IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

GENIOVOX Project: Computational generation of expressive voice

Oriol Guasch, Francesc Alías, Marc Arnela, Joan Claudi Socoró, Marc Freixes and Arnau Pont*

GTM-Grup de Recerca en Tecnologies Mèdia, La Salle, Universitat Ramon Llull
C/Quatre Camins 30, 08022 Barcelona, Catalunya, España.

{oriol.guasch, *francesc.alias, marc.arnela, joanclaudi.socoro, marc.freixes, arnau.pont}@salle.url.edu

Abstract

The GENIOVOX project: "Computational synthesis of expressive voice", with ref. TEC2016-81107-P and funded by the Ministerio de Economía, Industria y Competitividad (Plan Nacional de I+D Excelencia) was carried out in the period 2016-2019. Its two main objectives were the following ones. On the one hand, diphthongs and hiatuses were simulated in three-dimensional (3D) geometries using the finite element method (FEM), based on the resolution of the underlying wave equations. Likewise, techniques were developed to simulate syllables with fricative consonants that did not require the use of high-performance computing. The trick was to approximate the interdental flow acoustic source terms using quadrupole, dipole and monopole distributions instead of getting them from a computational fluid dynamics simulation. In addition to generating diphthongs and syllables with fricatives, the project proposed a first attempt to incorporate some expressive effects through modifications of the vocal tract geometry and the glottal source model. Vowel sounds were computationally generated by convoluting the impulse response of 3D FEM vocal tracts with glottal pulses that incorporated tense, neutral and lax phonations from expressive speech corpora.

1. Introduction

Pronouncing a sound as simple as a vowel is extremely easy for us. However, in doing so we are unaware of the large number of physical phenomena involved. The turbulent flow of air exhaled from the lungs induces self-oscillations of the vocal cords. These generate acoustic waves that propagate inside the vocal tract and are emitted outwards. As we vary the shape of the vocal tract (VT) we will perceive one sound or another. At present, the numerical simulation of voice from the modeling of its underlying physics by means of realistic three-dimensional (3D) geometries of the VT is a notable challenge, making it necessary to resort to supercomputing centers to simulate certain sounds. For example, it is not possible to produce a fricative using the finite element method (FEM) with a desktop computer, although we can synthesize a vowel or a diphthong, with good equipment. On the other hand, so far, numerically generated sounds have lacked expressiveness. This aspect is not only important to correctly emulate communication between people, but it can also be related to certain medical issues. For instance, it is possible to detect pathological aspects of the voice from the sustained pronunciation of a simple vowel, or from a diphthong.

The two main contributions of the GENIOVOX project are as follows. To begin with, computational simulations of diphthongs and hiatuses have been carried out, enriching the number and type of computational sounds generated to date with dynamic 3D VTs. These sounds have been generated through interpolation between static vowel VT geometries and, in a more realistic way, by resorting to the biomechanical model ArtiSynth (https://www.artisynth.org/). More importantly, methods for synthesizing fricatives in 3D geometries that do not require fluid dynamics calculations in supercomputing centers have been developed. Aerodynamic sources of noise have been simulated using quadrupole and/or dipole and monopole source distributions, inspired to some extent by 1D-based approaches. This has allowed us to generate syllables and sequences such as /sa/ and /asa/ for 3D VTs with desktop computers and at a cost not much higher than that of generating diphthongs.

The second major contribution of the GENIOVOX project is that, for the first time, the possibility of including some expressive effects on FEM generated voice has been attempted. Such effects can be achieved from changes in phonation and/or in the VT geometry. Both aspects have been taken into account. On the one hand, the tense, modal and lax voice continuum in the FEM generation of vowels has been analyzed using simplified and realistic 3D geometries of the vocal tract. Also, numerical methods have started being developed to modify the geometry of the vocal tract in order to achieve effects such as the grouping of the formants of a vowel (i.e. the so-called singing formant), an important element, for example, to simulate the projection of the sung voice.

2. Goals and main results

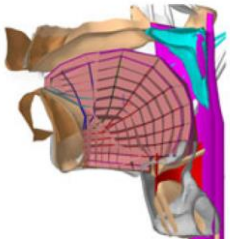
The six main objectives of the GENIOVOX project and the results achieved during its execution are summarized below.

2.1. O1: Characterization of expressive parameters in recorded voice

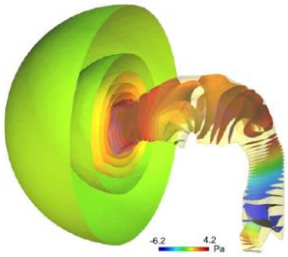
The GTM has five voice corpora corresponding to the following expressive styles: neutral, happy, aggressive, sad and sensual. Moreover, we analyzed a storytelling corpus containing diverse expressive categories. The analyses focused on the set of words that appear in all corpora, and more specifically, on the vowels. Prosodic parameters (fundamental frequency, duration and energy), perturbation parameters (jitter and shimmer) and several voice quality parameters (e.g., harmonic-to-noise ratio, Hammarberg index, spectral slope, etc.) were extracted to characterize the main components of expressive speech.

On the other hand, a proposal was developed to add singing capabilities to a neutral corpus-based text-to-speech synthesis system. A text-to-speech-and-singing synthesis framework that integrates speech-to-singing conversion was developed and tested, achieving reasonable quality compared to corpus-based singing approaches to eventually address singing needs in storytelling applications according to the conducted MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) perceptual test.

10.21437/IberSPEECH.2021-33



a)



b)

Figure 1: *a) Obtaining the vocal tract for the vowel /a/ from the ArtiSynth biomechanical model. b) Propagation of acoustic waves inside the vocal tract of an /a/, obtained with FEM. Image published in [5].*

2.2. O2: Simulation of diphthongs / hiatuses

This objective focused on the simulation of the diphthongs /ai/, /au/ and /ui/ in 3D dynamic geometries of the VT by means of FEM. This requires solving the wave equation in a domain that evolves from the geometry of the initial vowel to that of the final vowel. Two different strategies were developed for this purpose. First, static geometries generated from MRI (Magnetic Resonance Imaging) were discretized using an adaptive grid and a semi-polar grid, which allowed us to simplify the interpolation between 3D meshes. Second, a methodology was developed to extract closed cavities from the ArtiSynth biomechanical model and thus unify, for the first time, biomechanical and acoustic simulations in three dimensions. All the acoustic simulations were carried out by means of an in-house developed FEM code (see Figure 1).

2.3. O3: Simulation of syllables with fricative consonants

In this objective, the fricatives /s/ and /z/ were simulated with FEM. On the one hand, some computational aeroacoustics (CAA) simulations that required the use of supercomputing were completed. On the other hand, and given that precisely the computational cost of generating a syllable that contains a fricative is prohibitive, even in a supercomputing center, a method was developed to avoid the computational fluid dynamics (CFD) calculation of the simulation, which allows treating only with the wave equation. The contribution consists in approximating the source term that we would obtain from the CFD by means of a random distribution of Kirchhoff vortices (see Figure 2). In this way, the cost of synthesizing a syllable like /sa/ is not excessively higher than that of generating a diphthong. As a third alternative, monopolar and dipole sources excited using Gaussian noise were implemented. This allowed us to synthesize the sequence /asa/ with simplified 3D geometries.

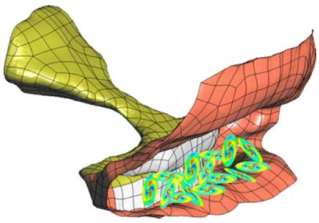


Figure 2: *Key idea for the generation of fricatives without CFD calculation: approximation of the flow source term by means of a random distribution of Kirchhoff's vortices.*

2.4. O4: Variation and adjustment in time of the geometry of the vocal tract

Varying the shape of the vocal tract to, say, concentrate formants and thus achieve certain expressive effects on the voice, is usually carried out in the context of 1D articulatory synthesis using sensitivity functions. These are justified from the non-linear phenomenon of radiation pressure. Throughout the project, some theoretical advances in this line were carried out. In particular, it was shown that the sensitivity functions can be obtained from a perturbation analysis of the VT eigenmodes, without the need to resort to the non-linear phenomenon of radiation pressure to justify them. This totally different theoretical approach is more prone to updating techniques in FEM and will facilitate optimizing 3D VT shapes to achieve expressive effects in the future.

2.5. O5: Glottal pulse modification

Throughout the project, the LF (Liljencrants-Fant) glottal pulse model was used to generate the excitation signal. This parametric model allowed us to modify the shape of the pulses using the control parameter Rd, in order to simulate lax, modal and tense phonations. To that goal, an aliasing-free LF model was adapted incorporating the control parameter Rd. Furthermore, this implementation was extended by integrating an aspiration model to emulate the aspiration noise found in vowels. This allowed us to obtain more realistic glottal pulses and generate more natural voice. The modification of the glottal pulses using the herein detailed method was evaluated in the generation of different vowels. The frequency responses obtained from a realistic geometry and a simplified one by FEM were used. The evaluation was made from the spectral analysis of the results and by quantifying the high frequency energy content.

2.6. O6: Numerical generation of expressive voice

The first steps were taken for the inclusion of expressive effects in the generation of computational voice. Specifically, simplified and realistic 3D vowel tracts were considered for the vowels /a/, /i/ and /u/. The former were based on models with radial symmetry while the latter consisted of realistic MRI-based models. The impulse responses of the vocal tracts were calculated by FEM and then convoluted with the glottal pulses corresponding to the tense, neutral and lax phonations (see Figure 3). From there, the influence that the type of phonation has on the high-frequency energy content of the different vowels according to the considered geometries was analyzed.

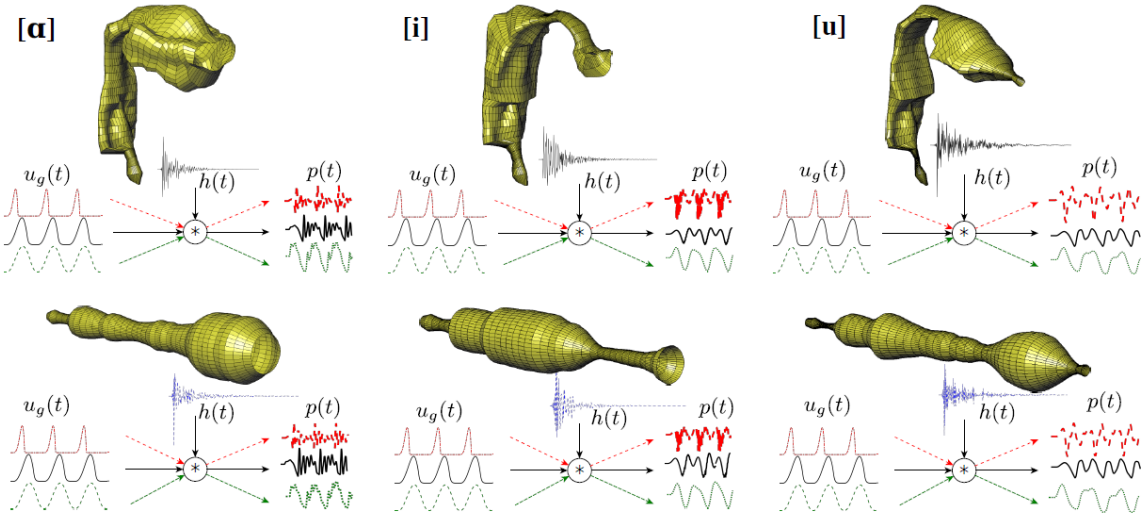


Figure 3: *Synthesis of the vowels /a/, /i/ and /u/ with realistic and simplified vowel tracts. The acoustic signal p(t) is calculated from the convolution of the glottal source $u_g(t)$ with the impulse response h(t) obtained from the 3D finite element model. Three types of phonation are considered: Tense (dashed red line), modal (black line) and lax (dashed green line). Image published in [6].*

2.7. O7: Evaluation of the quality of the generated voice

From the voice corpus analyzed in O1, the three corresponding to modal and tense voice were chosen from neutral, happy and aggressive styles, respectively. Parallel realizations in these three expressive styles were analyzed using a glottal vocoder that decomposes the speech signal into glottal excitation and VT response, and independently parameterizes them. From these parameters, two expressive conversion models (LF-based glottal excitation and VT) were trained. The neutral words were converted to the happy and aggressive styles using expressive prosody and, through different configurations of the conversion models, the contribution of the spectral characteristics of arousal and the vocal tract in the generation of a happy and aggressive voice was studied. The objective evaluation focused on the spectral analysis of the results and on the calculation of spectral distances between each configuration and the expressive reference on vowels using carrier words. The subjective evaluation was carried out using a MUSHRA perceptual test.

3. Publications

As a result of the project, 11 articles were published in journals from different fields (from numerical methods to acoustics and speech processing). Likewise, 13 papers were presented at several international conferences such as InterSpeech2017, ISSP2017, ECCM-ECFD2018, IberSPEECH2018, Inter-Noise2019, ECCOMAS-YIC2019, ICA2019, InterSpeech2019 and SSW10, and 1 book chapter was also published. The list of all publications is presented in the References' section of this article.

Finally, it is also worth mentioning that the PhD Thesis of Marc Freixes has been partially developed within the GENIOVOX project, under the supervision of Dr. Francesc Alías and Dr. Joan Claudi Socoró.

4. Conclusions

The GENIOVOX project has constituted a first step towards the numerical generation of diphthongs, hiatuses and fricative consonants through 3D FEM without resorting to supercomputer facilities, as well as a preliminary attempt to introduce expressiveness in vowels, specifically, modal and tense phonation types extracted from neutral and happy plus aggressive vowels.

In future works we aim at developing finite element computational strategies on dynamic 3D VTs to simulate spoken utterances containing fricative sounds as well as velar and bilabial stops. Moreover, we will keep working towards improving the naturalness of the generated expressive voice by properly modifying the glottal flow model and the VT shape to generate different voice qualities and vocal effects such as the Lombard effect or the singing formant, considering also inverse filtering techniques applied to expressive speech corpora.

5. Acknowledgments

The authors acknowledge the Agencia Estatal de Investigación (AEI) and FEDER, EU, for funding the project GENIOVOX (ref. TEC2016-81107-P).

6. References

Journal papers

[1] R. Montaño and F. Alías; "The Role of Prosody and Voice Quality in Indirect Storytelling Speech: A cross-narrator perspective in four European Languages," *Speech Communication*, vol. 88, pp. 1-16, 2017.

[2] M. Arnela and O. Guasch, "Finite element synthesis of diphthongs using tuned two-dimensional vocal tracts," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25 (10), pp. 2013-2023, 2017.

[3] A. Pont, R. Codina, J. Baiges and O. Guasch, "Unified solver for fluid dynamics and aeroacoustics in isentropic gas flows," *Journal of Computational Physics*, 363, pp 11-29, 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/," *International Journal for Numerical Methods in Biomedical Engineering*, 35 (1), e3153, pp. 1-17, 2019.
[5] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch, "Reconstruction of vocal tract geometries from biomechanical simulations," *International Journal for Numerical Methods in Biomedical Engineering*, 35 (2), e3159, pp. 1-19, 2019.
[6] M. Freixes, M. Arnela, J.C. Socoró, F. Alías and O. Guasch, "Glottal source contribution to higher order modes in the finite element synthesis of vowels," *Applied Sciences*, 9 (21), 4535, pp. 1-12, 2019.
[7] M. Arnela, S. Dabbaghchian, O. Guasch and O. Engwall, "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 27 (2), pp. 2173-2182, 2019.
[8] M. Freixes, F. Alías and J.C. Socoró, "A Unit Selection Text-to-Speech-and-Singing Synthesis Framework from Neutral Speech: Proof of concept," *EURASIP Journal on Audio, Speech, and Music Processing*, 2019:22, pp. 1-14, 2019.
[9] A. Pont, O. Guasch and M. Arnela, "Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff's vortices," *International Journal for Numerical Methods in Biomedical Engineering*, 36 (2), e3302, pp. 1-20, 2020.
[10] O. Guasch, M. Arnela and A. Pont, "Resonance tuning in vocal tract acoustics from modal perturbation analysis instead of nonlinear radiation pressure," *Journal of Sound and Vibration*, 493, 115826, 2021.
[11] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch (2021), "Synthesis of vowels and vowel-vowel utterances using a 3D biomechanical-acoustic model," *International Journal for Numerical Methods in Biomedical Engineering*, *Accepted*, 2021.

Conference articles (A) and presentations (P)

[12] M. Arnela, S. Dabbaghchian, O. Guasch and O. Engwall, "A semi-polar grid strategy for the three-dimensional finite element simulation of vowel-vowel sequences", *Interspeech 2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[13] N.C. Degirmenci, J. Jansson, J. Hoffman, M. Arnela, P. Sánchez-Martín, O. Guasch and S. Ternström, "A unified simulation of vowel production that comprises phonation and the emitted sound," *Interspeech 2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[14] S. Dabbaghchian, M. Arnela, O. Engwall and O. Guasch, "Synthesis of VV utterances from muscle activation to sound with a 3D model", Interspeech *2017*, August 20-24, Stockholm, (Sweden), 2017. (A)
[15] O. Guasch and S. Ternström, "Some current challenges in unified numerical simulations of voice production: from biomechanics to the emitted sound," ISSP2017, the 11th International Seminar on Speech Production, October 16-19, Tianjin, (China), 2017. (A)
[16] R. Codina, A. Pont, J. Baiges and O. Guasch, "Split boundary conditions for computational aeroacoustics of isentropic flows," *6th European Conference on Computational Mechanics and 7th European Conference on Computational Fluid Dynamics (ECCM-ECFD 2018)*, June 11-15, Glasgow (UK), 2018. (P)
[17] M. Freixes, M. Arnela, J.C. Socoró, F. Alías and O. Guasch, "Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]," *IberSpeech2018*, November 21-23, Barcelona, Catalonia (Spain), 2018. (A)
[18] O. Guasch, M. Arnela and A. Pont, "Modal perturbation analysis instead of nonlinear radiation pressure to derive the area sensitivity function for resonance tuning in an axisymmetric duct with variable cross-section," *InterNoise2019*, June 16–19, Madrid, (Spain), 2019. (A)
[19] A. Pont, O. Guasch and M. Arnela, "Modal perturbation analysis instead of nonlinear radiation pressure to derive the area sensitivity function for resonance tuning in an axisymmetric duct with variable cross-section," *InterNoise2019*, June 16–19, Madrid, (Spain), 2019. (A)
[20] A. Pont, M. Arnela and O. Guasch, "Finite element generation of vowel-sibilant utterances using random distributions of Kirchhoff's vortices and simplified vocal tract geometries", *ECCOMAS YIC*, September 1-6, Krakow (Poland), 2019. (P)
[21] M. Arnela and O. Guasch, "Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model," *ICA 2019, 23th International Congress on Acoustics*, September 9-13, Aachen (Germany), 2019. (A)
[22] O. Guasch, Survey lecture on "Realistic physics-based computational voice production," Co-contributors: Marc Arnela, Arnau Pont, Francesc Alías, Marc Freixas and Joan-Claudi Socoró, *Interspeech 2019*, September 15-19, Graz (Austria), 2019. (P)
[23] M. Freixes, M. Arnela, F. Alías, J.C. Socoró, "GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]," Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW10), pp. 132-136, 20-22 September 2019, Vienna, Austria, 2019. (P)
[24] Marc Arnela, Oriol Guasch, Arnau Pont (2020), "Tuning MRI-based vocal tracts to modify formants in the three-dimensional finite element production of vowels", Proc. of 12th International Conference on Voice Physiology and Biomechanics (ICVPB), March Edition, Grenoble, France. (A)

Book Chapters

[25] O. Guasch, A. Pont, J. Baiges and R. Codina, "Simultaneous Finite Element Computation of Direct and Diffracted Flow Noise in Domains with Static and Moving Walls", in: Ciappi E. *et al*. (eds) Flinovia—Flow Induced Noise and Vibration Issues and Aspects-II. FLINOVIA 2017 (ISBN: 978-3-319-76779-6), pp. 179-194, Springer, Cham Hastie, 2019.

PhD Thesis

[26] M. Freixes. "Adding expressiveness to unit selection speech synthesis and to numerical voice production", PhD Thesis, La Salle – Universitat Ramon Llull, 2021.