An animated realistic head with vocal tract for the finite element simulation of vowel /a/

Marc Arnella, Leonardo Pereira-Vivas, Jorge Egea
La Salle, Universitat Ramon Llull, C/ Quatre Camins 30, 08022 Barcelona, Catalunya (Spain)
marc.arnella@salle.url.edu, leonardo.pereira@salle.url.edu, jorge.egea@salle.url.edu

Abstract

Three-dimensional (3D) acoustic models can accurately simulate the voice production mechanism. These models require detailed 3D vocal tract geometries through which sound waves propagate. A few open source databases typically based on magnetic resonance imaging (MRI) are already available in literature. However, the 3D geometries they contain are mainly focused on the vocal tract and remove the head, which limits the computational domain of the simulations. This work develops a unified model consisting of an MRI-based vocal tract geometry set in a realistic head. The head is generated from scratch based on anatomical data of another subject, and contains different layers that add an organic appearance to the character. It is then not only designed to allow accurate finite element simulations of vowels, but more importantly, it can also be animated to add a realistic visual layer to the generated sound. This is expected to help in the dissemination of results and also to open potential applications in the audiovisual and animation sector. This paper shows the first results of the model focusing on the vowel /a/.

Index Terms: vocal tract acoustics, animation, human head, finite element method, vowels.

1. Introduction

When we speak, the vocal folds vibrate and generate sound waves that propagate through the vocal tract and emanate outside from the mouth. This complex physical phenomenon can be simulated using 3D acoustical models. The most extended approaches are those based on finite elements [1, 2, 3, 4, 5, 6], finite differences [7, 8], multimodal approaches [9, 10], and digital waveguide models [11]. All the above models require detailed geometries of the vocal tract to generate a sound. A few open source databases based on Magnetic Resonance Imaging (MRI) are already available in literature [12, 13]. However, they only include the vocal tract and remove the head. Accurate simulations can only be performed within the vocal tract, terminating the computational domain at the lips and imposing a zero pressure release condition [4] or more realistic boundary conditions based on analytical models such as the piston set in an infinite baffle [1]. It is also possible to extend the computational domain outwards to allow sound waves emanate from the mouth. The vocal tract is usually set in a flat baffle [14], which constitutes one side of the radiation computational domain. Perfectly Matched Layers [15], infinite elements [16], or Sommerfeld boundary conditions [4] are imposed on the other sides to absorb the incoming sound waves. However, this flat baffle is a rough approximation of the human head.

It is worth mentioning that in [17] a realistic head geometry was manually connected to a vocal tract to study the radiation effects in vowels. However, the geometry construction for each vowel required an extensive manual process, not feasible for the generation of dynamic sounds. Moreover, although the head was realistic, the vocal tract was simplified because there were not 3D vocal tract geometries available in literature at that time. 3D vocal tracts were generated using 1D vocal tract area functions [18] and taking elliptical cross-sections and an approximate vocal tract midline.

This work presents the first steps of a model that contains, in a unified domain, a realistic head connected with a 3D vocal tract generated from MRI. The model is designed to be easily coupled with 3D acoustic models, and allows one to automatically generate a large variety of geometries, both static and dynamic. But not only it generates realistic 3D geometries, but it also provides a realistic visual layer to the simulation results. The model contains textures and shaders that emulate the human skin, and hair simulation through Xgen and a Facial Action Coding System FACS-based rigging system [19]. This allows one to animate the model following correct anatomical concepts that add realism and even a certain expressiveness to the simulation. This work presents the first results using as example the vowel /a/. The FEM model in [4] is used for the acoustic simulations.

The work is structured as follows. Section 2 first presents the animated model of the head and vocal tract, it next provides some details about the animation procedure, and finally reviews the 3D finite element acoustic model. Section 3 shows the results obtained for the vowel /a/. Section 4 closes the work with the conclusions.

2. Methods

2.1. Realistic head and vocal tract model

The 3D vocal tract of vowel /a/ is based on the geometry in [12] generated from MRI. This geometry is slightly simplified as in [4] to facilitate the generation of dynamic sounds (in future works). In a nutshell, side branches such as the piriform fossae are first removed. Next, the vocal tract is discretized in a set of 40 cross-sections, and the shape of each cross-section is resampled to have 48 vertices. Although this work focuses on (static) vowels, dynamic sounds could then be easily generated by directly interpolating these cross-sections from the starting to the ending sound.

The realistic head model was created from scratch. The first step involves generating a 3D model through digital sculpting. This was done using the Zbrush program. In general, the sculpting represents the character’s skin. However, for correct anatomical proportions, it is necessary to create a skull model that guarantees the 3D placement of the vocal tract. Starting from the upper jaw and, precisely, the palatal vault, it defines the upper part of the oral cavity, while the lower jaw or mandible, articulated with the temporal bones of the skull, forms the lower part of this cavity [20]. The tubular space through which the vocal tract runs is defined from this point towards the trachea as it passes through the larynx. This allows
us to adapt the MRI models to the created face.

Next, the position of the vocal tract with respect to the lips is determined as in [17]. This is done through two planes created from the positions of 2 specific points in the lips. First, the outermost points for the upper and lower lips are identified. From these points, it is necessary to create a plane for each point, which must be inclined at 45°. These two planes are next cut. Their intersection determines the depth position of the vocal tract. The mouth opening is taken from the last section of the vocal tract. To join the vocal tract and head geometries, the last two cross-sections of the vocal tract are removed, and the vertices of the inner part of the lip are merged with the vocal tract section.

The digital sculpting of the head, however, results in a 3D high poly model with a high density of points that cannot be directly applied for animation. This geometry is thus lightened following a retopology process. This consists in redesigning the mesh of the 3D model following the facial patterns. This process is done using the Autodesk Maya program. These patterns are given as a consequence of the muscular movement of the face. In a retopology process, we first identify the muscles involved in this movement and the direction of the movement. From here, we create the patterns in the mesh, termed as loops, which are in charge of giving order to the facial deformation system to achieve the animation. The retopology is then a simplified and ordered model in coherent loops to obtain a light file that allows us to make animations. Figure 1 shows the resulting head and tract model.

In addition, for a correct operation with the finite element acoustic model, it is also necessary to take care that the model does not have intersections between its vertices, nor the overlapping of the faces of a polygon, to avoid problems in the simulations. This implies that, during this process, we obviate all the details of the face, such as wrinkles, skin pores, and others. These digital sculpting details are transformed into textures that are then applied to the model with the simplified topology.

A UV map is created for the correct placement of the texture on the face. UV mapping is the 3D modeling process of projecting a 2D image onto the surface of a 3D model [21]. From the 3D model, the seams are unfolded and placed flat on a 2D surface (this process is similar to the creation of seam patterns). Once the mapping is completed, an image of the face is fitted based on the “pattern” obtained (see Fig. 2). This allows the details of the face to be given to the model. This skin can finally be transferred to the head geometry, obtaining the realistic model shown in Figure 3.

2.2. Animation

To perform an animation, the geometry is prepared to apply a rigged system that allows the deformations of the face according to the muscular structure of the face. For this, based on FACS, the controllers are generated to guide the group of selected vertices. The essential points where the muscles are located are fixed to obtain and recognize the information of the key emotions [22]. These points are grouped into the nose, eyebrow tips, lip commissure, and eyelids.

The animation is done at 24fps and the opening of the mouth must match the initial opening of the 3D vocal tract model, given by the shape of the vowel /a/. For a readable and realistic animation, the face is divided into two parts: the upper mask of the face (eyes, eyebrows, and forehead) and the lower mask (nose, mouth, and jaw) (see Fig. 4). In this way, we can combine different expressions by combining the Action Units and reflecting different emotions in the character using the same sound.
2.3. Finite element simulations

Finite element simulations are performed in a computational domain $\Omega$ consisting of the 3D head geometry with the vocal tract and a radiation domain (see Fig. 5). The latter is made of two hemispheres of radius 0.2 m, 0.1 m apart, that are connected through a cylinder. $\Omega$ is meshed using tetrahedra of sizes ranging from 0.004 m within the vocal tract to 0.012 m within the radiation domain.

![Figure 5: Computational domain $\Omega$ used in the finite element simulation of vowel /A/. $\Gamma_G$ is the glottal cross-section, $\Gamma_W$ is the vocal tract wall, $\Gamma_H$ the head and $\Gamma_\infty$ the outer boundary of the radiation domain.](image)

The 3D acoustic model described in [4] is adopted for the finite element simulations. This model can generate vowels and diphthongs using realistic vocal tracts. In the particular case of vowels, the following problem is numerically solved

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

\[
\rho_0 \partial_t u + \nabla p = 0, \quad \text{in } \Omega, \ t > 0, \quad (1b)
\]

\[
p(x, 0) = 0, \ u(x, 0) = 0, \quad \text{in } \Omega, \ t = 0, \quad (1c)
\]

\[
u \cdot n = u_g(t), \quad \text{on } \Gamma_G, \ t > 0, \quad (1d)
\]

\[
u \cdot n = p/Z_w, \quad \text{on } \Gamma_W, \ t > 0, \quad (1e)
\]

\[
u \cdot n = p/Z_w, \quad \text{on } \Gamma_H, \ t > 0, \quad (1f)
\]

\[
u \cdot n = p/Z_0, \quad \text{on } \Gamma_\infty, \ t > 0, \quad (1g)
\]

where $p(x, t)$ is the acoustic pressure, $u(x, t)$ is the acoustic particle velocity, $\rho_0$ is the air density, $c_0 = 350$ m/s is the speed of sound, and $Z_0 = \rho_0 c_0$ is the air characteristic impedance. With regards to boundary conditions (see Fig. 5), a particle velocity $u_g(t)$ is imposed at the glottal cross-section $\Gamma_G$ to emulate the vocal cords, a wall impedance $Z_w = 83666$ kg/m$^2$s [23] is set at the vocal tract walls $\Gamma_W$ and head $\Gamma_H$, and a Sommerfeld boundary condition is prescribed on the outer boundary $\Gamma_\infty$ of the radiation domain to absorb sound waves reaching that boundary.

A vowel sound can be generated introducing a train of glottal pulses at $\Gamma_G$ and running an FEM simulation. In this work, however, it is generated through the convolution of the vocal tract impulse response with a glottal source excitation signal, which requires shorter FEM simulations. To do so we have proceeded as follows. First, the following Gaussian pulse is used for $u_g(t)$ in Eq. (1d)

\[
u_g(t) = e^{-[(t-T_{gp})/0.2\sigma_{gp}]^2} [\text{m}/\text{s}], \quad (2)
\]

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, $f_{max} = 10$ kHz. An FEM simulation is next performed with a sampling frequency of $f_s = 400$ kHz for a time period of 50 ms. The acoustic pressure $p_g(t)$ is tracked at the mouth exit. A vocal tract transfer function is next computed as

\[
H(f) = \frac{P_r(f)}{U_g(f)}, \quad (3)
\]

where $P_r(f)$ and $U_g(f)$ respectively stand for the Fourier transform of $p_g(t)$ and of the Gaussian pulse $u_g(t)$. The vocal tract impulse response is obtained as the inverse Fourier transform of $H(f)$, and finally used to generate the vowel sound by convolving it with a train of glottal pulses. The latter is generated using an LF model [4].

3. Results

Figure 6 shows the vocal tract transfer function $H(f)$ obtained for the vowel /A/. The typical formant frequencies of vowel /A/ are obtained below 5 kHz. Note that above this frequency the spectrum becomes more complex. This is because not only planar modes but also higher order modes propagate in this range. The large drop between 5-6 kHz is indeed originated by a transverse mode [2].

![Figure 6: Vocal tract transfer function for vowel /A/.](image)

![Figure 7: (a) Waveform and (b) spectrogram of the generated vowel /A/.](image)
The vocal tract transfer function $H(f)$ was next used to generate a vowel sound by convolving its inverse Fourier transform with a train of glottal pulses (see Section 2.3). The acoustic waveform of the generated vowel is depicted in Figure 7(a). A linear fade in and fade out was applied to the train of glottal pulses to emulate the onset and offset of the vocal cords. As observed, this effect is translated to the waveform signal. Also interesting is to examine the spectrogram (see Figure 7(b)). A preemphasis filter was applied in the spectrogram computation to improve the visualization of higher frequencies, as typically done in speech analysis. As expected, the frequency content does not change in time given that the vocal tract geometry is static. However, some slight fluctuations can be appreciated. These are produced because a pitch curve was applied to the glottal pulses to improve the naturalness of the generated sound. In what concerns the frequency content, observe that most of the energy is concentrated below 5 kHz, where plane wave propagation dominates. Between 5 and 6 kHz there is not energy because of the large drop observed in $H(f)$, and above 6 kHz some energy is observed again, although with a reduced level. This is in line with the observations reported in [24].

Finally, Figure 8 shows some snapshots of the character animated producing the vowel /a/. The first snapshot corresponds to the character producing the vowel /a/ in a frontal position. Observe that subsequent snapshots humanize the character by introducing a head movement together with a blink.

4. Conclusions

A unified model containing an MRI-based vocal tract and a realistic head has been constructed in this work. The vowel /a/ has been simulated as example, generating not only the corresponding sound but also a visual layer that contains the character animated. The face has been animated by humanizing the pronunciation of the vowel, through head movement, blinking and labial movement. Future work includes the synthesis of more static vowel sounds and the production of dynamic sounds like diphthongs. It is expected that the developed model could help in examining the acoustics of voice, such as directivity, but it may also facilitate the dissemination of results thanks to its visual layer. It also opens up potential applications in the audiovisual and animation sector, which may be interested in a solution that combines character animation with acoustics.

5. Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Innovation - Agencia Estatal de Investigación through the project FEMVoQ (PID2020-120441GB-I00 / AEI / 10.13039/501100011033). The authors would like to thank Gerard Grifell for his support and work in the development of the 3D head model.

6. References


Figure 8: Frame sequence of the final animation.


