative. Details about the implementation of this code can be found in Arnela and Guasch (2013).

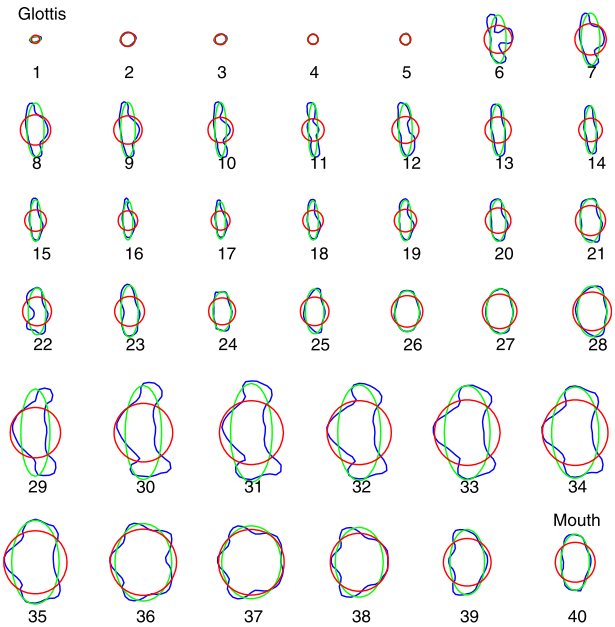

FIG. 4. (Color online) Simplification to realistic, elliptical, and circular cross-sectional shapes. In this example, 40 cross-sections are considered, with section 1 denoting the glottal end of the vocal tract and section 40 the mouth exit.

Several computational domains were generated. They consist of the vocal tract models shown in Fig. 1 set in a rigid flat baffle of dimensions 0.3×0.3 m, an attached rectangular volume with dimensions $0.3 \times 0.3 \times 0.2$ m that allows sound waves radiate from the mouth exit, and a PML of length 0.1 m and a relative reflection coefficient of 10^{-4} surrounding the free-field radiation volume that absorbs the incoming sound waves. Tetrahedral elements were used to generate the finite element meshes, with sizes comprising from $h = 0.001$ m within the vocal tract geometry, to $h = 0.0025/0.005$ in the free-field volume and $h = 0.0075$ m in the PML.

As far as the boundary conditions are concerned, a constant boundary admittance of $\mu = 0.005$ was used to take into account vocal tract wall losses. A Gaussian pulse of the type

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

with $T_{gp} = 0.646/f_0$ and $f_0 = 10$ kHz, was prescribed at the glottal cross-section as an input volume velocity $q_i(t)$. This pulse was low pass-filtered at the maximum frequency of analysis (10 kHz) to avoid spurious numerical errors. The rest of the boundaries were assumed to be acoustically rigid.

An FEM numerical simulation was then carried out with a speed of sound of $c_0 = 350$ m/s and an air density of $\rho_0 = 1.14\,\mathrm{kg/m}^3$. A sampling frequency of $f_s = 8000$ kHz was used for the evolution of the time scheme. Such a high value is needed to fulfill a very restrictive stability condition of the Courant-Friedrich-Levy type required by explicit numerical schemes. Time events of 20 ms were simulated, capturing the evolution of the acoustic pressure p_o at a node located in the free-field volume, 0.04 m from the center of the vocal tract exit. A vocal tract transfer function was finally computed as

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

where $P_o(f)$ and $Q_i(f)$, respectively, stand for the Fourier transforms of $p_o(t)$ and $q_i(t)$.

In addition, the acoustic pressure distribution for a formant or antiresonance was also eventually computed. To do so, the Gaussian pulse in Eq. (2) introduced at the glottal area was replaced with a sinusoidal signal, its frequency matching that of the formant or antiresonance to be analyzed. The time evolution of the acoustic pressure within the vocal tract can then be visualized after running a 20 ms simulation.

III. RESULTS

The results obtained for each one of the considered simplifications of the MRI-based vocal tract geometries for vowels [ɑ], [i], and [u] are presented next. As detailed in Sec. II A, they have been generated by linear interpolation of cross-sections with realistic, elliptical or circular shape through a bent or a straight midline (see Fig. 1). For simplicity, hereafter they will be termed, e.g., as elliptical-bent or circular-straight cases. The first step to generate these simplifications consists in discretizing the MRI geometry in a finite set of cross-sections. In Sec. III A, the effects produced when doing so are analyzed by comparing the acoustic behavior of the MRI geometry to that of the realistic-bent case. In addition, the MRI geometry without branches is also included in the comparisons since the analyzed simplifications only entail the main conduct of the vocal tract. Since a different number of cross-sections can also be used during this process, its influence on the realistic-bent case was next studied in Sec. III B by comparing the results obtained with 40, 60, and 80 cross-sections. Once a proper number of cross-sections were determined, the effects of cross-sectional shape in a bent configuration were analyzed in Sec. III C. Realistic, elliptical, and circular cross-sections were considered for this purpose. The effects of bending were next studied in Sec. III D. Finally, in Sec. III E, the results obtained from a multimodal method and a 1D model were compared to those coming from FEM. All results were analyzed in terms of vocal tract transfer functions. In addition, formant locations and bandwidths were also extracted for each vowel, method, and vocal tract model. They are listed in Table I.

A. Simplification to a vocal tract with realistic cross-sections

First, the effects produced on the vocal tract acoustics when the MRI geometry is simplified by a finite set of cross-sections were analyzed. The simplification that best approximates this geometry was chosen for this purpose. Among the considered cases, this corresponds to the one containing 80 cross-sections of realistic shape in a bent configuration. The vocal tract transfer functions obtained for this simplification are represented in Fig. 5 together with those of the MRI case and the MRI configuration without branches ("MRI-nb" in the figure).

As far as the low frequency region below 4 kHz is concerned, a small deviation of the formants towards higher frequencies is produced when the MRI geometry is simplified for all the vowels (see Fig. 5 and Table I). This is observed whenever side branches are removed from the MRI case and when the vocal tract is discretized by a finite set of cross-sections. Indeed, the side branches seem to play a determinant role in the correct location of some of the formants. For instance, for vowel [i] the shift of the second formant (F2) is stronger when side branches are removed than for the realistic case (see Table I). However, this is not a general rule. For instance, for vowel [ɑ] the third formant (F3) is more affected by the simplification procedure of the main conduct than for the removal of the side branches (see Table I). Putting aside the influence of the side branches, in general it can be observed that the realistic configuration can shift some of the formants to higher frequencies when compared to the MRI-nb case (see, e.g., F4, F5, and F6 of vowels [ɑ], [i], and [u]). With regard to the formant bandwidths, no significant deviations are observed between the herein analyzed cases (see Table I).

Beyond 4 kHz the differences become more apparent. Comparing first the MRI-nb case to the MRI one, as expected,

TABLE I. First formant frequencies and bandwidths for the different vocal tract models of vowels [ɑ], [i], and [u]. They consist of two reference models, the original MRI-based vocal tract with side branches (MRI) and without them (MRI nb), and six simplifications, generated from the combination of realistic, elliptical and circular cross-sections with bent and straight vocal tract midlines. The number of cross-sections for each simplification is set to 80, unless it is specifically denoted in the table between brackets. Values are extracted from the computation of vocal tract transfer functions using the FEM, a multimodal method (MM), or a one-dimensional model (1D).

			Formant frequencies (Hz)						Formant bandwidths (Hz)					
Vowel	Method	Vocal tract model	F_1	F_2	F_3	F_4	F_5	F_6	BW_1	BW_2	BW_3	BW_4	BW_5	BW_6
[ɑ]	FEM	MRI	638	1140	2297	3080	3792	4369	123	115	123	162	197	165
		MRI-nb	670	1156	2294	3160	3820	4302	119	115	123	157	184	170
		Bent-realistic (80)	673	1164	2323	3259	3861	4349	121	116	122	159	185	173
		Bent-realistic (60)	681	1172	2339	3302	3885	4402	119	115	122	163	189	166
		Bent-realistic (40)	678	1176	2362	3441	3969	4454	118	117	122	169	205	178
		Bent-elliptical	678	1173	2353	3205	3895	4423	111	111	119	156	203	170
		Bent-circular	681	1172	2388	3320	3925	4418	91	102	106	149	189	136
		Straight-realistic	646	1115	2264	3157	3800	4275	116	108	118	155	232	233
		Straight-elliptical	653	1127	2300	3190	3852	4347	106	103	114	150	231	209
		Straight-circular	653	1123	2325	3227	3872	4372	88	95	102	141	212	179
	MM	Straight-realistic	645	1123	2282	3133	3825	4332	15	47	31	15	138	94
		Straight-elliptical	652	1131	2305	3149	3862	4369	12	49	33	17	141	92
		Straight-circular	657	1135	2343	3171	3879	4397	16	49	32	16	142	93
	1D	Area functions	643	1099	2308	3241	3848	4318	71	88	93	141	203	165
[i]	FEM	MRI	214	2036	3036	3255	3822	4878	76	94	—	—	131	147
		MRI-nb	221	2113	3080	3290	4017	5432	77	94	—	—	116	196
		Bent-realistic (80)	220	2134	3150	3356	4065	5464	78	93	—	—	118	194
		Bent-realistic (60)	221	2145	3183	3385	4091	5474	80	92	—	—	118	195
		Bent-realistic (40)	222	2164	3235	3452	4143	5520	80	90	—	—	119	217
		Bent-elliptical	225	2136	3117	3340	4071	5482	73	92	—	—	114	192
		Bent-circular	227	2153	3180	3370	4096	5487	73	86	—	—	112	210
		Straight-realistic	223	2057	3055	3297	3947	5431	76	92	—	—	120	213
		Straight-elliptical	222	2068	3076	3309	3975	5466	70	89	—	—	117	232
		Straight-circular	222	2073	3098	3330	3985	5476	68	85	—	—	115	234
	MM	Straight-realistic	222	2063	3056	3370	3997	5491	9	10	27	110	37	122
		Straight-elliptical	223	2072	3063	3378	4012	5514	10	8	25	113	37	138
		Straight-circular	223	2074	3085	3388	4014	5523	10	13	30	112	38	143
	1D	Area functions	220	2095	3075	3275	4005	5559	90	64	—	—	115	394
[u]	FEM	MRI	264	704	2165	2438	3296	4310	91	83	136	—	115	100
		MRI-nb	284	752	2187	2560	3453	4451	92	84	164	—	107	93
		Bent-realistic (80)	289	756	2228	2640	3500	4467	96	82	147	—	106	92
		Bent-realistic (60)	289	762	2254	2762	3554	4475	92	81	130	—	110	91
		Bent-realistic (40)	297	772	2282	2904	3609	4485	94	82	123	—	119	93
		Bent-elliptical	288	758	2237	2632	3535	4508	93	80	148	—	104	90
		Bent-circular	288	768	2253	2645	3542	4523	84	75	132	—	91	82
		Straight-realistic	281	743	2200	2607	3444	4396	89	82	150	—	106	92
		Straight-elliptical	284	745	2212	2619	3476	4449	92	79	145	—	102	89
		Straight-circular	283	751	2218	2618	3473	4457	83	74	131	—	91	81
	MM	Straight-realistic	283	762	2201	2629	3491	4457	10	9	11	13	11	15
		Straight-elliptical	283	761	2205	2630	3512	4493	10	6	14	12	7	11
		Straight-circular	284	768	2220	2642	3521	4502	12	14	12	8	10	8
	1D	Area functions	267	753	2220	2545	3529	4493	116	84	202	—	88	64

some antiresonances have disappeared when the side branches are removed. For instance, for vowel [ɑ] that close to 4 kHz is not present, as well as the strong antiresonances close to 6 kHz of vowels [i] and [u]. In addition, some formants have again been shifted. Yet, some antiresonances are still present such as the deepest drop close to 6 kHz of vowel [ɑ], although for the configuration without side branches it is not so prominent. In order to better understand its behaviour, the pressure distribution pattern for this antiresonance has been computed for each configuration (see Fig. 6). As expected, in the MRI case the strongest pressure values are produced within the side branches. However, when they are removed it can be observed that a transverse mode is also excited at this frequency, which contains a very low pressure area within the oral cavity at the transverse plane centered to the mouth exit. Therefore, one could then say, that this strong dip of vowel [ɑ] results from the combination of the effects of the side branches with this transverse mode, although the former

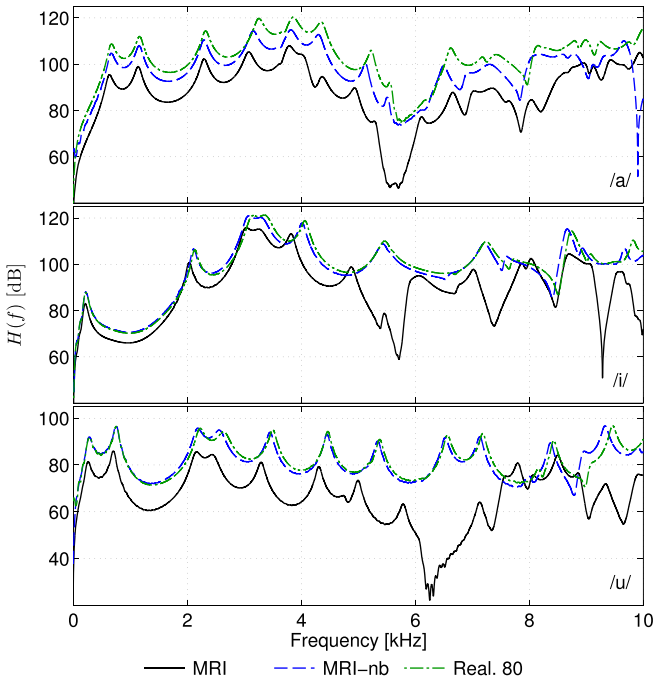

FIG. 5. (Color online) Vocal tract transfer functions for the MRI-based geometry, the MRI geometry without branches (MRI-nb), and the realistic case with 80 cross-sections.

seems to be the predominant one. Focusing now on the simplification with realistic cross-sections, note in Fig. 6 that it preserves the acoustic pressure distribution of this transverse mode. Indeed, it produces a very similar dip in the vocal tract transfer function to that of the MRI-nb case (see top of Fig. 5). Note that this is a particular case for the analyzed vowel [ɑ], since vowels [i] and [u] do not present an equivalent antiresonance around 6 kHz when the side branches are removed. For these vowel sounds, the oral cavity is narrower than for [ɑ] (see Fig. 1), which moves the cut-on frequency for non-planar mode propagation to higher frequencies. As a consequence, the first antiresonances for the MRI-nb case of [i] and [u] appear at higher frequencies than for [ɑ]. Note, however, that these are correctly reproduced when using realistic cross-sections, although some of them have slightly been shifted.

In general, some small differences can be observed in the vocal tract transfer functions between the realistic and the MRI configuration without side branches for all the

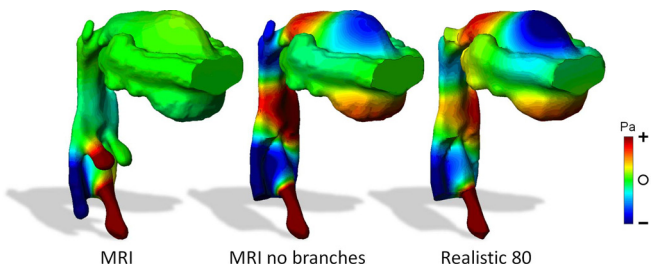

FIG. 6. (Color online) Snapshot of the acoustic pressure distribution for the stronger antiresonance located between 5–6 kHz within the MRI-based geometry, the MRI geometry without branches (piriform sinuses and vallecullae), and the simplification with 80 cross-sections with realistic shape in a bent configuration.

vowels. Despite them, it seems that this simplification with realistic cross-sections can emulate to a large extent the acoustics of the main conduct of the vocal tract.

B. Effects of the number of cross-sections

One of the parameters that needs to be determined when discretizing a 3D vocal tract geometry is the number of cross-sections. This is a problem that has already been addressed for classical 1D simulations, which make use of the so called vocal tract area functions. Typical values in the literature range between 40 and 80 cross-sections (for instance, 44 and up to 75 cross-sections for vowel sounds are, respectively, reported in Story, 2008, and Takemoto *et al.*, 2006). However, their influence seems to be not well established for 3D approaches. To examine this point, the realistic-bent configuration has been generated using 40, 60, and 80 cross-sections. The computed vocal tract transfer functions for vowels [ɑ], [i], and [u] are presented in Fig. 7 and compared to the MRI-nb cases.

Let us first focus on the low frequency region below 4–5 kHz. Looking at Fig. 7, the main effect that can be observed in the vocal tract transfer functions is a shifting of the formant frequencies for all the vowels. In general, the smaller the number of cross-sections the larger the deviation towards higher frequencies compared to the MRI-nb case (see also Table I). As far as the formant bandwidth is concerned, the differences produced by a different number of cross-sections do not follow any specific behavior and can be considered small (see Table I).

Focusing now on the frequency range above 4–5 kHz, the formants also shift to higher frequencies for all vowels when the number of cross-sections is reduced (see Fig. 7).

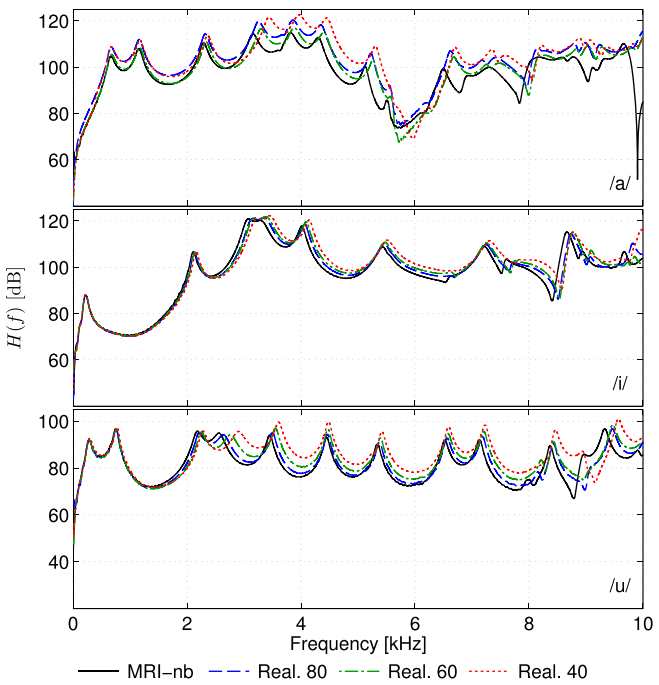

FIG. 7. (Color online) Effects of the number of cross-sections on the vocal tract transfer function $H(f)$. Comparison between the MRI geometry without branches (MRI-nb) and the realistic-bent case with 80, 60, and 40 cross-sections.

Moreover, some antiresonances have also shifted to higher frequencies, such as that close to 8 kHz for [ɑ] or those between 8 and 9 kHz for [i] and [u]. However, some antiresonances are not so influenced such as the strong drop close to 6 kHz of vowel [ɑ]. The 80 and 60 cross-section configurations seem to correctly locate this antiresonance when compared to the MRI-nb case, in contrast to the configuration with 40 cross-sections, which moves this antiresonance slightly to higher frequencies.

All in all, and as expected, the configuration that minimizes the deviations with respect to the case without branches is that of 80 cross-sections. In the following, simplifications will be directly constructed using this number of cross-sections.

C. Effects of cross-sectional shape in a bent configuration

Figure 8 shows the obtained vocal tract transfer functions when realistic, elliptical, and circular cross-sections are used to generate the vocal tract simplifications. The number of cross-sections is set to 80 for all configurations (see Sec. III B). The different types of cross-sections are then combined with the original vocal tract midline so as to obtain a bent configuration (see Fig. 1).

In the low frequency region below 4 kHz, plane wave propagation dominates and no important differences are observed when the cross-sectional shape is modified, independently of the vowel sound. Only some small formant deviations (see also Table I) are produced for vowel [ɑ] (e.g., the formants F4, F5, and F6) and for vowel [u] (e.g., the formant F6). These differences can be attributed to the bending phenomena of the propagating front waves produced at large area discontinuities (see, e.g., Kang and Ji, 2008), which should be

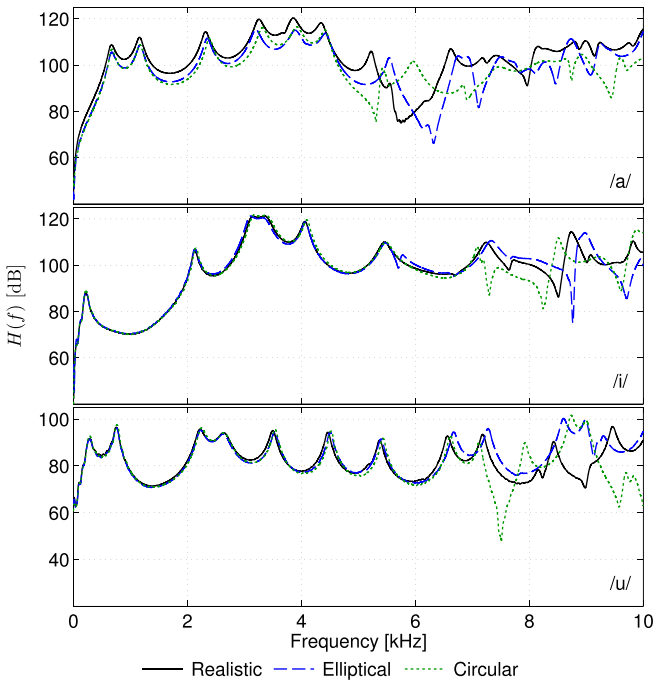

FIG. 8. (Color online) Effects of cross-sectional shape on the vocal tract transfer function $H(f)$. A bent vocal tract combined with realistic, elliptical and circular cross-sections is considered.

different depending on the cross-sectional shape. These phenomena are typically considered in 1D approaches by introducing inner-length corrections at the sudden expansions/constrictions of the vocal tract in order to correct the formant location (see, e.g., Sondhi, 1983, where this effect is approximated with expressions that consider circular cross-sections). Note that with 3D approaches this is naturally taken into account. As far as the formant bandwidths are concerned, results for the elliptical configuration are much closer to the realistic one than those of the circular configuration, for all of the analyzed vowels, obtaining in general smaller values for the circular cross-sectional shapes (see Table I).

In contrast, for frequencies beyond 4 kHz higher order modes become apparent, so not only resonances but also some antiresonances appear in the vocal tract transfer functions. As observed in Fig. 8, their behavior strongly depends on the cross-sectional shape. For vowel [ɑ], neither the elliptical case nor the circular approximation fit the results obtained for the realistic case in the high frequency range. However, the elliptical case seems to better approximate the realistic configuration compared to the circular vocal tract results. For instance, the seventh formant (∼5 kHz) and the strong antiresonance around 6 kHz of the realistic case are, respectively, replaced in the circular configuration by an antiresonance and a resonance, while they are still present in the elliptical case although shifted to higher frequencies. As far as the vowel [i] is concerned, the elliptical case generates a small antiresonance close to 6 kHz that is not present in the realistic and circular configurations. However, beyond 6 kHz it better fits the realistic case when compared to the circular one. Similar results are found for vowel [u]. In this occasion, no significant deviations are produced below 6 kHz, but for the frequency ranges up to 8 kHz, the elliptical case closely matches the realistic configuration, while the circular case introduces a strong dip. For frequencies beyond 8 kHz, none of them resemble the realistic case. As observed, the influence of cross-sectional shape is stronger for vowel [ɑ] than for vowels [i] and [u]. This can be attributed again to the large oral cavity of vowel [ɑ] (see Fig. 1), which produces a lower cut-on frequency of non-planar mode propagation. Moreover, the shape of the oral cavity for [ɑ] is more intricate than those of [i] and [u], which makes it more difficult to approximate this region by elliptical or circular cross-sections, which results in stronger variations.

D. Effects of bending with different cross-sectional shapes

The effects of bending on the vocal tract acoustics are next analyzed. Realistic, elliptical, and circular cross-sections are considered, which are combined with the original vocal tract midline to obtain bent vocal tracts and with a straightened midline to remove the bending (see Fig. 1). Eighty cross-sections are used for all simplifications (see Sec. III B). The obtained vocal tract transfer functions are presented in Fig. 9.

Concerning the low frequency region below 4 kHz, some formant shifts to lower frequencies are observed when the vocal tract is straightened, whatever the used

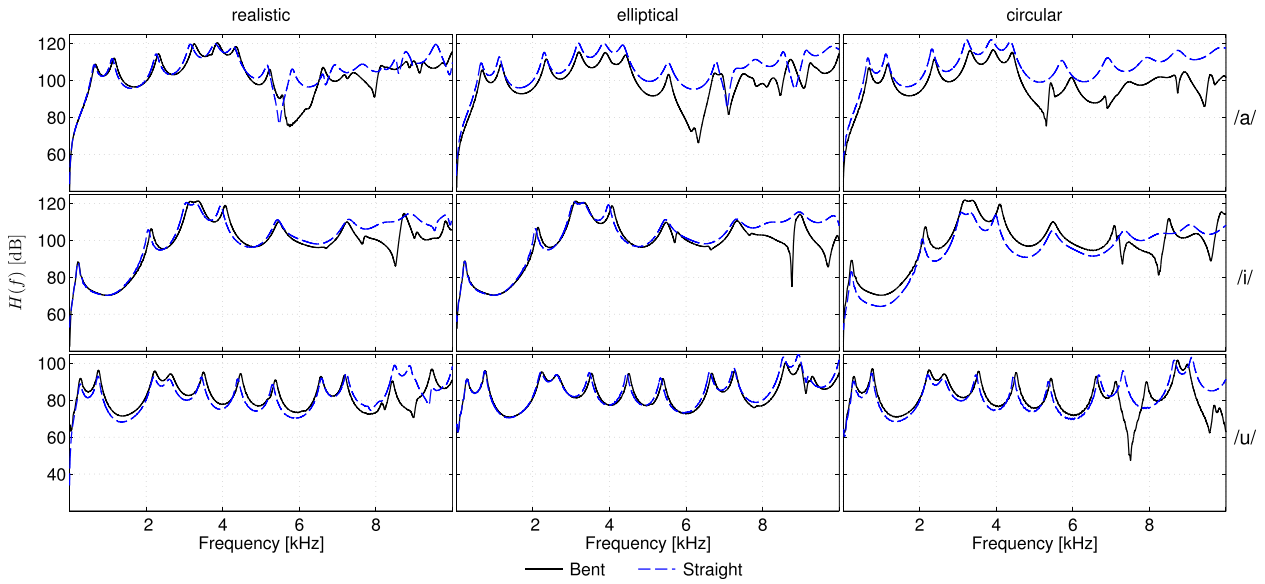

FIG. 9. (Color online) Effects of bending on the vocal tract transfer function $H(f)$. Realistic, elliptical, and circular cross-sections are considered and combined with a bent and a straightened vocal tract midline to generate each vocal tract geometry.

cross-sectional shape and analyzed vowel sound (see also Table I). This downward shift of the formant frequencies is in line with the observations reported by Motoki (2002) and also contradicts the analytical results obtained by Sondhi (1983) for bent rectangular tubes of constant cross-section, which presented some upward shifts compared to straight configurations. The differences in wavefront curvature between straight and bent vocal tracts may be a possible explanation for the observed deviations. Yet, they could also be attributed to the midline extraction procedure used to generate the straightened vocal tracts. Although a standard approach has been used for it, further work is probably still needed to improve its accuracy (see, e.g., Mochizuki and Nakai, 2007, where the centerline is extracted by using curved pressure contours computed by FEM instead of standard flat surfaces). However, this topic is out of the scope of this work.

In contrast, the effects of bending are more important for frequencies above 4 kHz (see Fig. 9). Focusing first on the circular case, only resonances appear in the transfer function when the vocal tract is straightened, while antiresonances are also present when it is bent, independently of the analyzed vowel. This is due to the radial symmetry of the straight circular vocal tract, which prevents the onset of higher order propagation modes (see also Blandin *et al.*, 2015). The use of elliptical shapes breaks this radial symmetry so that some higher order modes can also be excited in a straight vocal tract. Note, however, that this mainly occurs for vowel [ɑ], although some small variations can also be observed for vowels [i] and [u] above 8 kHz (e.g., the small dip for [u]). Some of these modes seem to match with the bent configuration (see, e.g., the antiresonance around 7 kHz for vowel [ɑ], and the small dip for [u] above 8 kHz). However, for vowels [ɑ] and [i] there are many other propagation modes that do not appear, such as the strong antiresonance close to 6 kHz of [ɑ] or those of vowel [i] above 8 kHz. For the particular case of vowel [u], no significant deviations are observed when bending is considered in the elliptical configuration. The complexity of the acoustic field logically increases when realistic cross-sections are used. The cut-on frequency of the non-planar propagation modes can be reduced compared to the elliptical case and, due to the complex shape of the vocal tract, a larger number of higher order modes can appear. This is the case of vowel [ɑ], which presents many high order modes even for the straight configuration. However, the matching of these modes with those of the bent configuration is very poor. For vowels [i] and [u] the influence of bending is weaker, which can be attributed again to the simplicity of their cross-sectional shape (see Sec. III C).

E. Comparisons with a 3D multimodal method and a 1D model

Finally, FEM results are compared to alternative approaches that can speed up numerical simulations, but that need to make use of some of the vocal tract geometry simplifications analyzed in this work. A 3D multimodal method and a classical 1D technique have been selected for this purpose. With regard to the multimodal method, the implementation in Blandin *et al.* (2015) has been followed and adapted to use cross-sections of arbitrary shape (in Blandin *et al.*, 2015, circular and elliptical cross-sections were considered). In such a situation the propagation modes cannot be obtained analytically, so the 2D Helmholtz equation for the cross-sectional shape was numerically solved with finite differences considering a hard wall boundary condition. The multimodal method can only consider straight vocal tracts (of arbitrary shape) and cannot account for wall losses, but it does account for 3D radiation losses. As far as the 1D model is concerned, a standard frequency-domain model based on transfer matrices has been used. The model implemented is similar to that of Sondhi and Schroeter (1987), but it has been adapted to consider the same wall losses as in FEM (see Sec. II C). This method requires vocal tract area functions, which are somewhat analogous to a 3D straight vocal

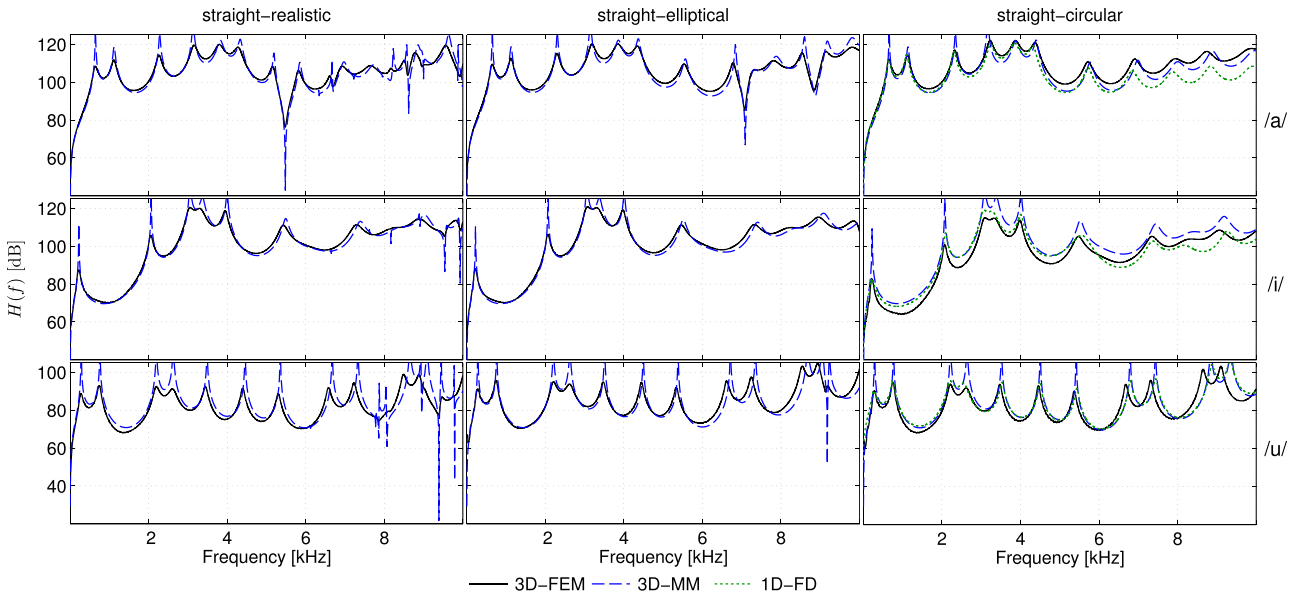

FIG. 10. (Color online) Comparison between the FEM approach and the multimodal method (MM) using a straightened vocal tract with realistic, elliptical, and circular cross-sections.

tract with circular cross-sections, as will be next observed. Therefore, only the straight vocal tract models will be considered in the comparisons. Figure 10 presents the vocal tract transfer functions computed using the multimodal method (3D-MM) for the straight vocal tracts with realistic, elliptical, and circular cross-sections, the equivalent curves for the finite element method (3D-FEM), and the results for the 1D frequency domain model (1D-FD).

As far as the 3D-MM is concerned, there is very good agreement between 3D-FEM and 3D-MM in all cases (see Fig. 10). The multimodal method is not only able to reproduce the formant resonances in the low frequency range (see also Table I), but also the higher order modes that appear above 5 kHz. However, the obtained formant bandwidths are underestimated (see Table I). Note also that the 3D-MM produces higher and sharper peaks in Fig. 10, since wall losses are not considered in the 3D-MM implementation. This also produces the onset of some small resonances and antiresonances in the high frequency range (see, e.g., the straight-realistic case for vowel [ɑ]) that are not present in 3D-FEM, which may have been mitigated if wall losses were taken into account.

Results for the 1D-FD are also shown in Fig. 10 but only in the figures corresponding to the straight-circular configurations. This simplified 3D model is the one that better resembles the behavior of a 1D model, since, as discussed in Sec. III D, its radial symmetry prevents the onset of higher order modes. This is confirmed when comparing 3D-FEM to 1D-FD over the whole examined frequency range. Very close curves are obtained, although some small discrepancies can be observed in the high frequency range. It is also remarkable that for comparison purposes, the levels of the 1D-FD curves were normalized to the first formant of 3D-FEM. Original 1D-FD curves had offsets of $\sim +16$, $+22$, and $+36$ dB for vowels [ɑ], [i], and [u], respectively. These increments were produced because 1D models can only capture the acoustic pressure within the vocal tract, while 3D models can also do it outside in the free-field space, the latter being the option adopted in this work. Smaller pressure values are obtained outside the vocal tract than inside, which justifies this increment. Moreover, the reported values seem to be reasonable when thinking in terms of radiated power. Vowel [ɑ] has the largest mouth aperture so it is the one that radiates more power, which results in smaller differences between the pressure inside and outside the vocal tract. The opposite occurs for [u], while for [i] we get an intermediate increment.

IV. CONCLUSIONS

In this work, finite element simulations have been conducted to analyze the acoustic response of several vowel vocal tract geometry simplifications. These consisted of cross-sections with realistic, elliptical, and circular shape interpolated through a bent or a straight vocal tract midline. These cross-sections were extracted from the main conduct of MRI-based vocal tracts for vowels [ɑ], [i], and [u]. The influence of discretizing the vocal tract geometry by a finite set of cross-sections, and the importance of the cross-sectional shape and vocal tract bending have been examined.

For frequencies below 4–5 kHz, the vocal tract shape and bending have shown a weak influence on the vocal tract acoustic response given that mainly plane waves propagate. Cross-sectional shape has hardly affected the formant locations ($<3\%$), although formant bandwidths were more sensitive to it showing a better performance for the elliptical shape compared to the circular one. Vocal tract bending has produced some formant shifts to lower frequencies, but all of them below 5%. The use of 80 cross-sections has also been recommended. Otherwise, some significant formant shifts above 2–3 kHz have been observed (up to 14% with 40 cross-sections). Despite of these deviations, at low frequencies all simplifications have shown a reasonable performance. Depending on the desired degree of accuracy one can choose any of them with good confidence.

The situation has become more intricate for higher frequencies. Both the vocal tract bending and cross-sectional shape have played a significant role in the correct generation of the higher order modes. Their onset depends on the analyzed vowel sound, being vowel [ɑ] the one showing the highest presence of these modes due to its large oral cavity. This has resulted in a stronger influence of the vocal tract bending and shape for this vowel sound, than for vowels [i] and especially [u]. Considering the most restrictive vowel, i.e., vowel [ɑ], the realistic shape has shown a good matching with the original MRI geometry, but neither the elliptical nor the circular shapes were able to correctly emulate the high frequency behavior, the former performing slightly better. The vocal tract bending has produced stronger variations than the vocal tract shape, which have been clearly exemplified with the circular configuration, in which no antiresonance was present for a straight vocal tract. At high frequencies it is then recommended to include bending and to consider realistic cross-sectional shapes. A perceptual evaluation would be valuable to evaluate the importance of the observed modifications in the high frequency range, which may extend some of the given recommendations to other vocal tract shapes (elliptical and/or circular). However, such a study lays out of the scope of the current work.

Finally, FEM results have been compared to those generated using a 3D multimodal model and a 1D frequency-domain method. 1D was found to correctly emulate the behaviour of a 3D straight-circular vocal tract, although some small variations were observed at higher frequencies. The 3D multimodal approach allowed us to also consider arbitrary cross-sectional shapes in a straight configuration and to capture the acoustic pressure outside of the vocal tract. A good matching was observed with FEM. However, formant bandwidths were underestimated because the multimodal approach does not consider wall losses in its current stage of development.

ACKNOWLEDGMENTS

This research has been supported by EU-FET Grant EUNISON 308874.

[1]A preliminary version of the work in this paper was presented in Arnela *et al*. (2015).

Aalto, D., Aaltonen, O., Happonen, R.-P., Jääsaari, P., Kivelä, A., Kuortti, J., Luukinen, J.-M., Malinen, J., Murtola, T., Parkkola, R., Saunavaara, J., Soukka, T., and Vainio, M. (**2014**). "Large scale data acquisition of simultaneous MRI and speech," Appl. Acoust. **83**, 64–75.

Arnela, M., Blandin, R., Dabbaghchian, S., Guasch, O., Alías, F., Pelorson, X., Van Hirtum, A., and Engwall, O. (**2016**). "Influence of lips on the production of vowels based on finite element simulations and experiments," J. Acoust. Soc. Am. **139**(5), 2852–2859.

Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Pelorson, X., and Van Hirtum, A. (**2015**). "Effects of vocal tract geometry simplifications on the numerical simulation of vowels," in *11th Pan-European Voice Conference (PEVOC)*, Florence, Italy.

Arnela, M., and Guasch, O. (**2013**). "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method," J. Acoust. Soc. Am. **133**, 4197–4209.

Arnela, M., and Guasch, O. (**2014**). "Two-dimensional vocal tracts with three-dimensional behaviour in the numerical production of vowels," J. Acoust. Soc. Am. **135**, 369–379.

Arnela, M., Guasch, O., and Alías, F. (**2013**). "Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations," J. Acoust. Soc. Am. **134**, 2946–2954.

Arnela, M., Guasch, O., Codina, R., and Espinoza, H. (**2014**). "Finite element computation of diphthong sounds using tuned two-dimensional vocal tracts," in *Proceedings of the 7th Forum Acousticum*, Kraków, Poland.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," J. Acoust. Soc. Am. **90**, 799–828.

Birkholz, P. (**2013**). "Modeling consonant-vowel coarticulation for articulatory speech synthesis," PLoS One **8**, e60603.

Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Van Hirtum, A., and Labal, X. (**2015**). "Effects of higher order propagation modes in vocal tract like geometries," J. Acoust. Soc. Am. **137**, 832–843.

Dabbaghchian, S., Arnela, M., and Engwall, O. (**2015**). "Simplification of vocal tract shapes with different levels of detail," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, Scotland, UK.

Doel, K. v. d., and Ascher, U. (**2008**). "Real-time numerical solution of Webster's equation on a nonuniform grid," IEEE Trans. Audio Speech Lang. Process. **16**, 1163–1172.

Engwall, O., and Badin, P. (**1999**). "Collecting and analysing two- and three-dimensional MRI data for Swedish," Tal Musik Hörsel Quart. Prog. Status Rep. Stockholm **3**, 11–38.

Fant, G. (**1970**). *Acoustic Theory of Speech Production*, 2nd ed. (Mouton, Paris), pp. 1–328.

Guasch, O., Arnela, M., Codina, R., and Espinoza, H. (**2015**). "Stabilized finite element formulation for the mixed convected wave equation in domains with driven flexible boundaries," in *Noise and Vibration: Emerging Technologies (NOVEM2015)*, Dubrovnik, Croatia.

Guasch, O., Arnela, M., Codina, R., and Espinoza, H. (**2016**). "A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production," Acta Acust. Acust. **102**, 94–106.

Kang, Z., and Ji, Z. (**2008**). "Acoustic length correction of duct extension into a cylindrical chamber," J. Sound Vib. **310**, 782–791.

Kelly, J., and Lochbaum, C. (**1962**). "Speech synthesis," in *Proceedings of the Fourth ICA* (Copenhagen, Denmark), pp. 1–4.

Kröger, B. J., Winkler, R., Mooshammer, C., and Pompino-Marschall, B. (**2000**). "Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results," in *5th Seminar on Speech Production*, pp. 333–336.

Mochizuki, K., and Nakai, T. (**2007**). "Estimation of area function from 3-D magnetic resonance images of vocal tract using finite element method," Acoust. Sci. Tech. **28**, 346–348.

Motoki, K. (**2002**). "Three-dimensional acoustic field in vocal-tract," Acoust. Sci. Tech. **23**, 207–212.

Rokkaku, M., Hashimoto, K., Imaizumi, S., Nimi, S., and Kirtani, S. (**1986**). "Measurements of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique," Ann. Bull. RILP **20**, 47–54.

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

Sondhi, M. M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Audio Speech Lang. Process. **35**, 955–967.

Speed, M., Murphy, D. T., and Howard, D. M. (**2013**). "Three-dimensional digital waveguide mesh simulation of cylindrical vocal tract analogs," IEEE Trans. Audio Speech Lang. Process. **21**, 449–455.

Story, B. H. (**2005**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**, 3231–3254.

Story, B. H. (**2008**). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J. Acoust. Soc. Am. **123**, 327–335.

Story, B. H. (**2013**). "Phrase-level speech simulation with an airway modulation model of speech production," Comput. Speech Lang. **27**, 989–1010.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.

Sundberg, J., Johansson, C., Wilbrand, H., and Ytterbergh, C. (**1987**). "From sagittal distance to area. A study of transverse, vocal tract cross-sectional area," Phonetica **44**, 76–90.

Švancara, P., and Horáček, J. (**2006**). "Numerical modelling of effect of tonsillectomy on production of Czech vowels," Acta Acust. Acust. **92**, 681–688.
Takemoto, H., Adachi, S., Mokhtari, P., and Kitamura, T. (**2013**). "Acoustic interaction between the right and left piriform fossae in generating spectral dips," J. Acoust. Soc. Am. **134**, 2955–2964.
Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am. **119**, 1037–1049.
Takemoto, H., Mokhtari, P., and Kitamura, T. (**2010**). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," J. Acoust. Soc. Am. **128**, 3724–3738.
Vampola, T., Horáček, J., and Švec, J. G. (**2008**). "FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels," Acta Acust. Acust. **94**, 433–447.
Vampola, T., Horáček, J., and Švec, J. G. (**2015**). "Modeling the influence of piriform sinuses and valleculae on the vocal tract resonances and antiresonances," Acta Acust. Acust. **101**, 594–602.