Speech Communication: Paper ICA2016-699

Copy synthesis of running speech based on vocal tract imaging and audio recording

Benjamin Elie[(a)], Yves Laprie[(a)]

[(a)]Loria, Inria/CNRS/Université de Lorraine, Nancy, France, benjamin.elie@loria.fr

August 29, 2016

Abstract

This study presents a simulation framework to synthesize running speech from information obtained from simultaneous vocat tract imaging and audio recording. The aim is to numerically simulate the acoustic and mechanical phenomena that occur during speech production given the actual articulatory gestures of the speaker, so that the simulated speech reproduces the original acoustic features (formant trajectories, prosody, segmentic phonation, etc). The result is intended to be a copy of the original speech signal, hence the name copy synthesis. The shape of the vocal tract is extracted from 2D midsagittal views of the vocal tract acquired at a sufficient framerate to get a few images per produced phone. The area functions of the vocal tract are then anatomically realistic, and also account for side cavities. The acoustic simulation framework uses an extended version of the single-matrix formulation that enables a self-oscillating model of the vocal folds with a glottal chink to be connected to the time-varying waveguide network that models the vocal tract. Copy synthesis of a few French sentences shows the accuracy of the simulation framework to reproduce acoustic cues of natural phrase-level utterances containing most of French natural classes while considering the real geometric shape of the speaker. This is intended to be used as a tool to relate the acoustic features of speech to their articulatory or phonatory origins.

Keywords: Copy synthesis

Copy synthesis of running speech based on vocal tract imaging and audio recording

1 Introduction

Copy synthesis [9] consists in numerically reproducing the physical and acoustical phenomena that occur during a natural utterance. Thus, the numerical simulation gives access to physical quantities that are hardly experimentally observable, such as the pressure distribution along the vocal tract, the oscillation cycle of the vocal folds, etc. The availability of such simulation frameworks may be useful for speech researchers to study the relationships between the acoustic features that are observed in natural speech and the corresponding articulatory and phonatory configurations of the speaker.

The speech production needs to be modeled at several distinct levels. An articulatory [11, 20] model of the vocal tract should reproduce the complexity of its shapes and its deformation over time with a few parameters. A glottal source model [5, 6, 15] should reproduce the mechanical behavior of the vocal folds, namely the self-oscillating cycle of the vocal folds. An acoustic wave simulation framework [12, 20] is then used to solve the equations driving the acoustic propagation along the vocal tract, yielding to a simulated speech signal that reproduces the acoustic features of the natural speech.

In the presented study, the shape of the vocal tract is derived from Xray imaging [10]. This consists in sequences of 2D midsagittal slices of the vocal tract acquired at a sufficient framerate to guarantee at least 3 images by phoneme. Area functions are then derived from the articulatory contours extracted from the image sequence and corrections are applied afterwards using the corresponding audio recording. The method is described in Sec. 3.1. The glottal source is computed via a self-oscillating model of the vocal folds, derived from classic mass-spring lumped systems [15], to which a glottal chink is branched in parallel [4]. The possibility of connecting a glottal chink has been proven to be interesting for simulating some features of natural speech, such as breathy voice [21], or voiced fricatives [5]. The model is presented in Sec. 2.1. The acoustic propagation, described in Sec. 2.2, is simulated via the *Transmission Line Circuit Analog* (TLCA) model [12] including recent reformulations and extensions [4, 14]. The choice of TLCA is motivated by the fact that it easily deals with geometrical variations of the vocal tract over time, including length variations, which is not the case for the other widely used *Reflection Type Line Analog* model [7].

2 Acoustic model

2.1 Glottal source

2.1.1 Self-oscillations of the vocal folds

The acoustic model considers an unsteady inviscid and incompressible flow through the glottis, corrected with terms accounting for viscous losses [1]. The pressure distribution in the glottal constriction, along the x axis, is

$$\begin{array}{rcll} P(x) & = & P_{sub} - \frac{\rho U_g^2}{2l_g^2}\left[\frac{1}{h^2(x)} - \frac{1}{h^2(x_0)}\right] - \frac{12\mu U_g}{l_g}\int_{x_0}^{x}\frac{dx}{h^3(x)} - \rho\frac{\partial}{\partial t}\left[\frac{U_g}{l_g}\int_{x_0}^{x}\frac{dx}{h(x)}\right] & x < x_s \\ P(x) & = & P_{sup} & x > x_s, \end{array} \tag{1}$$

where l_g is the length of the vocal folds, ρ and μ are respectively the mass density and the shear viscosity of the air, and x_s is the position of the mobile separation point, as defined in [15].

Considering a classic lumped mass-spring system for this study, the geometry of the glottal constriction, defined by the positions of the vocal folds, is computed at each simulation step according to the system of differential equations

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{R}\dot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{F}, \tag{2}$$

with $\mathbf{M} \in \mathbb{R}_+^{4\times4}$, $\mathbf{R} \in \mathbb{R}_+^{4\times4}$, $\mathbf{K} \in \mathbb{R}^{4\times4}$, and $\mathbf{F} \in \mathbb{R}^{4\times4}$ are matrices containing the values of respectively the mass, the damping, the stiffness and the pressure forces applied to each mass, and $\mathbf{y} \in \mathbb{R}^4$ is the vector containing the displacement of each mass from its rest position.

The model that is used in this paper is a 2×2-mass with smooth contours, as in [15]. It is extended to include a glottal chink [5], acting as a "zip-like" structure, represented by an acoustic waveguide branched in parallel to the oscillating part of the vocal folds. The glottis is then divided into two component: the posterior part, composed by the glottal chink, and the anterior part, that acts as a self-sustaining oscillator.

2.2 Acoustic propagation in the vocal tract

The model for the acoustic propagation is based on the *transmission line circuit analog* (TLCA) method [12], and the more recently improved *single-matrix formulation* [14]. These methods are based on an electric-acoustic analogy, where the elementary acoustic tubelets that model the vocal tract (VT) are seen as lumped circuit elements.

The frication noise is generated via the activation of random pressure sources inside each tubelet. The amplitude of the noise source P_{n_i} at section i is

$$P_{n_i} = \max\left\{0, \xi w\left(Re^2 - Re_c^2\right)\frac{U_{DC}^3}{a_{i-1}^{3/2}}\right\}, \tag{3}$$

where ξ is an arbitrarily adjustable real constant used to control the noise level and w is a random number taken in the range $[0,1]$. Re is the Reynolds number of the air flow inside the VT, Re_c is an arbitrary threshold above which the frication noise is generated [18], U_{DC} is the low-frequency component of the air flow inside the VT, and a_{i-1} is the area of the upstream tubelet. For this study, Re_c is set to 1700.

3 Data

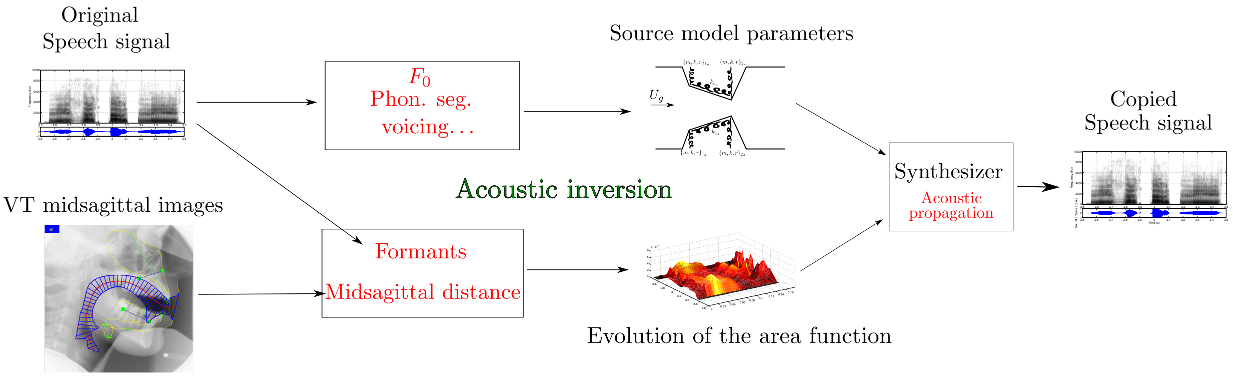


Figure 1: Simulation framework for copy synthesis.

Fig. 1 summarizes the presented copy synthesis framework. Input parameters, both for the glottis model and the VT geometry, are derived from the acoustic features and the VT images. The acoustic propagation is then computed at each time step to simulate the utterance.

3.1 Area functions

The contours of the VT in the midsagittal plane were derived from X-ray films comprising several short French sentences [17] uttered by a 25 years old French native female speaker. Area functions were obtained by dividing the VT shape in tubelets perpendicular to the VT centerline [9], and then applying α β transformations to recover the area [19]. Area functions are then optimized so that the resonance frequencies, computed via an independent frequency-based technique [18], match the observed formant frequencies of the original utterance. The inversion technique is an iterative method [3] with biomechanic constraints. It is used in order to minimize the potential discrepancies between generated and observed formants due to the arbitrary choice of α β transformations. By setting the biomechanic constraints to high values, it guarantees minor modifications from the arbitrary α β choice.

The time sampling between two successive images is 20 ms, which is larger than the time duration of some phonetic events, such as the production of stop consonants. Thus, a second step consists in interpolating the estimated area functions, seen as temporal targets. The time location of these targets are set according to a preliminary manual phonetic segmentation. This follows the technique introduced in [13].

The oronasal coupling is defined via the 2D velum model introduced in [8]. This enables degree of nasality to be realistically taken into account during the production of nasal phonemes. In this model, contours of the velum are also derived from X-ray images.

3.2 Glottal parameters

Glottal source parameters are derived from the original speech signal and from nominal values found in the literature. The fundamental frequency contour, extracted from the original signal, gives information about the mechanical parameters of the lumped-element model. Variations of fundamental frequency are simulated by multiplying the value of the stiffness by a factor $Q_f^2 = 2\pi^2 F_0^2 \frac{m_i}{k_i}$, where $m_i = 0.1$ g, and $k_i = 80$ N/m are the nominal values. Since the acoustic coupling between the glottis and the VT may significantly modify the expected F_0 of the simulated utterance, an iterative method is used to modify the input F_0 so that the pitch contour of the simulated utterance matches the original one. The correction term is simply the difference between the obtained and the desired F_0. Basically, 2 iterations suffice.

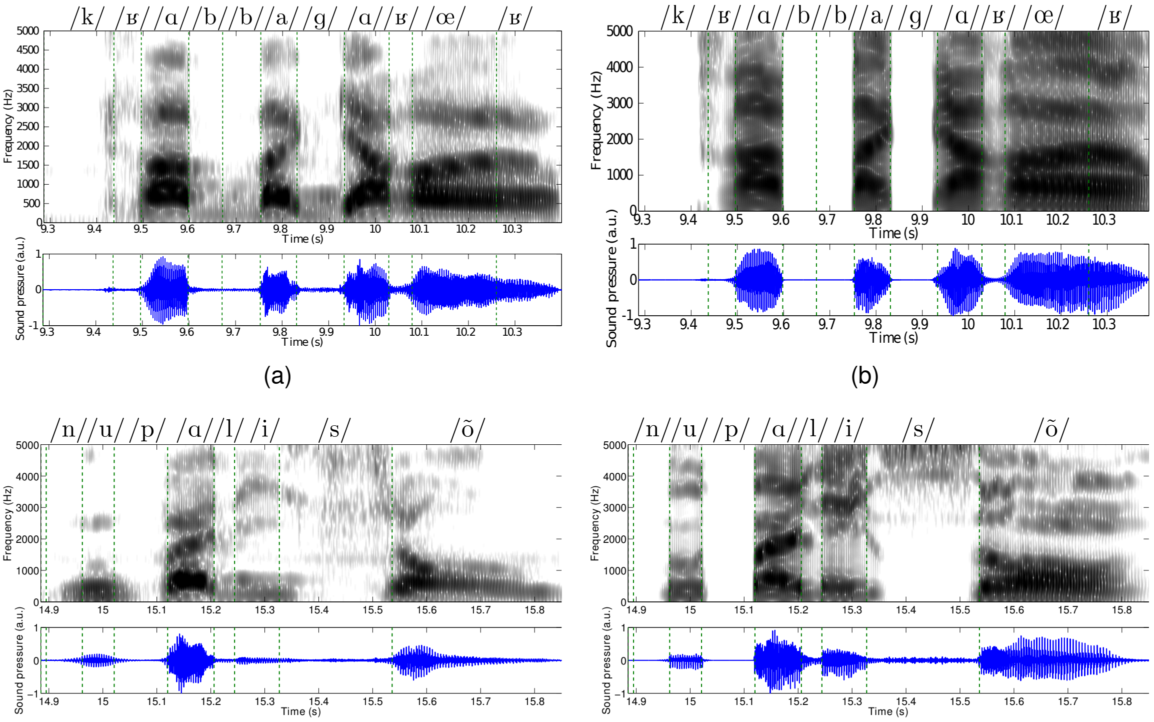
Finally, abduction is derived from the phonetic segmentation [9]: total abduction occurs for voiceless consonants, while a partial abduction occurs for voiced consonants. The partial abduction consists in a zip-like opening of the glottis [2,5], where only a portion of the vocal folds' length vibrates, while the other part is abducted. The abduction parameters is then defined by the quantity l_{ch}, corresponding to the opening length of the glottis. During the production of voiceless consonants, $l_{ch} = l_g$, namely the vocal folds are completely abducted, while it is less than l_g during the production of voiced fricatives. Note that for simulating breathy voice, l_{ch} may be set to a small but non-null value during the production of sonorants.

4 Copy synthesis

First, two copied sentences are compared with their original audio signal. Then, the evolution of a few aeroacoustic quantities as a function of time is computed to highlight the relevance of the method to investigate the impact of the vocal tract configuration on the oscillation of the vocal folds.

4.1 Simulated utterances

Utterances are chosen so that a large variety of French natural classes is represented. As a visual example, Fig. 2 shows the wide-band spectrograms and audio signals of two original utterances and the copied ones: "Crabes bagarreurs" (/kʁɑb.ba.gɑ.ʁɛʁ/) and "Nous palissons" (/nu.pɑ.li.sɔ̃/). In comparison with the original utterances, the copy-synthesized signals reproduce most of the acoustic features: formant trajectories are similar, as well as the temporal envelope, the phonetic segmentation, and the frication noise generation. This validates the interest of the method for quantitatively investigating the relationship between the articulatory and phonatory configurations and these acoustic features. However, spectrograms in Fig. 2 also highlights discrepancies. Apart from the presence of external noise and acoustic reverberation in the natural signal, the latter exhibits larger contrasts between formantic and non-formantic areas than in copy-synthesized signals. Besides, spectral tilts are different: simulated speech signals contain more energy in the high frequency area than natural signals. Several factors may explain these differences. First, TLCA-based methods are known to fail at accurately modeling acoustic losses in the vocal tract, especially because of its inability to account for frequency dependence. Secondly, although vocal tract parameters are optimized to match the observed fundamental frequency and phonetic segmentation, 2×2-mass models may fail at quantitatively reproducing the glottal source behavior [16].



(a)

(b)

(c)

(d)

Figure 2: Left column: Wide-band spectrograms and acoustic signals of the original utterances. Right column: Wide-band spectrograms and acoustic signals of the copied utterances. Top figures correspond to /kʁɑb.ba.gɑ.ʁœʁ/, and bottom figures to /nu.pɑ.li.sɔ̃/. The phonetic segmentation is represented by vertical dashed lines.

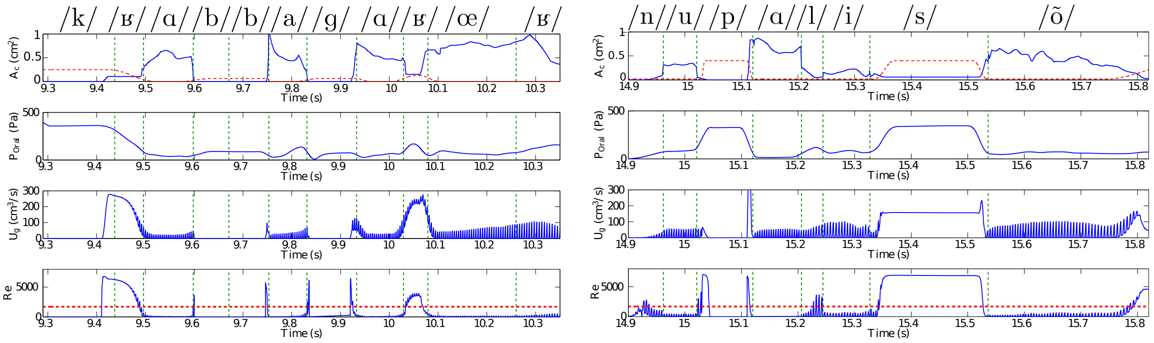
4.2 Aeroacoustic quantities

Fig. 3 shows the evolution of some aeroacoustic quantities as a function of time for the two simulated utterances: the intraoral pressure P_{Oral}, the glottal flow, and the Reynolds number. As previously observed [20], P_{Oral} raises when the supraglottal constriction area a_c is very small. This can be seen for the consonants /k/, /ʁ/, /b/, /g/, /p/, /l/, and /s/, and also for the close vowels /u/ and /i/. Note that the rise of P_{Oral} is more important for voiceless consonants, such as /k/, /p/, and /s/.

The release of occlusives systematically leads to a peak in the Reynolds number curve. Since the peak is above the threshold Re_c, the characteristic plosive noise may be generated.

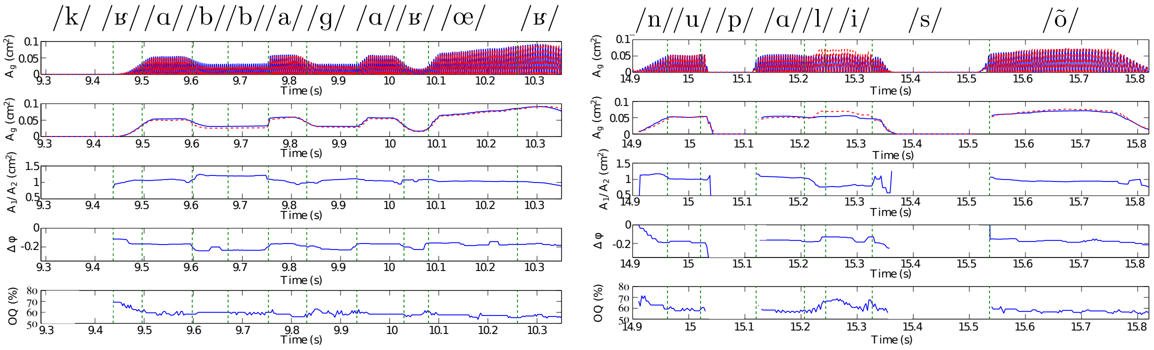
4.3 Oscillation of the vocal folds

The evolution of the oscillatory cycles of the vocal folds is shown in Fig. 4. It shows that, due to the opening of the glottal chink, the maximal opening area in the vibrating part of the vocal folds is smaller for consonants than for sonorants. The ratio between the amplitudes of mass 1 and mass 2, denoted A_1 and A_2 respectively, shows that the amplitude of A_1 is generally larger



(a) (b)

Figure 3: From top to bottom: supraglottal constriction area, intraoral pressure, glottal flow, and Reynolds number, as a function of time, computed during the simulation of (left) /kʁɑb.ba.gɑ.ʁœʁ/, and (right) /nu.pɑ.li.sɔ̃/. The frication noise threshold Re_c is denoted by the horizontal dotted line in the Reynolds number. The phonetic segmentation is represented by vertical dashed lines.



(a) (b)

Figure 4: From top to bottom: opening area of the oscillating vocal folds at the locations of mass 1 and 2, absolute amplitude of oscillations of mass 1 and 2, amplitude ratio of oscillation, phase differences, and open quotient, as a function of time, computed during the simulation of (left) /kʁɑb.ba.gɑ.ʁœʁ/, and (right) /nu.pɑ.li.sɔ̃/. Parameters are not computed during voiceless sounds /k,p,s/. The phonetic segmentation is represented by vertical dashed lines.

than A_2. However, in a few cases, $A_2 > A_1$: this occurs when the supraglottal constriction is relatively narrow, such as for the lateral /l/, the close vowels /u/, /i/, and the mid-close nasal vowel /ɔ̃/. This may be related to the rise of P_{Oral}. Indeed, the displacement of mass 2 is directly connected to the pressure forces distribution downstream of the glottal constriction, hence a larger amplitude of oscillation as P_{Oral} increases. Note that this also modifies the phase difference and the open quotient, which both decrease when P_{Oral} increases. The open quotient exhibits a hump from 60% to 70% during the sequence /li/. In other configurations, it remains constant, around a value slightly smaller than 60%.

5 Conclusions

The simulation framework presented in this paper allows the acoustical phenomena involved in speech production to be numerically simulated. From simultaneous acquisition of both audio speech signal and vocal tract imaging, it is possible to reproduce, or copy, the original speech signal by considering the observed vocal tract geometries and articulatory gestures in the framework input. The numerical examples provided in the paper show that salient acoustic features of natural speech, such as the formant trajectory, the phonetic segmentation, and the frication noise generation, may be quantitatively copy-synthesized from the articulatory and phonatory conditions of the speaker. However, due to the lack of information about the glottal source, differences are still observable between natural speech and its copied version, especially for the spectral tilt.

Despite this limitation, the presented simulation framework proves to be a useful tool for quantitatively study the relationships between the articulatory configurations and the formant trajectory, and/or the phonetic characteristics of the produced utterance, in running speech context. In the next future, more data of the aerodynamic configurations at the glottis should be acquired to define accurate time scenario for glottis/vocal tract coordination.

Acknowledgements

This study is supported by the ANR (*Agence Nationale de la Recherche*) ArtSpeech project.

References

[1] L. Bailly, X. Pelorson, N. Henrich, and N. Ruty. Influence of a constriction in the near field of the vocal folds: Physical modeling and experimental validation. *J. Acoust. Soc. Am.*, 124(5):3296–3308, 2008.

[2] B. Cranen and J. Schroeter. Physiologically motivated modelling of the voice source in articulatory analysis/synthesis. *Speech Communication*, 19(1):1–19, 1996.

[3] B. Elie and Y. Laprie. Audiovisual to area and length functions inversion of human vocal tract. In *Eusipco, Lisbon*, 2014.

[4] B. Elie and Y. Laprie. Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. *Speech Communication*, 82:85–96, 2016.

[5] B. Elie and Y. Laprie. A glottal chink model for the synthesis of voiced fricatives. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5240–5244, March 2016.

[6] K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Syst. Tech. J.*, 51(6):1233–1268, 1972.

[7] J. L. Kelly and C. C. Lochbaum. Speech synthesis. In *Proceedings of the Fourth International Congress on Acoustics*, pages 1–4, 1962.

[8] Y. Laprie, B. Elie, and A. Tsukanova. 2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes. In *Proceedings of the International Congress of Phonetic Science (ICPhS)*, 2015.

[9] Y. Laprie, M. Loosvelt, S. Maeda, E. Sock, and F. Hirsch. Articulatory copy synthesis from cine X-ray films. In *Interspeech 2013 (14th Annual Conference of the International Speech Communication Association)*, pages 1–5, Lyon, France, 2013.

[10] Y. Laprie, R. Sock, B. Vaxelaire, and B. Elie. Comment faire parler les images aux rayons X du conduit vocal (How to make X-ray images speak). In *SHS Web of Conferences*, pages 1285–1298. EDP Sciences, 2014.

[11] Y. Laprie, B. Vaxelaire, and M. Cadot. Geometric articulatory model adapted to the production of consonants. In *10th International Seminar on Speech Production (ISSP)*, pages 1–4, Köln, Allemagne, 2014.

[12] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1:199–229, 1982.

[13] S. Maeda. Phoneme as concatenable units: VCV synthesis using a vocal tract synthesizer. In *Sound Patterns of Connected Speech: Description, Models and Explanation, Proceedings of the symposium held at Kiel University, Arbeitsberichte des Institut für Phonetik und digitale Spachverarbeitung der Universitaet Kiel:31*, pages 145–164, 1996.

[14] P. Mokhtari, H. Takemoto, and T. Kitamura. Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches. *Speech Communication*, 50(3):179 – 190, 2008.

[15] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands, and Y. Auregan. Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model. *J. Acoust. Soc. Am.*, 96(6):3416–3431, 1994.

[16] N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, and A. Hirschberg. An in vitro setup to test the relevance and the accuracy of low-order vocal folds models. *J. Acoust. Soc. Am.*, 121(1):479–490, 2007.

[17] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, and J. Sturm. DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models. In *The Ninth International Seminar on Speech Production - ISSP'11*, pages 41–48, Canada, Montreal, 2011.

[18] M. M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. Acoust. Speech Sig. Process.*, 35(7):955–967, 1987.

[19] A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36(3):169–180, 2002.

[20] B. H. Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer Speech & Language*, 27(4):989–1010, 2013.

[21] M. Zañartu, G. E. Galindo, B. D. Erath, S. D. Peterson, G. R. Wodicka, and R. E. Hillman. Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction. *J. Acoust. Soc. Am.*, 136(6):3262–3271, 2014.