Copy-synthesis of phrase-level utterances

Benjamin Elie
LORIA
INRIA/CNRS/université de Lorraine
Email: benjamin.elie@inria.fr

Yves Laprie
LORIA
INRIA/CNRS/université de Lorraine
Email: yves.laprie@loria.fr

***Abstract*—This paper presents a simulation framework for synthesizing speech from anatomically realistic data of the vocal tract. The acoustic propagation paradigm is appropriately chosen so that it can deal with complex geometries and a time-varying length of the vocal tract. The glottal source model designed in this paper allows partial closure of the glottis by branching a posterior chink in parallel to a classic lumped mass-spring model of the vocal folds. Temporal scenarios for the dynamic shapes of the vocal tract and the glottal configurations may be derived from the simultaneous acquisition of X-ray images and audio recording. 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.**

Index Terms: Copy synthesis, Coordination, Glottal chink, Vocal folds

I. Introduction

The different methods for synthesizing human speech may be classified along an axis going from fully signal-based methods, such as concatenative synthesis [1], which requires almost no physical assumptions, to fully physics-based techniques, such as finite-element methods [2]. Intermediate methods integrate signal cues or physical assumptions according to the desired objective of the synthesis.

The aim of articulatory synthesis [3]–[5] is to simulate the physical and articulatory phenomena involved in speech production. The approach uses simplified physical models to approximate the realism of finite-element methods with lower computing complexity so that it allows phrase-level utterances to be simulated in a reasonable amount of time. The simulation frameworks are then a useful analysis-by-synthesis tool to study speech production. In this study, the aim is to reproduce the original formant trajectories, as well as the original prosody, and also to guarantee the phonetic contrasts.

Tackling the simulation of the physical and articulatory phenomena involved in speech production implies several challenges. First, one should model realistic vocal tract (VT) geometries by using only a few parameters and accurate time scenarios in order to reproduce the acoustic features of natural speech. Dealing with realistic geometries of the vocal tract is essential if the aim of the simulation framework is to investigate the relationships between articulatory configurations, or articulatory gestures of the speaker and the resulting acoustic features. Then, the glottal source must be finely modeled in order to simulate any kind of phonation, such as breathy or pathological voices, for instance. Finally, the acoustic coupling between the VT and the articulators should be taken into account to study the aeroacoustic conditions for the production of the self-oscillating movements of the vocal folds (VFs).

Recent models of articulatory synthesizers generate intelligible phrase-level utterances [4], [5], but generally at the expense of strong assumptions. For instance, the artificial talker in [5] uses simple VT geometries, does not consider any variation in the length of the VT, and imposes the glottal input area at the VT entrance. In [4], the VT geometry is modeled with simplified geometrical primitives, without considerations of side cavities.

The presented simulation framework aims at overcoming the aforementioned limitations. First, the chosen acoustic propagation model [6] enables complex VT geometries to be taken into account, including any side cavities and length variation. This acoustic model is then slightly modified to account for more realistic aeroacoustic conditions at the glottis in order to integrate self-oscillating models of the VFs. Finally, the glottis model is also modified to integrate a posterior glottal chink in order to simulate glottal partial closure [7]. The acoustic model is detailed in Sec. II. This paper also presents simulations of a few phrase-level utterances copied from the simultaneous recordings of natural speech and X-ray films of the VT (Sec. IV). The method for defining the temporal scenarios, based on both the recorded X-ray images of the VT and the original audio signal, is presented in Sec. III. The aim is to reproduce the original formant trajectories, as well as the original prosody, and also to guarantee the phonetic contrasts.

II. Acoustic model

A. Glottal source

1) Self-oscillations of the vocal folds: The airflow is considered as an unsteady inviscid and incompressible flow through the glottis, corrected with terms accounting for viscous losses [8]. A mobile separation point, denoted x_s is considered in the divergent downstream part of the VFs [9]. It is located at the position x along the flow direction such that the glottal height $h(x_s) = 1.2 \min[h(x)]$, where $h(x)$ denotes the height of the glottal constriction at position x. Following these assumptions, the pressure drop inside the glottis writes

$$\begin{aligned} P(x) &= P_{sub} + Be(x) + Po(x) + In(x) \qquad & x < x_s \\ P(x) &= P_{sup} & x > x_s, \end{aligned} \tag{1}$$

where $Be(x)$, $Po(x)$, and $In(x)$ are respectively the steady term of the Bernoulli equation, the Poiseuille corrective term and the unsteady term of the Bernoulli equation. They are defined as:

$$\begin{aligned} Be(x) &= -\frac{\rho U_g^2}{2l_g^2}\left[\frac{1}{h^2(x)} - \frac{1}{h^2(x_0)}\right], \\ Po(x) &= -\frac{12\mu U_g}{l_g}\int_{x_0}^{x}\frac{dx}{h^3(x)}, \\ In(x) &= -\frac{\rho}{l_g}\frac{\partial}{\partial t}\left[U_g\int_{x_0}^{x}\frac{dx}{h(x)}\right], \end{aligned} \tag{2}$$

where l_g is the length of the VFs, ρ and μ are respectively the mass density and the shear viscosity of the air. Note that the defined quantities are all time-dependent, and that for the sake of clarity, the term (t) is omitted in this paper.

When a self-oscillating model of the VFs is used, as in this paper, the geometry of the glottal constriction needs to be computed at each simulation time step. Distributed, or lumped, models are common tools to compute the glottis geometry driven by the aeroacoustic conditions. In such models, the VFs are modeled by mass-spring systems moving in the y direction, perpendicular to the airflow. The present study uses a 2 × 2-mass model with smooth contours [9]. The mass positions at each new simulation step are derived from the pressure forces, computed via Eq. (1), following the classic system of differential equations

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{R}\dot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{F}, \tag{3}$$

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.

2) Glottal chink: The glottis model also supports the integration of a posterior glottal chink, as in [7]. The glottis is then modeled as a "zip-like" structure, where the posterior partial abduction of the glottal folds leads to a glottal leakage in the triangular region of the chink, and the anterior part of the VFs vibrates according to their natural motion computed via the model detailed in the previous section. The partial closure of the glottis during the VFs oscillation enables voiced fricatives to be realistically simulated [10]. Besides, it may also be useful to simulate breathiness, which is an important acoustic cue, especially for gender identification [11].

B. Acoustic propagation in the vocal tract

The model for the acoustic propagation is based on the *transmission line circuit analog* (TLCA) method [3], and the more recently improved *single-matrix formulation* [6]. These methods are based on an electric-acoustic analogy, where the elementary acoustic tubelets that model the VT are seen as lumped circuit elements. In comparison with the other classic technique *reflection type line analog* (RTLA) method [12], TLCA-based methods [3], [6], [7] easily deal with time-varying shapes of the VT, and especially with length variations. Besides, the single-matrix formulation [6] enables the various side branches of the VT (nasal tract, piriform fossae…) to be simultaneously taken into account, as well as bilateral channels [7]. Consequently, TLCA-based methods are more suitable for time-domain continuous speech synthesis based on anatomically realistic VT geometries.

1) Connecting the glottis model: The integration of the glottis model detailed in Sec. II-A into the single-matrix formulation has been introduced and validated in [7], where the reader may find all computational details.

Basically, the introduction of Eq. (1) into the system of equations governing the values of the volume velocities U_i inside the N elementary acoustic tubelets leads to a quadratic equation at the glottis:

$$\mathbf{f} = \mathbf{Z}\mathbf{u}_Z + \mathbf{Q}\mathbf{u}_Q, \tag{4}$$

where $\mathbf{f} \in \mathbb{R}^{(N+1)}$ is a vector containing pressure terms, $\mathbf{Z} \in \mathbb{R}^{(N+1)\times(N+1)}$ is a tridiagonal matrix containing impedance and loss terms associated to each tubelet, $\mathbf{u} \in \mathbb{R}^{N+1} = [U_1, \dots, U_{N+1}]^T$ is the vector containing the volume velocities inside each tubelet, $\mathbf{Q}$ is a square matrix the same size as $\mathbf{Z}$ having only one non-zero element, that is $Q_{(1,1)} = Be/U_g^2$, and $\mathbf{u}_Q \in \mathbb{R}^{(N+1)} = [U_1^2, U_2^2, \dots, U_{N+1}^2]^T$ is the vector containing the square power of the volume velocities. Solving the system in Eq. (4) is quite straightforward after elimination of the sole quadratic equation by solving it after a preliminary rearrangement

$$\mathbf{Z}^{-1}\mathbf{f} = \mathbf{I}\mathbf{u}_Z + \mathbf{Z}^{-1}\mathbf{Q}\mathbf{u}_Q, \tag{5}$$

where $\mathbf{I}$ is the identity matrix.

2) Frication noise modeling: The model to generate frication noise is important to approach the naturalness of the synthesized speech. The method presented in this paper uses bandpass filtered Gaussian white noise sources [13]. The amplitude of the noise source P_{n_i} at section i is

$$P_{n_i} = \max\left\{0, \xi w\left(Re^2 - Re_c^2\right)\frac{U_{DC}^3}{a_{i-1}^{3/2}}\right\}, \tag{6}$$

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 air flow is turbulent [13], and consequently, above which the frication noise is generated, U_{DC} is the air flow volume velocity inside the VT, and a_{i-1} is the area of the upstream tubelet. For this study, Re_c is set to 1700. The computation of U_{DC} follows the method proposed by Maeda [14].

III. Input parameters

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.

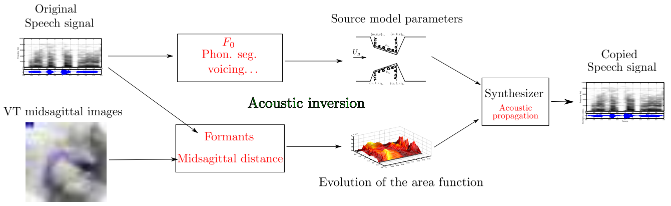


Figure 1. Simulation framework for copy synthesis.

Table I
INPUT PARAMETERS FOR THE VOCAL FOLDS MODEL

Parameter	Unit	Value	Parameter	Unit	Value
Subglottal pressure P_{sub}	Pa	800	Opening at point 0 h_0	mm	40
Position of mass 1 x_1	mm	0.2	Position of mass 2 x_2	mm	3.2
VF thickness d_g	mm	3	VF length l_g	mm	10
Nominal mass m_{1_i}	g	0.1	Nominal mass m_{2_i}	g	0.125
Nominal stiffness k_{1_i}	N/m	80	Nominal stiffness k_{2_i}	N/m	80
VF Abduction h_{ab}	mm	2.5			

A. Getting the area functions

The contours of the VT in the mid-sagittal plane were derived from X-ray films comprising several short French sentences [15] uttered by a 25 years old female speaker. French is her native language. Area functions were obtained by dividing the VT shape in tubelets perpendicular to the VT centerline, and then applying α β transformations to recover the area [16]. The determination of the centerline plays a critical role in synthesis since it influences the VT length, and consequently the resonance frequencies. A specific algorithm was designed purposely [17].

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 [14].

The oronasal coupling is defined via the 2D velum model introduced in [18]. 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.

Since the arbitrary α β parameters [16] may lead to discrepancies between simulated formant frequencies and that of the original sentence, a final step consists in adjusting the area functions. Thus, starting from the area functions derived from α β parameters, an iterative method [19] modifies them so that they generate an evolution of the resonance frequencies of the VT, computed by a independent frequency-based technique [13], that matches the formant trajectories of the original speech signal. Since the frequency-based technique to compute the resonance frequencies is independent, this can be used as a benchmark for assessing the ability of the presented method to reproduce observed formant trajectories.

B. Phonatory parameters

Since glottal source parameters cannot be derived from the X-ray images, they are derived from the original speech signal and from nominal values found in the literature.

The nominal values define the geometry of the glottis: they correspond to typical data for adult female subjects, and are displayed in Tab. I.

The fundamental frequency contour, extracted from the original signal, gives information about the mechanical parameters for the VFs model. Variations of fundamental frequency are simulated by multiplying the value of the stiffness by a factor $Q_f^2 = 2\pi^2 F_0^2 \frac{m_i}{k_i}$. 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 F_0 contour of the simulated utterance matches that of 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 [17]: 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 [10], [20], where only a portion of the VFs 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 VFs 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.

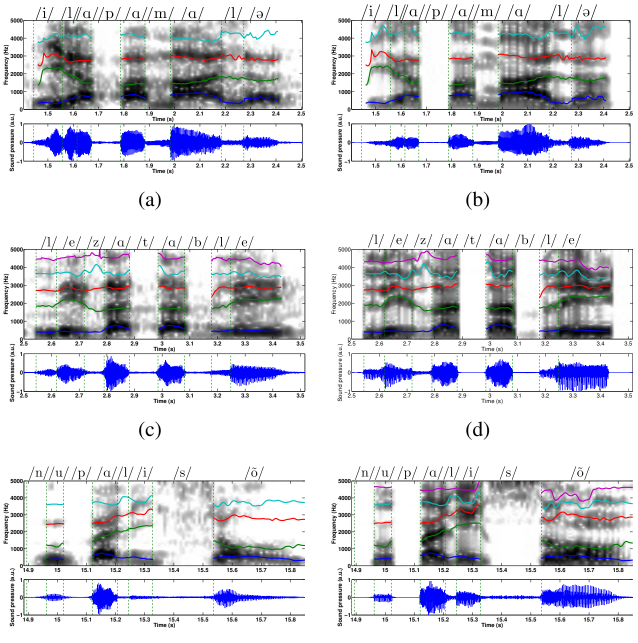
IV. NUMERICAL SIMULATIONS

A. Presentation of a few examples

A corpus of 11 eleven short phrase-level utterances have been copy-synthesized[1] and are summarized in Tab. II. They are chosen so that a large variety of French natural classes is represented. Unfortunately, the current database does not allow us, at this time, to extend the amount of copied utterances. However, we are currently developing cineMRI methods to acquire articulatory data at high spatiotemporal resolution. This is intended to increase the size of the corpus of simulated utterances in the next future. As a visual example, Fig. 2 shows the wide-band spectrograms and audio signals of three original utterances and the copied ones. They are chosen among the analyzed corpus, and correspond to "Il a pas mal" (/i.lɑ.pɑ.mɑ/lə/), "Les attablés" (/le.zɑ.tɑ.ble/), and "Nous palissons" (/nu.pɑ.li.sɔ̃/).

The acoustic features of the simulated signals agree with the original ones: both formant trajectories and temporal envelopes are similar. This demonstrates the good ability of the acoustic paradigm to reproduce the target acoustic clues. The effect of the coordination between the vocal tract configuration and the glottis is highlighted by the simulation of /s/ in Fig. 2 f). At the beginning of the phoneme, since the glottis is not sufficiently

[1] For the sake of briefty, only a few examples are presented. More audio files are available at: www.loria.fr/~belie/pages/artsynthesis.html



(a) (b)

(c) (d)

(e) (f)

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 /i.lɑ.pɑ.mɑ.lə/, middle figures to /le.zɑ.tɑ.ble/, and bottom figures to /nu.pɑ.li.sɔ̃/. In order to compare the formant trajectories, formants of the copied utterances are displayed in the spectrograms on the original ones, and formants of the original utterances are displayed on the spectrograms of the copied ones.

open, the voiced source is predominant, hence a small part (30 ms) of the /s/ is voiced. Then, when the glottal chink opening is large enough, the noise source becomes large compared to the voiced source, resulting in a devoiced sound. It is also the case in the original signal.

However, some differences are still observable. First, the original signals are a bit noisy, and echoes degrade the recording quality. It is clearly visible during the stop consonants /p/ and /t/ of the original signals. Another visible difference lies in the spectral tilt: the copied utterances exhibit flatter spectral tilts than original ones, i.e. they contain more energy in the high frequency domain. The main reason which would explain these discrepancies is that, since the glottal parameters of the speaker are unknown, some input parameters are arbitrarily chosen. Consequently, they are likely to be very different from the original speaker.

Although those approximations, that cannot be overcome so far, the synthesis quantitatively reproduces similar acoustic features of natural speech considering realistic vocal tract geometries. The acoustic propagation model deals correctly with self-oscillating vocal folds and the side branches that model the vocal tract. Note that the simulation of the nasalized sound /m/ is in good agreement with the original one.

B. *Quantitative evaluation*

The quality of copied utterances is evaluated via the computation of cepstral distances. It consists in computing the NRMSE values (*Normalized Root Mean Square Error*) between the MFCC (*Mel-Frequency Cepstral Coefficients*) of both the copied utterance and the original one. Tab. II shows the computed NRMSE values between a copied utterance and its corresponding original audio signal. It confirms that the synthesized utterances are correctly simulated. There is little variation in NRMSE values among the corpus (between 0.11 and 0.15).

Table II
CORPUS OF FRENCH UTTERANCES

Phrase	IPA	NRMSE	Phrase	IPA	NRMSE
Il a pas mal	/i.lɑ.pɑ.mɑ.lə/	0.131	Les attablés	/le.zɑ.tɑ.ble/	0.148
Très acariatres	/tʁɛ.zɑ.kɑ.ʁjat/	0.136	Il zappe pas mal	/il.zɑp.pɑ.mɑl/	0.124
Il l'a daté	/i.lɑ.dɑ.te/	0.109	Crabe bagarreur	/kʁɑb.bɑ.gɑ.ʁœʁ/	0.140
Trois sacs carrés	/tʁwɑ.sɑk.kɑ.ʁe/	0.129	Pas de date précise	/pɑd.dɑt.pʁe.siz/	0.117
Blague garantie	/blɑg.gɑ.ʁɑ̃.ti/	0.114	Nous palissons	/nu.pɑ.li.sɔ̃/	0.108
Elle a tout faux	/ɛ.lɑ.tu.fo/	0.125			

Fig. 3 shows the similarity matrix containing cross-NRMSE values, corresponding to the difference between a copied utterance and any original audio signal. As expected, it has smallest values in the main diagonal. The diagonal values are way below the off-diagonal ones: their mean value is 0.207 (the median value is 0.205) and the standard deviation is 0.026. Mean value of the diagonal is 0.126 (same as the median), with a standard deviation of 0.013.

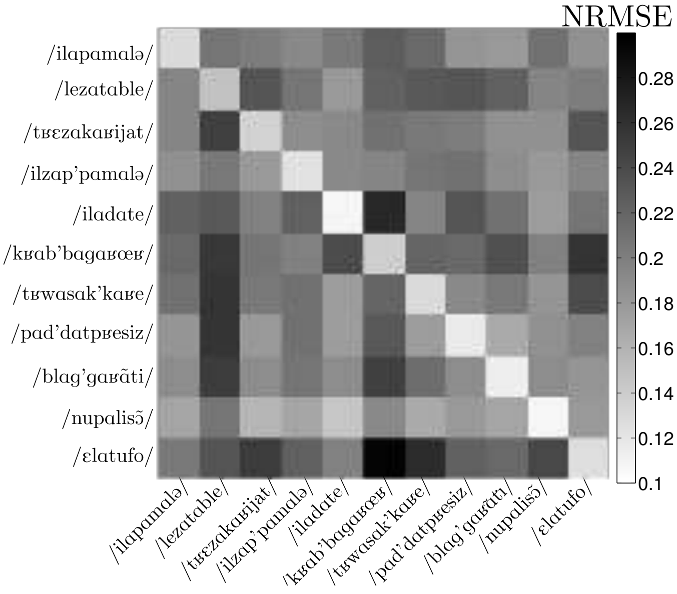


Figure 3. Matrix showing the cross-NRMSE values for the 11 French utterances of the corpus and the corresponding copied utterances.

V. CONCLUSIONS AND FURTHER WORKS

The simulation framework for speech synthesis presented in this paper has been shown to accurately reproduce acoustic features of natural French phrase-level utterances. In comparison with existing artificial talkers, it can support realistic geometries and dynamic deformations of the VT: the transmission line circuit analog model enables the dynamic variations of the VT geometry and its complexity to be accurately taken into account. Besides, it supports the connection with a glottis model that is able to reproduce self-oscillations of the VFs, as well as glottal leakage, which is a major contribution for the simulation of different types of phonation, such as breathy voice, and for the synthesis of voiced fricatives [10].

The ability to copy acoustic features of natural speech is shown from the copy of a corpus of 11 French utterances: temporal envelopes, phonetic contrasts, and formant trajectories are similar to those of the original signals. These similarities are quantitatively confirmed by the fact that the cepstral distance between original and copied utterances is significantly lower than cross-distance. This shows that the method can be exploited to evaluate physical parameters in the VT that are not easily accessible in practice, such as the internal pressures or the VFs motion. More specifically, it may be used to relate acoustic features of the produced speech signal to their articulatory or phonatory origins, thanks to analysis by synthesis techniques. This could be a great benefit for phonetic sciences, and/or language training.

References

[1] Chung-Hsien Wu, Yi-Chin Huang, Shih-Lun Lin, and Chia-Ping Chen, "Natural speech synthesis based on hybrid approach with candidate expansion and verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 250–254.

[2] F. Alipour, D. A. Berry, and I. R. Titze, "A finite-element model of vocal-fold vibration," *J. Acoust. Soc. Am.*, vol. 108(6), pp. 3003–3012, 2000.

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

[4] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in the time-domain simulation of the vocal tract system.," in *Proc. of the Interspeech 2004-ICSLP*, 2004, pp. 1125–1128.

[5] Brad H. Story, "Phrase-level speech simulation with an airway modulation model of speech production," *Computer Speech & Language*, vol. 27(4), pp. 989–1010, 2013.

[6] Parham Mokhtari, Hironori Takemoto, and Tatsuya Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches," *Speech Communication*, vol. 50(3), pp. 179 – 190, 2008.

[7] Benjamin Elie and Yves Laprie, "Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of vocal folds with glottal chink," Sept. 2015.

[8] 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.*, vol. 124(5), pp. 3296–3308, 2008.

[9] 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.*, vol. 96(6), pp. 3416–3431, 1994.

[10] B. Elie and Y. Laprie, "A glottal chink model for the synthesis of voiced fricatives," in *ICASSP*, 2016.

[11] Dennis H. Klatt and Laura C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87(2), pp. 820–857, 1990.

[12] John L Kelly and Carol C Lochbaum, "Speech synthesis," in *Proceedings of the Fourth International Congress on Acoustics*, 1962, pp. 1–4.

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

[14] 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*, 1996, pp. 145–164.

[15] 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*, Canada, Montreal, 2011, pp. 41–48.

[16] Alain Soquet, Véronique Lecuit, Thierry Metens, and Didier Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36(3), pp. 169–180, 2002.

[17] 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)*, Lyon, France, 2013, pp. 1–5.

[18] Yves Laprie, Benjamin Elie, and Anastasiia 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.

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

[20] Bert Cranen and Juergen Schroeter, "Physiologically motivated modelling of the voice source in articulatory analysis/synthesis," *Speech Communication*, vol. 19(1), pp. 1–19, 1996.