Computational aeroacoustics strategies for the simulation of voice sequences involving sibilants

Oriol Guasch, Marc Arnela, Arnau Pont

COMPUTATIONAL AEROACOUSTICS STRATEGIES FOR THE SIMULATION OF VOICE SEQUENCES INVOLVING SIBILANTS

Oriol Guasch, Marc Arnela, Arnau Pont
GTM Grup de recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull, Barcelona, Spain
oriol.guasch@salle.url.edu

ABSTRACT

Realistic finite element (FEM) simulations of voice sequences containing sibilant consonants like /sa/, /za/ or /asa/, to mention a few, turns to be a very demanding problem from a computational point of view. The sole reproduction of a single sibilant implies resorting to computational aeroacoustics (CAA) strategies. One can either employ a direct numerical simulation (DNS) of the full compressible Navier-Stokes equations or use a hybrid approach in which the incompressible or compressible Navier-Stokes equations are first solved to obtain an acoustic source term from the flow jet motion. After that, an acoustic wave operator is used to synthesize the sibilant sound. If the sibilant is part of a voice sequence that includes vowels, the whole problem needs to be solved in a moving vocal tract (VT) geometry and a train of glottal pulses must be prescribed at the glottis for vowel generation. Dealing with dynamic VTs leads the computational cost of the problem prohibitive, even if one resorts to supercomputer facilities.

In this work, we review CAA approaches and some recently proposed alternatives to produce sibilants and voice sequences with sibilants at a reasonable computational cost. The key idea is to approximate the acoustic flow source term to avoid solving the Navier-Stokes equations in the first step of CAA. Two options are considered. The first one consists in using a random distribution of spinning Kirchhoff's vortices to emulate the quadrupole acoustic sources between the lower incisors and lower lips. The second one simply uses a white noise amplitude monopole to simulate the interdental flow jet and a dipole to account for incisors' diffraction effects. The wave equation in mixed form is solved to produce the final sounds. Numerical and implementation details are outlined, as well as the pros and cons of the proposed options. Some videos and audios of numerically generated voice sequences will be shown in the conference presentation.

1. INTRODUCTION

To produce sibilant sounds like /s/, the tongue blade is moved upwards to the alveolar ridge generating a small constriction where the air flow emanating from the lungs gets accelerated and develops into a turbulent jet. The latter passes through the inter-dental space between upper and lower incisors and hits the cavity in front of the lower lips. Turbulent eddies in these regions are responsible for the production of aerodynamic sound which gets diffracted by the upper incisors and radiated outwards (see Fig. 1). Some theoretical models have been developed to explain the basic mechanisms of sibilant sound production [1, 2]. In this work, however, we will concentrate on the three-dimensional (3D) numerical computations of sibilant sounds and vowel-sibilant sequences. Current approaches and future developments of realistic physical-based simulations will be reviewed.

2. SIMULATION OF SIBILIANTS IN STATIC GEOMETRIES

Given that turbulence is essentially a 3D phenomenon (no vortex stretching is allowed in 2D turbulence which yields to self-organized solutions), understanding the physics driving the generation of sibilant sounds implies simulating the glottal flow jet evolution in static 3D VTs. Despite some large eddy simulations (LES) of the flow dynamics inside VTs of /s/ have been performed in past years, it has only been until recently that the aeroacoustics of sibilant sounds have been reproduced with CAA strategies. For instance, in [3] the authors carried out a compressible LES simulation of the flow inside the VTs of an /s/ and a /ʃ/, from which they extracted both, the flow and the acoustic

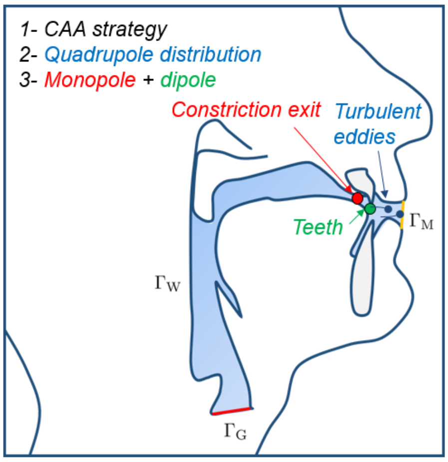

Figure 1: Strategies for simulating sibilant sounds like /s/: full CAA simulations, source approximation through turbulent eddy quadrupoles, source approximation through a monopole at the constriction exit and a dipole

fields. The authors also computed Lighthill's tensor approximation for low Mach numbers to determine the location of sibilant sound sources within the VT. 3D rectangular geometries were employed for the simulations which exhibit similar mean flow patterns than those of realistic 3D VTs. Comparisons were also performed with experiments. The closure of the compressible Navier-Stokes equations was achieved by means of an LES model with a one equation type subgrid-scale model. A finite volume method was implemented to solve the equations that was second order accuracy both in space and time. Use was made of the software OpenFOAM with a grid of 4.7×10^7 points. The results of the simulations agreed well with experiments and confirmed that the sound pressure spectrum of /s/ shows a characteristic peak around 5 kHz and exhibits a strong broadband content up to 15 kHz. The aeroacoustic sources of /s/ were mainly located at the gap between the upper and lower incisors. The VT was modified by shifting the tongue in the posterior direction to generate sibilant /ʃ/. This resulted in a cavity between the tongue and the lower lips where flow recirculation takes place diminishing the turbulence intensity in the interdental gap. As a consequence, the characteristic peak of /ʃ/ occurs at lower frequencies (close to 3 kHz) than that of /s/ and the high frequency content of the acoustic spectrum diminishes substantially.

Another work relying on CAA strategies is that in [4]. There, the authors resorted to an in-house developed finite element (FEM) code (FEMUSS from CIMNE-UPC) to simulate the production of sibilant /s/. The goal in that work was to locate the sound sources within the VT and determine the separate influence from the direct turbulent eddies' sound and from the aerodynamic sound diffracted by the incisors. To that goal the hybrid CAA strategy in [5] was followed. The incompressible Navier-Stokes equations were solved using an orthogonal subgrid-scale FEM, which acted as an implicit LES method and allowed one to use equal interpolating spaces for the hydrodynamic pressure and flow velocity. The acoustic source term was computed as the fluid simulation evolved and inserted in the linear acoustic wave equation to get the sound pressure at the far field. The acoustic pressure was split into a direct component and a diffracted one, giving rise to two coupled acoustic problems that provided the independent contributions of the sound generated by the turbulent flow and by its diffraction by the upper incisors. About 4.5×10^7 P1/P1 elements were used in the simulation and a third-order backward differentiation formula was implemented for the time discretization. As in [3], the results showed that the sound sources concentrate at the interdental gap but also in the cavity between the lower incisiors and the lower lips. As regards the sound pressure spectrum at the far field, it was shown to be dominated by the direct turbulent sound for frequencies $f < 2$ kHz, it had similar contributions from the direct and diffracted acoustic fields between 2 kHz< f <8-9 kHz, and it was driven by the latter for f >8-9 kHz.

The CAA approaches in [3, 4] can provide noticeable insight into the physics of sibilant sound generation. However, they involve huge computational models that need to be solved resorting to supercomputer facilities. If one aims at numerically generating audible voice sequences involving sibilant sounds in 3D VT geometries, the computational cost of CAA soon becomes prohibitive. It is therefore worthwhile exploring less demanding methods. Given that the bottle neck of CAA is the fluid dynamics computation due to the Navier-Stokes equations inherent non-linearity, a reasonable option would be that of emulating the flow acoustic source term. In this way, it only becomes necessary to solve the linear acoustic wave equation to produce sibilants. Different options have been recently proposed following this line of reasoning.

Taking advantage of Lighthill's acoustic analogy which interprets unbounded flow noise sources as distributions of quadrupole sources, the authors in [6] proposed to use a random distribution of Kirchhoff spinning vortices as a substitute for the source term obtained from fluid dynamic simulations in CAA hybrid approaches (see Fig. 1). The source distribution was designed so that its energy spectrum matches with that of the flow velocity of the incompressible LES in [4]. Therefore, and as said before, the production of /s/ only required solving the linear wave equation. The latter was actually expressed in mixed form to avoid the singular behavior of the source term and because that is needed when addressing the generation of vowel-sibilants sequences (see next section). With this strategy, the computational mesh only involved 7×10^5 P1/P1 elements instead of the 4.5×10^7 elements in [4]. A second order backward differentiation formula sufficed for the time discretization as only first order time derivatives appear in the wave equation in mixed form. Overall, that resulted in a drastic reduction of the computational cost and the possibility to generate long time sequences of sibilant /s/ and also /z/, by imposing a train of glottal pulses at the glottis of the VT computational domain.

An even simpler approach for the generation of sibilants within 3D VTs is inherited from 1D voice generation strategies. In the framework of the Ffowcs Williams-Hawkings equation, one can understand the turbulent noise production in a VT constriction as dominant contributions from monopole and dipole sources. The first one accounts for turbulent velocity fluctuations in the jet leaving the constriction, while the second considers the fluctuating forces the flow exerts on the incisors when impinging on them at almost normal direction [7–9]. 1D VT models have the severe limitation of plane wave propagation, which limits the frequency range up to $\sim$ 5kHz. However, one can consider the inclusion of white noise amplitude monopole and dipole sources in 3D VT geometries to cover frequencies up to 10 kHz or 15 kHz (see Fig. 1), where most acoustic energy of a sibilant sound like /s/ concentrates. That was the option chosen in [10] (monopole) and in [11, 12] (monopole plus dipole) In [12], a multimodal approach was used to simulate the generation of sibilant /s/ in a simplified 3D VT of rectangular cross-sections. Numerical results showed fairly good agreement with experiments, both

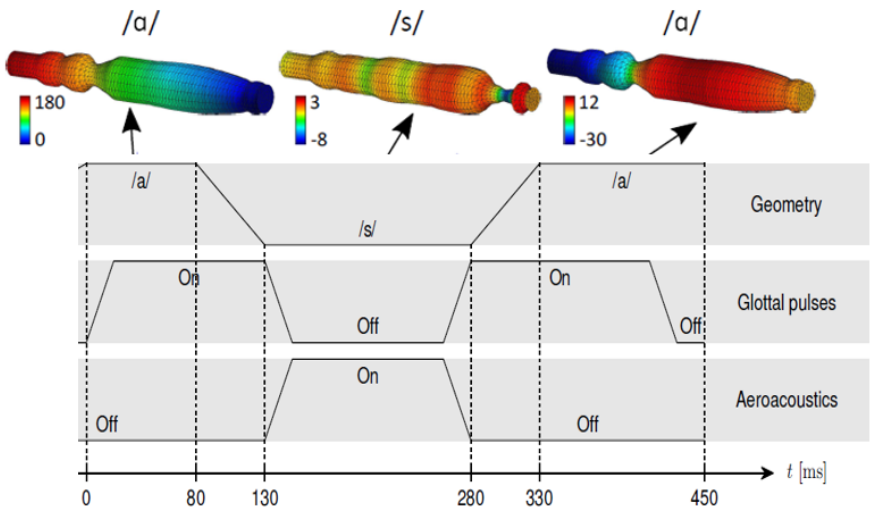

Figure 2: Generation of the vowel-sibilant-vowel sequence /asa/ using monopole and dipole sources to emulate fricative /s/.

in terms of acoustic pressure spectrum and directivity patterns. The approach in [11] will be reviewed in next section, dealing with the production of vowel sibilant voice sequences.

3. SIMULATION OF VOWEL-SIBILIANT SEQUENCES IN DYNAMIC GEOMETRIES

As mentioned in the previous section, the simulation of sibilants using CAA approaches which involve solving the Navier-Stokes equations become extremely costly. The situation worsens when one tries to generate vowel-sibilant sequences because the flow evolution has now to be solved in a VT with time varying geometry.

To date, and as far as we know, only the work in [13] has addressed such complex simulations. There, the authors were interested in analyzing the flow dynamics and noise generation when the tongue in a rectangular 3D VT was ascended and descended from the position of /u/ to /s/, at different speeds. To that aim, an implicit LES model was used for the compressible Navier Stokes equations, which were solved using a sixth-order accuracy finite difference scheme for the spatial derivatives and a third-order accuracy Runge-Kutta method for the time integration. A volume penalization method was employed as an immersed boundary method to account for the moving tongue. Most simulations were performed with a total number of 4.9×10^7 grid points. A simulation with a finer mesh of 10.3×10^7 was also carried out for validation.

The analysis of the tongue transition from /u/ to /s/ and then back to /u/ produced several interesting results. A significant one was an hysteresis effect on the generated aerodynamic sound. The overall sound pressure level (SPL) at a distant point from the mouth during the descent motion was higher than that of the ascent motion. The reason for it is that during the tongue ascent the flow in the constriction between the tongue and the hard palate gets accelerated and turbulence develop. When the tongue descends it takes some time for the turbulent eddies to dissipate, so more sound is generated than during the ascent. That was confirmed by comparison with experiments. The total duration of the tongue ascent and descent was ∼200 ms.

When it comes to the generation of an audible voice sequence involving sibilants, longer simulation times are usually required and several aspects need to be considered. As explained, resorting to the distribution of quadrupoles described in the previous section, or to simpler monopole and dipole sources can noticeably facilitate the computations. A first step in this line was recently presented in [11], where the sequence /asa/ was numerically simulated and the corresponding audio file was generated. The authors proceed by solving the wave equation in mixed form (i.e. the linear continuity and momentum equations for the acoustic pressure and acoustic particle velocity) in an arbitrary Lagrangian-Eulerian (ALE) frame of reference. This allows one to account for acoustic wave propagation inside a VT that evolves from the geometry of an /a/ to that of an /s/, and then get back to /a/. The equation was solved by means of the stabilized FEM strategy in [14]. As a first proof of concept, simple straight axisymmetric geometries were considered for the VTs. Linear interpolation was used to move from one VT geometry to another. The VT boundary evolution was therefore prescribed and the inner mesh modes were moved through diffusion, by solving a Laplacian equation for their displacements.

Different sound sources were considered. To generate vowel /a/, a train of glottal pulses was imposed as a boundary condition at the glottis Γ_G, while sibilant /s/ was simulated by inserting a point monopole at the constriction exit and a dipole in the axial direction at the teeth location (see Fig. 1). The monopole acted as a volume source term in the continuity equation and the dipole as a force one in the momentum equation. Both sources were assigned Gaussian white noise amplitudes.

The generation of /asa/ involves careful activation and deactivation of the sound sources, namely the train of glottal pulses to generate vowel /a/ and the monopole and dipole sources to produce sibilant /s/, in accordance with the VT evolution. This is represented in Fig. 2 extracted from the results in [11], where all fixed and transition times for sources and VT geometries are provided. Also, three snapshots of the VT geometries are shown together with their wall acoustic pressure distribution. As observed, a certain source overlap is imposed when transitioning from /a/ to /s/ and viceversa to get a more natural sound. The simulation in [11] only needed 1.5×10^4 elements and could be ran in laptop computer. As seen in Fig. 2, the simulated sequence /asa/ for which an audio file was recorded lasted 450 ms.

4. CONCLUSIONS

In this brief article, an overview has been given of current CAA methods to simulate sibilant sounds and voice sequences involving vowels and sibilants. CAA strategies that imply the simulation of the fluid dynamics provide deep insight on the underlying physics of sibilant sound generation, but are tremendously expensive in terms of computational cost. If one constrains to such approaches, research on different aspects of numerical voice generation may be slowed down. For instance, linking acoustic wave propagation inside a realistic 3D VT defined and driven by a 3D biomechanical model constitutes a big challenge in itself [15]. The use of approximations for the flow noise sources in terms of quadrupole distributions and/or point monopoles and dipoles can strongly facilitate the simulation of voice sequences including sibilants in such complex situations. Given that the overall sound spectrum of sibilants obtained from simplified flow noise sources capture the main characteristics of sibilants, using them opens the door to reliable 3D simulations up to 15 kHz, avoiding the plane wave limitation of 1D models. In this way, the high frequency energy content of vowel sounds and the high frequency dominant spectra of sibilants can be incorporated in the simulations, allowing for the numerical production of more natural vowel-sibilant voice sequences.

5. REFERENCES

[1] M. Krane, "Aeroacoustic production of low-frequency unvoiced speech sounds," *J. Acoust. Soc. Am.*, vol. 118(1), pp. 410–427, 2005.

[2] M. Howe and R. McGowan, "Aeroacoustics of [s]," *Proc. R. Soc. A*, vol. 461, pp. 1005–1028, 2005.

[3] T. Yoshinaga, K. Nozaki, and S. Wada, "Experimental and numerical investigation of the sound generation mechanisms of sibilant fricatives using a simplified vocal tract model," *Phys. Fluids*, vol. 30, no. 3, p. 035104, 2018.

[4] A. Pont, O. Guasch, J. Baiges, R. Codina, and A. Van Hirtum, "Computational aeroacoustics to identify sound sources in the generation of sibilant /s/," *Int. J. Numer. Method. Biomed. Eng.*, vol. 35, no. 1, p. e3153, 2019.

[5] O. Guasch, A. Pont, J. Baiges, and R. Codina, "Concurrent finite element simulation of quadrupolar and dipolar flow noise in low mach number aeroacoustics," *Comput. & Fluids*, vol. 133, pp. 129–139, 2016.

[6] A. Pont, O. Guasch, and M. Arnela, "Finite element generation of sibilants /s/ and /z/ using random distributions of kirchhoff vortices," *Int. J. Numer. Method. Biomed. Eng.*, vol. 36, no. 2, p. e3302, 2020.

[7] K. N. Stevens, *Acoustic phonetics*. The MIT Press, Cambridge, MA, 2000.

[8] S. Narayanan and A. Alwan, "Noise source models for fricative consonants," *IEEE transactions on speech and audio processing*, vol. 8, no. 3, pp. 328–344, 2000.

[9] P. Birkholz, D. Jackèl, and B. J. Kroger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1218–1226, 2007.

[10] T. Yoshinaga, A. Van Hirtum, and S. Wada, "Multimodal modeling and validation of simplified vocal tract acoustics for sibilant/s/," *J. Sound Vib.*, vol. 411, pp. 247–259, 2017.

[11] M. Arnela and O. Guasch, "Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 1802–1809, 2019.

[12] T. Yoshinaga, A. Van Hirtum, K. Nozaki, and S. Wada, "Acoustic modeling of fricative /s/ for an oral tract with rectangular cross-sections," *J. Sound Vib.*, p. 115337, 2020.

[13] T. Yoshinaga, K. Nozaki, and A. Iida, "Hysteresis of aeroacoustic sound generation in the articulation of [s]," *Phys. Fluids*, vol. 32, no. 10, p. 105114, 2020.

[14] O. Guasch, M. Arnela, R. Codina, and H. Espinoza, "A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production," *Acta Acust. united Ac.*, vol. 102, no. 1, pp. 94–106, 2016.

[15] S. Dabbaghchian, M. Arnela, O. Engwall, and O. Guasch, "Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model," *Int. J. Numer. Method. Biomed. Eng.*, p. e3407, 2020.