HAL
open science

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

Benjamin Elie, Yves Laprie

To cite this version:

Benjamin Elie, Yves 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, 2016, 82, pp.85-96. 10.1016/j.specom.2016.06.002 . hal-01199792v3

HAL Id: hal-01199792

https://hal.science/hal-01199792v3

Submitted on 28 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Benjamin Elie[1,*], Yves Laprie

LORIA, INRIA / CNRS / Université de Lorraine, Vandoeuvre-les-Nancy, France

Abstract

The paper presents extensions of the single-matrix formulation (Mokhtari *et al.*, 2008, Speech Comm. 50(3) 179 – 190) that enable self-oscillation models of vocal folds, including glottal chink, to be connected to the vocal tract. They also integrate the case of a local division of the main air path into two lateral channels, as it may occur during the production of lateral consonants. Provided extensions are detailed by a reformulation of the acoustic conditions at the glottis, and at the upstream and downstream connections of bilateral channels. The simulation framework is validated through numerical simulations. The introduction of an antiresonance in the transfer function due to the presence of asymmetric bilateral channels is confirmed by the simulations. The frequency of the antiresonance agrees with the theoretical predictions. Simulations of static vowels reveal that the behavior of the vocal folds is qualitatively similar whether they are connected to the

*Corresponding author. Tel: +33 383593036

Email address: `benjamin.elie@inria.fr` (Benjamin Elie)

single-matrix formulation or to the classic reflection-type line analog model. Finally, the acoustic effect of the glottal chink on the production of vowels is highlighted by the simulations: the shortening of the vibrating part of the vocal folds lowers the amplitude of the glottal flow, and therefore lowers the global acoustic level radiated at the lips. It also introduces an offset in the glottal flow waveform.
Keywords: Speech synthesis, Vocal folds, Glottal chink, Lateral consonants

1. Introduction

Time-domain continuous speech synthesizers are commonly based on simplified physical models to compute the acoustic propagation along the vocal tract [1–5] and/or the self-sustaining oscillations of the vocal folds [6–9]. In comparison with finite-element-based methods [10], which require a huge amount of time, their low computation time makes them interesting for continuous speech synthesis.

Simplified acoustic models use the plane wave assumption to compute the acoustic propagation along a set of acoustic tubes. The dimensions of the elementary tubes (or *tubelets*) approximate the geometry of the vocal tract. In regards to the typical dimensions of the human vocal tract, these models are valid up to frequencies around 5 kHz [11].

Articulatory synthesis bridges the gap between the articulatory and acoustic domains of speech. This is thus an invaluable tool to apprehend the acoustic impact of the speech articulator's gestures and their temporal coordination, that of the anatomic characteristics of the human vocal tract, and of the interactions between the vocal folds and the vocal tract. In

order to enable speech production to be studied via articulatory synthesis, several aspects should be covered by the numerical simulations of the speech aerodynamic/acoustic phenomena. First, the complexity of the vocal tract should be accurately modeled so that the various cavities (nasal tract, paranasal sinuses, sublingual cavities, etc) can be taken into account during the simulation. Then, the simulation framework should be able to deal with time-varying geometries of the vocal tract in order to simulate word-level or phrase-level utterances. This constrains the time trajectory of each articulator to be accurately modeled. Finally, the acoustic coupling between the glottal source, i.e. the vocal folds, and the vocal tract needs to be realistically modeled.

So far, there is no known time-domain continuous speech synthesizer that can deal with all these constraints. The scientific literature about speech synthesis based on simplified physical models identifies two main techniques: the *reflection-type line analog* method [1], which is called RTLA in this paper, and the *transmission line circuit analog* [2] method, called TLCA.

The reader may find a detailed review of existing techniques for speech synthesis in [12]. Basically, RTLA has the advantage of accurately accounting for the frequency dependence of acoustic losses, but suffers from the constraints on the tubelet dimensions in regards to the chosen simulation frequency. As a consequence, the total length of the vocal tract cannot be modified during the simulation. This is an important issue for continuous speech synthesis since the length of the vocal tract varies during natural speech production. Its use is usually limited to studies about the self-sustained motion of the vocal folds [13–15] coupled with simplified acoustic resonators. Using

RTLA to simulate running speech constrains the vocal tract to unrealistic simplified geometries [4]. On the other hand, many continuous speech synthesizers use TLCA [2, 6, 16, 17]. It is based on the electric-acoustic analogy: the vocal tract acoustics is seen as a lumped electric circuit. The main advantage of TLCA is its flexibility of use with time-varying geometries of the vocal tract, including length variation and uneven spatial sampling of the vocal tract. However, this analogy does not allow the frequency dependence of the acoustic losses and the acoustic radiation to be accurately taken into account. Recently, the *Single-Matrix Formulation* (SMF) [3] has been a major contribution to TLCA models: it is now possible to compute the acoustic propagation along a vocal tract modeled as a waveguide network, where each waveguide represents a side cavity. Consequently, using SMF is a useful tool to study the acoustic effects of the numerous side cavities in the context of continuous speech synthesis.

Another important challenge to tackle when dealing with articulatory synthesis is the glottal source model. Indeed, in order to thoroughly study the phonatory mechanisms, one should include a glottis model that is able to realistically account for the coupling between the vocal folds and the vocal tract. Many efforts to simulate the production of the glottal source have been made, and self-oscillating models of vocal folds, based on lumped mass-spring systems, have been of particular interest. The reader may find detailed reviews of these models in [18, 19]. Curiously, most of these studies have neglected the possibility of including a posterior glottal opening to simulate air leakage. This glottal opening, also called *glottal chink*, allows a DC component to appear in the glottal flow waveform, as commonly observed *in*

vivo [20]. A partial closure of the glottis during the oscillating cycle of the vocal folds may be useful to simulate voiced fricatives [21] and breathiness, which is an important acoustic cue, especially for gender identification [20].

First attempts at modeling a glottal chink have used parametric models [22, 23]. In these papers, two models of glottal leakage are proposed: firstly, the glottal chink is caused by a partial abduction of the vocal folds, i.e. only a portion of the vocal folds vibrates, the other part is abducted and forms a triangular glottal chink, and secondly, the glottal chink is formed in the inter-arytenoid portion of the glottis. In the second case, the vocal folds vibrate along their whole length. Later on, Wilhelms-Tricarico [24] proposed a modification of the classic two-mass model by Ishizaka and Flanagan [6] to include the glottal chink by connecting an electric branch in parallel to the vocal fold model. The glottal system is then connected to the first resonance of the vocal tract. Recently, Zañartu *et al.* [25] have studied the effect of the glottal chink on self-oscillating movements of the vocal folds. Although this study is a significant advance in glottis modeling, its connection with RTLA models of acoustic propagation does not make it suitable for continuous speech synthesis with realistic time-varying geometries of the vocal tract.

Starting from the single-matrix formulation presented in [3], this paper details the theory and the methodology for extending it by overcoming its limitations. The limitations of SMF are the following: it does not offer the possibility to connect a self-oscillating model of the vocal folds, and the configuration of anastomozing waveguides, i.e. the local division of the main oral tract into two lateral channels, as it may occur during the production

of lateral consonants, is not discussed. The aim is then to propose a complete simulation framework for speech synthesis that can account for the complexity of the vocal tract geometry and its numerous cavities taken simultaneously, that can deal with a time-varying realistic model of the vocal tract, including length variation, and that realistically models the acoustic coupling between the glottal source and the vocal tract, including a glottal leakage.

The paper is organized as follows. The transmission line circuit analog and the original single matrix formulation are detailed in Sec. 2. It also includes the required acoustic conditions at the glottis for integrating self-oscillating models of the vocal folds. The main aspects of the simulation framework, called *Extended Single-Matrix Formulation* (ESMF), are detailed in Sec. 3. They consist in the mathematical formulations for introducing the case of anastomozing waveguides into the single-matrix formulation, as well as the mathematical formulations to connect self-oscillating models of the vocal folds and a glottal chink to the single-matrix formulation. Finally, Sec. 4 presents numerical simulations that illustrate the accuracy of the extended single-matrix formulation to deal with the new features.

2. Theoretical background

This section summarizes the single-matrix formulation of the vocal tract by Mokhtari [3], which is itself derived from the *transmission line circuit analog* model by Maeda [2]. The present paper provides modifications in the formulation, taking into account the internal resistance of a noise source pressure to simulate frication noise. This reformulation is motivated by the

fact that many quantities introduced in this section are used to demonstrate the contributions detailed in the next section. Yet, for the sake of brevity, not all computation details have been provided, and one may refer to the original papers [2, 3] for more details.

2.1. Transmission line circuit analog model

The vocal tract is modeled as a concatenation of cylindrical tubes (or *tubelets*) along which plane waves propagate. The length and the cross-sectional areas of the tubelets are such that they approximate the vocal tract geometry. TLCA represents each tubelet as lumped circuit elements. Fig. 1 shows the lumped circuit elements of a single tube section and Tab. 1 details the acoustic-electric analogy.

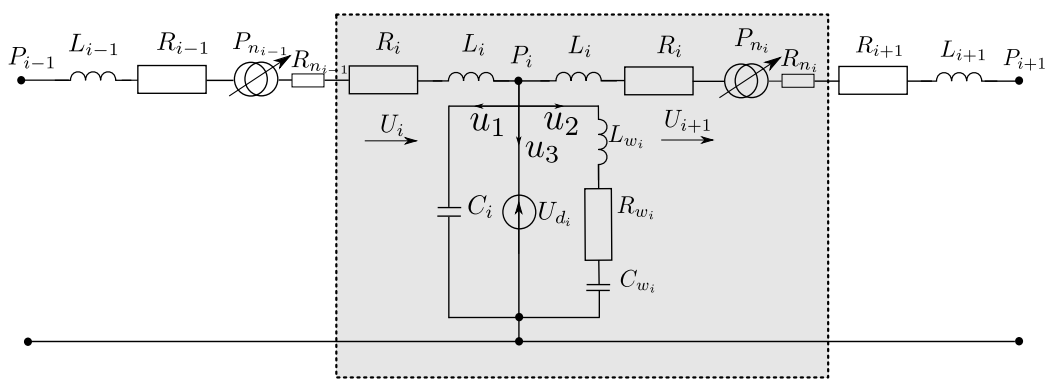


Figure 1: Lumped circuit elements of an acoustic tube. The acoustic-electric analogy is detailed in Tab. 1.

The terms W_R, W_C, and W_L are constant terms denoting respectively the resistance, the stiffness, and the mass of the vocal tract walls per area unit. Chosen values for this study are those provided in [16], namely $W_R = 8000$ kg.m^{-2}.s^{-1}, $W_C = 8.45 \times 10^6$ kg.m^{-2}.s^{-2}, and $W_L = 21$ kg.m^{-2}. By convention,

Table 1: Acoustic-electric analogy

Electric	Acoustic
Current	Volume velocity u
Voltage	Acoustic pressure p
R_i	Energy loss ($R_i = \frac{4\pi\mu l_i}{a_i}$)
C_i	Air compliance ($C_i = \frac{a_i l_i}{(\rho c_s)^2}$)
L_i	Air inertance ($L_i = \frac{\rho l_i}{2a_i}$)
R_{w_i}	Wall resistance ($R_{w_i} = \frac{W_R}{2l_i\sqrt{\pi a_i}}$)
C_{w_i}	Wall compliance ($C_{w_i} = \frac{2l_i\sqrt{\pi a_i}}{W_C}$)
L_{w_i}	Wall inertance ($L_{w_i} = \frac{W_L}{2l_i\sqrt{\pi a_i}}$)
U_{d_i}	Flow source ($-\frac{\partial}{\partial t} l_i a_i$)
P_{n_i}	Frication noise source ($P_{n_i} = \max\{0, \xi w \frac{U_{DC}^3}{a_{i-1}^{3/2}} (Re^2 - Re_c^2)\}$)
R_{n_i}	Internal resistance of noise source ($R_{n_i} = \kappa\rho \frac{U_{DC}}{a_{i-1}^2} + 8\pi\mu \frac{l_{i-1}}{a_{i-1}^2}$)

indices follow the air flow direction. For instance, considering the vocal tract, index 1 denotes the glottis connection, and index N denotes the lip termination.

Note that unlike in [2, 3], lumped circuit elements include a frication noise source. It is made up of a pressure source P_{n_i}, with an internal resistance R_{n_i} [5, 26, 27], that becomes active when the air flow is considered as turbulent, namely when the Reynolds number Re is above a certain threshold Re_c. Mathematical expressions for computing P_{n_i} and R_{n_i} are in Tab. 1, where ξ is an arbitrarily adjustable real constant used to control the noise level, and w is a Gaussian white noise to which first-order lowpass and third-

order highpass filters have been applied [26]. The chosen Re_c threshold for the study is 2700, and $\xi = 10^{-6}$, as suggested by Sondhi and Schroeter [27]. The frication noise source is usually located at the next point downstream of the supraglottal constriction [5, 26, 27], hence the spatial lag $i-1$. The term κ denotes a scaling coefficient set to 1.42, as used in [6, 26], and U_{DC} is the low-frequency component of the air flow [26]. Considering the lumped circuit elements displayed in Fig. 1, the continuous time equations relating the pressure and volume velocities inside all of the tubelets are

$$\begin{aligned} P_{i-1} - P_i &= \frac{\partial}{\partial t}\left[\left(L_{i-1} + L_i\right) U_i\right] + \left(R_{i-1} + R_i + R_{n_{i-1}}\right) U_i + P_{n_{i-1}} \\ U_i - U_{i+1} &= u_1 + u_2 + u_3, \end{aligned} \tag{1}$$

The paper follows the discrete representation of Eq. (1) introduced by Maeda [2], which yields to the following set of linear equations at any simulation time step:

$$\begin{cases} F_1 & = & Z_1 U_1 + b_1 U_2 \\ F_i & = & b_{i-1} U_{i-1} + Z_i U_i + b_i U_{i+1} \quad \text{for } 2 \leq i \leq N \\ F_{N+1} & = & b_N U_N + Z_{N+1} U_{N+1} \end{cases}, \tag{2}$$

where F_i, b_i, and Z_i are pressure forces, loss terms, and impedance terms, associated with tubelet i, and N is the number of tubelets modeling the considered tract. One may refer to [2] for detailed steps leading to Eq. (2).

The acoustic propagation is computed by estimating the values of the volume velocity U_i inside tubelets $i = 1, \ldots, N+1$ at each simulation time step. The previous system of equations forms a well-determined system, the

$N+1$ unknown volume velocities are governed by a set of $N+1$ linear equations. It can be rewritten into the following matrix form

$$\mathbf{f} = \mathbf{Z}\mathbf{u}, \tag{3}$$

where

$$\begin{aligned}
\mathbf{f} \in \mathbb{R}^{(N+1)} &= [F_1, \ldots, F_{N+1}]^T \\
\mathbf{u} \in \mathbb{R}^{N+1} &= [U_1, U_2, \ldots, U_{N+1}]^T, \\
\mathbf{Z} \in \mathbb{R}^{(N+1)\times(N+1)} &= \begin{bmatrix} Z_1 & b_1 & 0 & \\ b_1 & Z_2 & b_2 & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & b_N & Z_{N+1} \end{bmatrix}.
\end{aligned}$$

2.2. Single-matrix formulation of the vocal tract

The SMF, as defined by Mokhtari *et al.* [3], is a reformulation of the TLCA model to be used with a vocal tract seen as a waveguide network. The equations driving the acoustic propagation along every waveguide are merged into a single matrix. Each waveguide of the network represents a side cavity, represented as a parallel side branch in the analog lumped circuit.

In the waveguide network, the oral subcavity, going from the glottis to the lips, is called the *root node*. Quantities derived from the root node are denoted by the symbol (1) as exponent. Each waveguide connected to the oral tract is one of its *children*, and children of the root nodes may also have children themselves. The whole vocal tract can then be seen as a tree structure, where the root node is the oral subcavity. According to Mokhtari *et al.* [3], the system of equations driving the acoustic propagation inside $\mathcal{N}$ waveguides of the network is

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \vdots \\ \mathbf{f}^{(\mathcal{N})} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(2)T} & \dots & \mathbf{C}_1^{(\mathcal{N})T} \\ \mathbf{C}_1^{(2)} & \mathbf{Z}^{(2)} & & \mathbf{C}_2^{(\mathcal{N})T} \\ \vdots & & \ddots & \vdots \\ \mathbf{C}_1^{(\mathcal{N})} & \mathbf{C}_2^{(\mathcal{N})} & & \mathbf{Z}^{(\mathcal{N})} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \vdots \\ \mathbf{u}^{(\mathcal{N})} \end{bmatrix}, \tag{4}$$

where $\mathbf{f}^{(m)}$, $\mathbf{Z}^{(m)}$, and $\mathbf{u}^{(m)}$ contain the elements of Eq. (3) driving the acoustic propagation along the waveguide (m), independently of the rest of the network. Note that the number of waveguides $\mathcal{N}$ included in the network should not be confused with N, which denotes the number of tubelets included in a single waveguide. The number N may differ according to the waveguide it is related to.

The single-matrix formulation is then a concatenation of the linear systems driving the acoustic propagation inside each individual network, where sparse coupling matrices $\mathbf{C}_m^{(n)}$ are added to account for the junction between two waveguides (m) and (n). In [3], the formulation provided by the authors distinguishes three cases: i) the waveguides are not directly connected, hence $\mathbf{C}_m^{(n)} = \mathbf{0}$, ii) (n) is a child of (m), connected at the Kth tubelet of (m). Then $\mathbf{C}_m^{(n)}$ is a sparse matrix whose non-zero elements are $c_{1,K} = b_K^{(m)}$ and $c_{1,K+1} = Z_C^{(m,n)}$, and iii) (m) and (n) are twins, both connected to the waveguide (p). This is the case for the piriform fossae, for instance: they are usually both connected to the sampled oral tract at the same tubelet. Then $\mathbf{C}_m^{(n)}$ is a sparse matrix whose sole non-zero element is $c_{1,1} = Z_C^{(p,n)}$.

Details of the mathematical expressions for computing $\mathbf{C}_m^{(n)}$ may be found in [3]. The term $Z_C^{(m,n)}$ accounts for the coupling between the pair of waveg-

uides $\{m, n\}$ at the junction K of the waveguide (m), and

$$Z_C^{(m,n)} = -\left[b_K^{(m)} + R_K^{(m)} + R_{n_K}^{(m)} + \frac{2}{T} L_K^{(m)}\right].$$

2.3. Two-mass model and aeroacoustic considerations at the glottis

The method presented in this paper is intended to support any kind of self-oscillating model of the vocal folds. For the sake of brevity, only one will be used for the simulations as an example. We arbitrarily chose the 2×2-mass model with smooth contours [7, 14], but other models could be considered, such as the one-mass model [28], the three-mass model [29], or even models dealing with more masses [30]. It is based on two spring-mass systems, representing the rear and front ends of the vocal folds. It stems from the basic two-mass model by Ishizaka and Flanagan [6]. The model presented in this section considers recent improvements: contours are smooth, allowing a mobile separation point [7, 14], and it adds corrective terms to take into account the viscous losses and the unsteady flow effects [15, 31]. The geometry of the glottal constriction used in this paper is illustrated in Fig. 2.

The Bernoulli equation for unsteady flow, with an additive Poiseuille corrective term, gives the pressure $P(x,t)$ along the glottal constriction:

$$\begin{aligned} P(x,t) &= P_{sub} + Be(x,t) + Po(x,t) + In(x,t) \qquad && x < x_s \\ P(x,t) &= P_{sup} && x > x_s, \end{aligned} \tag{5}$$

where x_s is the flow separation point, and $Be(x,t)$, $Po(x,t)$, and $In(x,t)$ are respectively the steady term of the Bernoulli equation, the Poiseuille

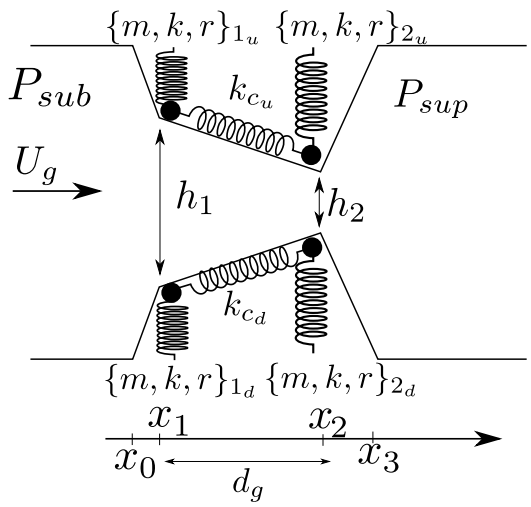


Figure 2: Geometry of the glottal constriction, as introduced by Lous *et al.* [14]. Indices u and d stand for *up* and *down* respectively.

corrective term and the unsteady term of the Bernoulli equation. They are defined as:

$$
\begin{aligned}
Be(x,t) &= -\frac{\rho U_g^2(t)}{2l_g^2}\left[\frac{1}{h^2(x,t)} - \frac{1}{h^2(x_0,t)}\right], \\
Po(x,t) &= -\frac{12\mu U_g(t)}{l_g}\int_{x_0}^{x}\frac{dx}{h^3(x,t)}, \\
In(x,t) &= -\frac{\rho}{l_g}\frac{\partial}{\partial t}\left[U_g(t)\int_{x_0}^{x}\frac{dx}{h(x,t)}\right],
\end{aligned}
\tag{6}
$$

where $h(x)$ is the glottal opening along the x-coordinate, l_g is the length of the vocal folds, ρ and μ are respectively the mass density and the shear viscosity of the air. The position of the flow separation point x_s varies according to the glottal constriction geometry: x_s is such that $h_s = h(x_s) = 1.2h_1$ if $1.2h_1 < h_2$, and $x_s = x_2$ otherwise.

The value 1.2 is an *ad-hoc* criterion, as used in [7, 14]. At the flow

separation point $x = x_s$, the pressure drop between the upstream and the downstream parts of the glottis is given by Eq. (7)

$$P_{sub}(t) - P_{sup}(t) = R_b(t)U_g^2(t) + R_v(t)U_g(t) + \frac{\partial}{\partial t}\left[L_g(t)U_g(t)\right], \tag{7}$$

where

$$\begin{aligned} R_b(t) &= \frac{\rho}{2l_g^2}\left[\frac{1}{h^2(x_s,t)} - \frac{1}{h^2(x_0,t)}\right], \\ R_v(t) &= \frac{12\mu}{l_g}\int_{x_0}^{x_s}\frac{dx}{h^3(x,t)}, \\ L_g(t) &= \frac{\rho}{l_g}\left[\int_{x_0}^{x_s}\frac{dx}{h(x,t)}\right], \end{aligned} \tag{8}$$

From the determination of the glottal flow U_g, Eq. (5) gives the pressure distribution $P(x)$ along the glottal constriction. The pressure forces are then used to derive the mass positions at each simulation step following the classic system of differential equations

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{R}\dot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{F}, \tag{9}$$

where $\mathbf{M} \in \mathbb{R}_+^{4\times4} = \mathrm{diag}\,(m_{1u}, m_{2u}, m_{1d}, m_{2d})$, $\mathbf{R} \in \mathbb{R}_+^{4\times4} = \mathrm{diag}\,(r_{1u}, r_{2u}, r_{1d}, r_{2d})$, and $\mathbf{F} \in \mathbb{R}^{4\times4} = \mathrm{diag}\,(F_{1u}, F_{2u}, F_{1d}, F_{2d})$ are diagonal matrices containing the values of respectively the mass, the damping and the pressure forces applied to each mass, $\mathbf{y} \in \mathbb{R}^4 = [y_{1u}, y_{2u}, y_{1d}, y_{2d}]^T$ is the vector containing the displacement of each mass from its rest position, and $\mathbf{K} \in \mathbb{R}^{4\times4}$ is a matrix containing stiffness coefficients. Due to the presence of a coupling spring k_c,

$$\mathbf{K} = \begin{bmatrix} k_{1u} + k_{cu} & -k_{cu} & 0 & 0 \\ -k_{cu} & k_{2u} + k_{cu} & 0 & 0 \\ 0 & 0 & k_{1d} + k_{cd} & -k_{cd} \\ 0 & 0 & -k_{cd} & k_{2d} + k_{cd} \end{bmatrix}$$

Pressure forces $F_{i,j}(t)$, with $j = \{u, d\}$, are derived from the pressure applied to the mass at instant t:

$$\begin{aligned} F_{1,j} &= l_g \int_{x_0}^{x_1} \frac{x - x_0}{x_1 - x_0} P(x)dx + l_g \int_{x_1}^{x_2} \frac{x - x_2}{x_1 - x_2} P(x)dx \\ F_{2,j} &= l_g \int_{x_1}^{x_2} \frac{x - x_1}{x_2 - x_1} P(x)dx + l_g \int_{x_2}^{x_3} \frac{x - x_3}{x_2 - x_3} P(x)dx \end{aligned}$$

3. Extended Single-Matrix Formulation of the Vocal Tract

General remark:

In this paper, the subglottal pressure is imposed as an input parameter. Although its temporal evolution should be considered for time-domain continuous speech synthesis, especially in order to simulate natural prosody, it has not been taken into account in this paper since the compatibility with the single-matrix formulation has already been proven, even for complex geometries, by Ho *et al.* [32].

3.1. Anastomozing waveguides: bilateral consonants

In some cases, the air path inside the vocal tract may be locally divided into two lateral channels. For instance, this is observed in the case of lateral consonants [33–35]. The acoustic effect is not fully apprehended, mainly because of the lack of appropriate acoustic models, and also because of insufficient relevant articulatory data. Indeed, lateral consonants are usually

modeled by a single lateral channel, and additionally a supralingual cavity [36, 37]. Zhang *et al.* [34, 35] studied the effect of bilateralization with a frequency-based method to compute the transfer function of bilateral vocal tracts. However, to the best of our knowledge, there is still no existing model of bilateral vocal tracts to be used in the context of time-domain continuous speech synthesis. This section details the theoretical and mathematical developments for integrating this condition into the single-matrix formulation. The aim is then to define the mathematical expression of the coupling matrix $C_m^{(n)}$ of Eq. (4) when (m) and (n) are two parallel waveguides, connected to each other at two different points.

Fig. 3 shows the waveguide connections in the case of bilateralization. It includes a secondary waveguide that is connected to the oral branch at two points.

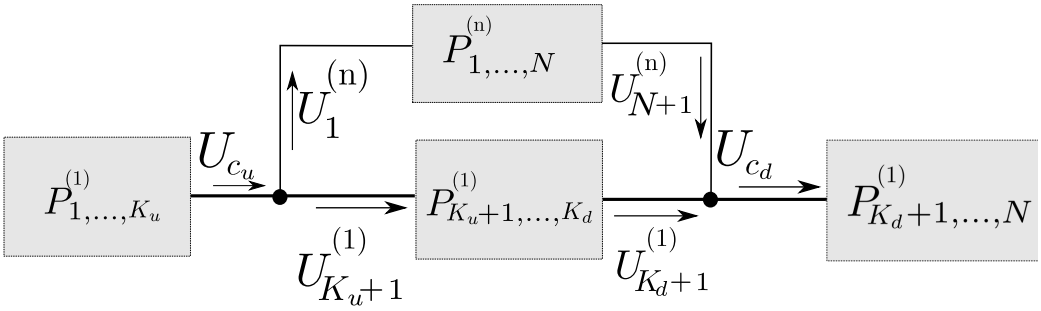


Figure 3: Equivalent diagram of anastomozing waveguides.

Following the medical and the hydrography terminology, this relation is called *anastomosis* in this paper. This relationship distinguishes a main anastomozing waveguide and a secondary one, called *anabranch*. For instance, the main anastomozing waveguide in Fig. 3 is waveguide (1), while (n) is its anabranch.

It is clear that the conditions at the upstream connection are similar to a parent-child relationship, hence similar expressions in the coupling matrix $C_m^{(n)}$. In that case, three equations of the system, corresponding to the three tubelets around the junction, are modified:

$$\begin{aligned} F_{K_u}^{(1)} &= b_{K_u-1}^{(1)} U_{K_u-1}^{(1)} + Z_{K_u}^{(1)} U_{K_u}^{(1)} + b_{K_u}^{(1)} U_{K_u+1}^{(1)} + b_{K_u}^{(1)} U_1^{(n)}, \\ F_{K_u+1}^{(1)} &= b_{K_u}^{(1)} U_{K_u}^{(1)} + Z_{K_u+1}^{(1)} U_{K_u+1}^{(1)} + b_{K_u+1}^{(1)} U_{K_u+2}^{(1)} + Z_C^{(1,n)} U_1^{(n)}, \\ F_1^{(n)} &= Z_1^{(n)} U_1^{(n)} + b_1^{(n)} U_2^{(n)} + b_{K_u}^{(1)} U_{K_u}^{(1)} + Z_C^{(1,n)} U_{K_u+1}^{(1)}. \end{aligned} \tag{10}$$

Similarly, at the downstream connection K_d, the equations corresponding to the tubelets around the junction are modified:

$$\begin{aligned} F_{K_d+1}^{(1)} &= b_{K_d}^{(1)} U_{K_d}^{(1)} + Z_{K_d+1}^{(1)} U_{K_d+1}^{(1)} + b_{K_d+1}^{(1)} U_{K_d+2}^{(1)} + Z_{C+1}^{(1,n)} U_{N+1}^{(n)}, \\ F_{K_d+2}^{(1)} &= b_{K_d+1}^{(1)} U_{K_d+1}^{(1)} + Z_{K_d+2}^{(1)} U_{K_d+2}^{(1)} + b_{K_d+2}^{(1)} U_{K_d+3}^{(1)} + b_{K_d+1}^{(1)} U_{N+1}^{(n)}, \\ F_{N+1}^{(n)} &= b_N^{(n)} U_N(n) + Z_{N+1}^{(n)} U_{N+1}^{(n)} + b_{K_d+1}^{(1)} U_{K_d+2}^{(1)} + Z_{C+1}^{(1,n)} U_{K_d+1}^{(1)}, \end{aligned} \tag{11}$$

where

$$\begin{aligned} Z_{C+1}^{(1,n)} &= -\left[b_{K+1}^{(1)} + R_{K+1}^{(1)} + \frac{2}{T} L_{K+1}^{(1)} \right], \\ Z_{N+1}^{(n)} &= -b_N^{(n)} - b_{K+1}^{(1)} - \frac{2}{T}\left(L_N^{(n)} + L_{K+1}^{(1)} \right) - R_N^{(n)} - R_{K+1}^{(1)} - R_{n_N}^{(n)}. \end{aligned} \tag{12}$$

Note that in that case, the expression of $Z_{N+1}^{(n)}$ differs from other configurations because of different boundary conditions at the waveguide termination.

The matrix formulation is then

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(n)T} \\ \mathbf{C}_1^{(n)} & \mathbf{Z}^{(n)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(n)} \end{bmatrix}, \tag{13}$$

where $\mathbf{C}_1^{(n)}$ is the matrix accounting for the coupling between the anastomozing waveguides. It is a sparse matrix with 4 non-zero elements $c_{1,K_u} = b_{K_u}^{(1)}$, $c_{1,K_u+1} = Z_C^{(1,n)}$, $c_{N+1,K_d+2} = b_{K_d+1}^{(1)}$, and $c_{N+1,K_d+1} = Z_{C+1}^{(1,n)}$, where N is the generic number of tubelets that model the anabranch (n), and K_u and K_d are the upstream and downstream locations on the main oral tract where (n) is connected.

3.2. Integration of the self-oscillating model of vocal folds

In the original papers [2, 3], the glottal source is generated by an imposed oscillating glottal input area. This may be sufficient to simulate utterances with a good quality, but it cannot be used to study the acoustic coupling between the vocal folds and the vocal tract. In [32], the authors connect the original two-mass model by Ishazaka and Flanagan [6] to the single-matrix formulation, in order to simulate the self-oscillating motion of the vocal folds. However, the quadratic term in Eq. (7) is not taken into account. The present section details the mathematical considerations to account for a more realistic aeroacoustic model at the glottis when the single-matrix formulation is used in the context of time-domain speech synthesis.

To connect the self-oscillating model of the vocal folds, the pressure distribution along the glottal constriction should be known at each simulation step. Thus, Eq. (7) should be integrated into the single-matrix formulation. Once U_g is known, pressure forces are derived from Eq. (5), and the mass positions are computed following Eq. (9).

Introducing Eq. (7) into the single-matrix formulation of Eq. (22) requires the first line of the system to be modified into a quadratic equation

$$F_1 = Z_1 U_1 + b_1 U_2 + R_b U_1^2. \tag{14}$$

The matrix form is then

$$\mathbf{f} = \mathbf{Z}\mathbf{u}_Z + \mathbf{Q}\mathbf{u}_Q, \tag{15}$$

where $\mathbf{Q}$ is a square matrix the same size as $\mathbf{Z}$ having only one non-zero element, i.e. $Q_{(1,1)} = R_b$, and $\mathbf{u}_Q \in \mathbb{R}^{(N+1)} = [U_1^2, U_2^2, \ldots, U_N^2]^T$ is the vector containing the square power of the volume velocities. Eq. (15) is also valid in the case of a waveguide network, since it does not directly modify the coupling equations between the different side cavities modeling the VT.

The system is almost entirely linear: only the first line is a quadratic equation. To solve the system, one should first solve the quadratic equation separately. It could be straightforward if the first line of $\mathbf{Z}$ contained only one non-zero element. Unfortunately, it is not the case since $Z_{(1,1)} \neq 0$ and $Z_{(1,2)} \neq 0$. A practical solution consists in finding an equivalent system in which the matrix of linear coefficients is a diagonal matrix. Left-multiplying both sides of Eq. (15) by $\mathbf{Z}^{-1}$ yields

$$\mathbf{Z}^{-1}\mathbf{f} = \mathbf{I}\mathbf{u}_Z + \mathbf{Z}^{-1}\mathbf{Q}\mathbf{u}_Q, \tag{16}$$

where $\mathbf{I}$ is the identity matrix. In this formulation, the first line of the system depends only on U_1. The quadratic equation can then be solved, and the value U_1 that is accepted is the largest positive solution. If both roots are negative, U_1 is set to 0. Once the glottal volume velocity U_1 is known, Eq. (15) can be rearranged

$$\tilde{\mathbf{f}} = \tilde{\mathbf{Z}}\tilde{\mathbf{u}}_Z, \tag{17}$$

where $\tilde{\mathbf{f}} \in \mathbb{R}^N = [f_2 - b_1 U_1, f_3, \ldots, f_n, \ldots, f_{N+1}]^T$, $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times N}$ is the matrix $\mathbf{Z}$ of which the first line and first column have been withdrawn, and $\tilde{\mathbf{u}}_Z \in \mathbb{R}^N = [U_2, U_3, \ldots, U_n, \ldots, U_{N+1}]^T$.

Eq. (17) is then a well-determined tridiagonal linear system. Any classical method to solve such systems gives the solutions U_i with $i = 2, \ldots, N+1$.

3.3. A glottal chink model

The self-sustaining model of the vocal folds presented in the previous section considers them to vibrate uniformly along their length. Consequently, at this stage, it is not possible to account for a partial closure of the glottis. This may be an issue for the synthesis of several types of phonation, including breathy voice and voiced fricatives. In these cases, the glottis is never completely closed, and an offset appears in the glottal flow waveform, so that it is never null.

The model presented in this paper connects an electric branch in parallel to the vocal fold model. The electric analogy and the parallel branch make this model interesting for the extended single-matrix formulation. Hence the equivalent electric circuit represented in Fig. 4. Unlike in [24], U_g and U_{ch} separate upstream of the glottal contraction, and merge downstream of the glottal expansion. These assumptions follow the model detailed in [23]. The formulation assumes that the partial closure of the glottis is due to a partial abduction of the vocal folds and that the chink is linked to the membranous part of the glottis, as shown in Fig. 4. Since the model considers

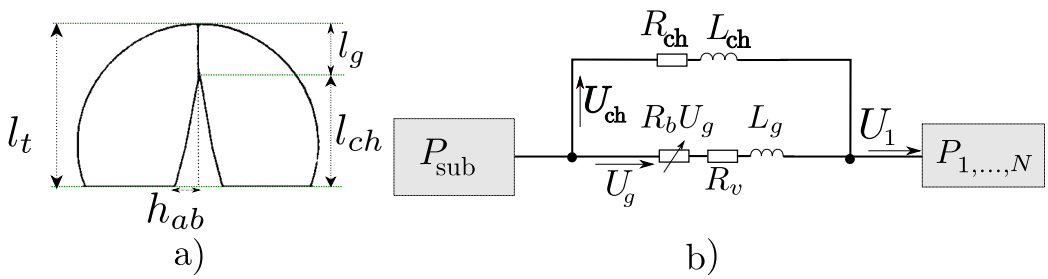


Figure 4: a) View of the partially closed glottis, extracted from Cranen and Schroeter [22]. In this model, the partial closure is due to a partial abduction of the vocal folds. l_g is the length of the vibrating part of the vocal folds, l_{ch} is the length of the glottal chink and $l_t = l_g + l_{ch}$ is the total length of the vocal folds. The abduction of the vocal folds is assumed to be constant and is denoted by h_{ab}. b) Electric-circuit analogy of the partially closed glottis. U_{ch}, R_{ch}, and L_{ch} are the volume velocity through the glottal chink, the energy loss, and the air inertance inside the glottal chink, respectively.

one-dimensionnal inviscid flow, the chink and the oscillating part of the vocal folds may be represented as independent channels.

The glottal chink area is $a_{ch} = l_{ch} h_{ab}$, where l_{ch} is the length of the glottal chink and h_{ab} is the abduction of the vocal folds. Note that in the original model [23], the glottal chink may be extended with a constant opening area, due to the inter-arytenoid portion of the glottis. The presented model does not account for it, but the implementation is straightforward: the glottal chink area is then $a_{ch} = l_{ch} h_{ab} + a_{ia}$, where a_{ia} is the opening area of the inter-arytenoid portion of the glottis. When the glottal chink is considered, U_1 is no longer connected to U_g. Indeed, as seen in Fig. 4 b), U_1 is the merged flow downstream the glottal expansion, and is therefore the sum of the volume velocities through the chink and through the vibrating part of the vocal folds ($U_1 = U_g + U_{ch}$).

The presence of the glottal chink modifies the first line of the system defined in Eq. (16). Indeed, applying Eq. (7) to both glottal branches and reorganizing equations yields to the following boundary conditions at the glottis

$$\begin{aligned} F_1 &= R_b U_g^2 + Z_1 U_g + b_1 U_2 + Z_C^{(1,ch)} U_{ch}, \\ F_2 &= b_1 U_g + Z_2 U_2 + b_2 U_3 + b_1 U_{ch}, \\ F_{ch} &= Z_{ch} U_{ch} + Z_C^{(1,ch)} U_g + b_1 U_2, \end{aligned} \tag{18}$$

where

$$Z_C^{(1,ch)} = b_1 + R_1 + \frac{2}{T} L_1 \tag{19}$$

$$Z_{ch} = b_1 + R_{ch} + R_1 + \frac{2}{T}(L_{ch} + L_1). \tag{20}$$

Introducing Eqs. (18) in the single matrix formulation yields

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ F_{ch} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(ch)T} \\ \mathbf{C}_1^{(ch)} & Z_{ch} \end{bmatrix} . \begin{bmatrix} \mathbf{u}^{(1)} \\ U_{ch} \end{bmatrix} + R_b U_g^2, \tag{21}$$

where $\mathbf{C}_1^{(ch)}$ is the matrix accounting for the coupling between the glottal chink and the vocal tract. It is a sparse row vector with two non-zero elements:

$$\mathbf{C}_1^{(ch)} = [Z_{(1,ch)}, b_1, 0, \ldots, 0].$$

3.4. General form of the extended single-matrix formulation of the vocal tract

Finally, the general form of the extended single-matrix formulation (ESMF) of the vocal tract is

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \vdots \\ \mathbf{f}^{(\mathcal{N})} \\ F_{ch} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(2)T} & \dots & \mathbf{C}_1^{(\mathcal{N})T} & \mathbf{C}_1^{(ch)T} \\ \mathbf{C}_1^{(2)} & \mathbf{Z}^{(2)} & & \mathbf{C}_2^{(\mathcal{N})T} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{C}_1^{(\mathcal{N})} & \mathbf{C}_2^{(\mathcal{N})} & & \mathbf{Z}^{(\mathcal{N})} & \mathbf{0} \\ \mathbf{C}_1^{(ch)} & \mathbf{0} & \dots & \mathbf{0} & Z_{ch} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \vdots \\ \mathbf{u}^{(\mathcal{N})} \\ U_{ch} \end{bmatrix} + \mathbf{Q}\mathbf{u}_Q \quad (22)$$

where $\mathcal{N}$ is the number of waveguides modeling the whole vocal tract. In addition to the single-matrix formulation defined in Eq. (4), two new configurations detailed in this section are considered: i) (n) is an anabranch of (m), i.e. it is connected to (m) at two different points, the upstream connection K_u and the downstream connection K_d. Then $\mathbf{C}_m^{(n)}$ is a sparse matrix whose non-zero elements are $c_{1,K_u} = b_{K_u}^{(m)}$, $c_{1,K_u+1} = Z_C^{(m,n)}$, $c_{N+1,K_d+2} = b_{K_d+1}^{(m)}$, and $c_{N+1,K_d+1} = Z_{C+1}^{(m,n)}$, and where N is the generic number of tubelets that model (n), and ii) (n) is the glottal chink. In that case, it is connected to (1) and $\mathbf{C}_1^{(ch)} = [Z_{(1,ch)}, b_1, 0, \dots, 0]$.

4. Numerical simulations

In order to validate the simulation framework presented in the previous section, this section provides examples of synthesized vowels in static configurations. Area functions are derived from X-ray films comprising short French sentences [38]. They are obtained by dividing the vocal tract shape into tubelets perpendicular to the vocal tract centerline, determined via a specified algorithm [39], and then applying α β transformations to recover the area [40]. Chosen parameters of the vocal folds model (see Tab. 2) lie in the typical range of values found in the literature [7, 41, 42].

Table 2: Input parameters for the vocal folds model

Parameter	Unit	Value
Subglottal pressure P_{sub}	Pa	800
Position of mass 1 x_1	mm	0.2
Position of mass 2 x_2	mm	3.2
Vocal fold thickness d_g	mm	3
Vocal fold length l_g	mm	10
Opening at point 0 h_0	mm	40
Vocal folds abduction h_{ab}	mm	2.5
Nominal mass m_1	g	0.1
Nominal stiffness k_1	N/m	80
Nominal mass m_2	g	0.125
Nominal stiffness k_2	N/m	80
Nominal damping coefficient r_i	kg.rad.s^{-1}	$0.2\sqrt{k_i m_i/2}$
Coupling spring k_c	N/m	$k/2$

In the following numerical simulations, the vocal folds are supposed symmetric, namely the upper vocal fold has the same mechanical parameters as the lower one.

Note that when a waveguide is closed at its end, e.g. the piriform fossae, a termination area set to 0 may cause the numerical computation to break because of infinite or NaN values. Setting the termination area to a very small value, for instance the order of magnitude of the unit roundoff, is sufficient to efficiently approximate a closed termination. A similar technique can be used

when sudden modifications of the global structure of the network occur, such as a sudden occurrence of nasalization or bilateralization: setting the input area function of a certain waveguide to very small values makes it temporally shunted from the network. Doing so prevents discontinuities that may cause undesired artifacts.

4.1. Anastomozing waveguides

The acoustics of lateral consonants has been previously studied mainly by using single lateral channel models [36, 37]. Consequently, only a few previous works focused on the acoustic effect of bilateralization. To validate the bilateral model presented in this paper, it is compared to results provided by Zhang *et al.* [34, 35], using the same area functions. Fig. 5 shows the single-tube model used for the simulation in [35]. The area functions of the different parts are derived from MRI [34], and are averaged over the length of each section of the single-tube model (see Fig. 5). Dimensions can be found in the original paper [34]. When the supralingual cavity is taken into account, the latter is connected to the main oral tract at the same location as the lateral channel. Consequently, the supralingual cavity is a twin waveguide of the lateral channel.

From the area functions provided by the original paper, transfer functions of the vocal tract are computed thanks to the method used in [2, 3], which consists in simulating a sudden glottal closure and computing the Fourier transform of the acoustic response to the step-down glottal excitation. The vocal tract acoustic response functions are computed for different configura-

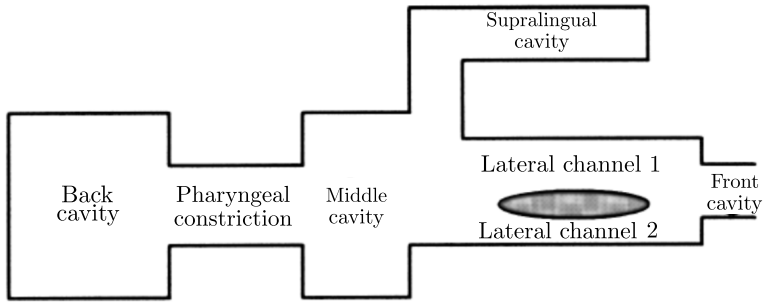


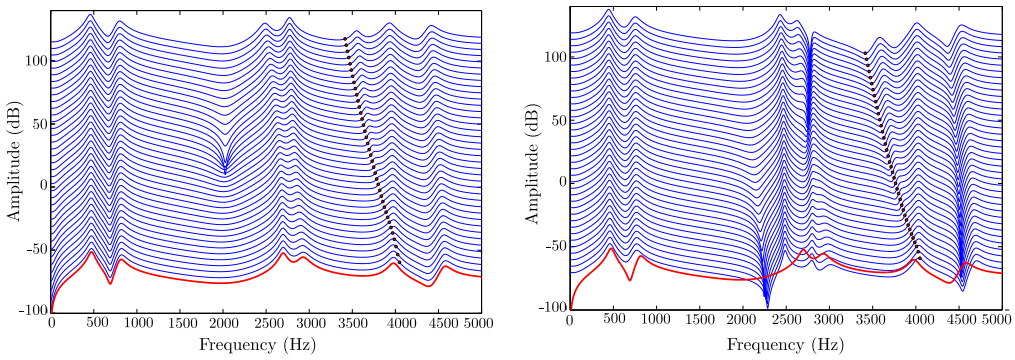
Figure 5: Single-tube model used for the simulation of bilaterals. Figure adapted from [35].

tions of bilateralization. They are defined by an asymmetry factor

$$\gamma = \frac{l_{c_2}}{l_{c_1}},$$

where l_{c_2} and l_{c_1} denote the length of both lateral channels. During the simulations, l_{c_1} is set to 4.2 cm, and l_{c_2} is increased to raise the asymmetric factor γ. Simulations include two main configurations: with and without the consideration of the supralingual cavity.

Fig. 6 shows transfer functions for several configurations obtained with ESMF. The asymmetric factor γ varies from 1 (bottom curve), corresponding to the symmetric configuration, to 1.37 (top curve). The increment step of γ between two successive curves is 0.01. The thick line at the bottom is the transfer function of the vocal tract where both lateral channels are merged into a single channel, and without the consideration of the supralingual cavity. The computed transfer functions agree with those obtained in [35] with an independent frequency-based method. For instance, in both configurations (with and without the suparalingual cavity), the bilateralization introduces a pole/zero around 4 kHz. The frequency of the pole/zero pair drops as the length of the anabranch increases. This is in agreement



(a) Without the supralingual cavity

(b) With the supralingual cavity

Figure 6: Transfer functions of the vocal tract in several configurations using the extended single-matrix formulation, with (right) and without (left) consideration of a supralingual cavity. The bottom curve (thick line) is the transfer function of the vocal tract where both lateral channels are merged into a single channel and without the supralingual cavity. The transfer functions of the vocal tract with bilateralization are plotted from bottom to top, where the asymmetric length factor varies from 1 (bottom) to 1.37 (top). The increment step of the length asymmetry factor between two successive curves is 0.01. Theoretical positions of zeros introduced by the bilateralization are denoted by circle marks.

with the theory, which relates the frequency of the pole/zero introduced by the bilateralization with the resonance frequency of an equivalent tube having a length equal to the sum of the lateral channels [43], hence a drop of the frequency as the length increases. The theoretical zero frequencies are represented in Fig. 6 by circle marks: they match the zero frequencies of the simulated vocal tract acoustic response functions.

The length asymmetry of the bilateral channels slightly impacts the formant frequencies. F_5 is the formant for which the effect is the most predominant. It seems to be due to its proximity to the introduced pole/zero pair.

Finally, the introduction of the supralingual cavity in the model gives rise to another pole/zero pair slightly above 2.7 kHz. As observed in [35], this corresponds to the first quarter-wavelength resonance of a 3.1 cm-long tube, around 2700 Hz. The supralingual cavity lowers the formant frequencies, and especially F_3 and F_4, which are close to the zero introduced by the supralingual cavity.

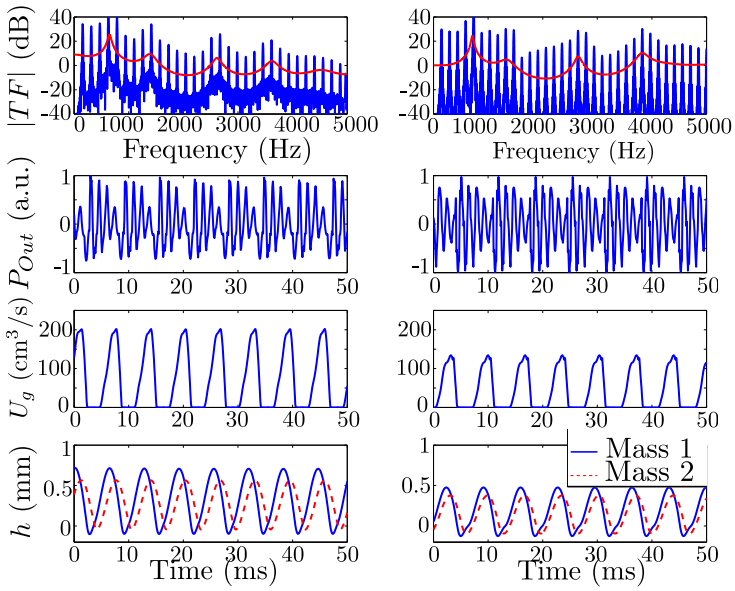
The agreement between the global effects of bilateralization, and the supralingual cavity on the transfer function obtained with ESMF and those obtained with the independent frequency based technique used in [35], proves the efficiency of the method to deal with bilateralization. Since it is suitable for time-domain continuous speech synthesis, the method could be useful to thoroughly investigate the acoustic effect of the tongue movement during the production of bilateral consonants.

4.2. Effect of the acoustic model on the motion of the vocal folds

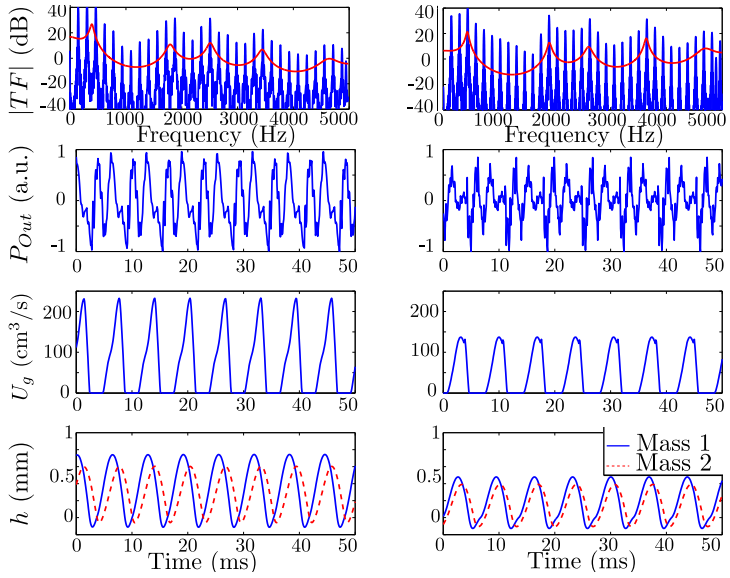
In this section, numerical simulations are used to study the validity of ESMF when connected to a self-sustaining model of the vocal folds. The method is compared with the concurrent approach using *reflection-type line analog* (RTLA) models [1, 44]. This approach is widely used and validated when connected with self-oscillating models of the vocal folds, including the 2×2-mass model with smooth contours used in this paper [14, 15, 31]. Thus, simulations consist in computing the motion of the vocal folds via the model described in Sec. 2.3 connected to several configurations of the vocal tract, using two models of acoustic propagation: ESMF and RTLA. The configurations of the vocal tract are area functions corresponding to 6 French vowels: /i/, /e/, /ɑ/, /ø/, /o/, and /u/.

Fig. 7 shows the spectrum and the spectral envelope of the synthesized output pressure (P_{Out}) radiated at the lips, as well as the glottal flow and the glottal opening at the location of mass 1 and mass 2, both by RTLA and ESMF, for 2 French vowels (/a/ and /e/). For each vowel, the peaks of the spectral envelopes, i.e. the formants, obtained from both methods are similar. One can also observe similarities in the global shape of the other waveforms, namely P_{Out}, U_g, and the motion of the vocal folds.

However, the comparison also highlights the differences between both approaches. For instance, the obtained fundamental frequency is lower with ESMF than with RTLA. The amplitude of the vocal folds' motion, and the phase shift between both masses are also lower with ESMF (see Fig. 8). This yields to a lower amplitude of the glottal flow. Such differences are somewhat expected since the coupling with the acoustic propagation may be



(a) /a/: RTLA (left) and ESMF (right)



(b) /e/: RTLA (left) and ESMF (right)

Figure 7: The results of the simulations for 2 French vowels, /a/ and /e/. From top to bottom, the left column shows the spectrum and spectral envelope of the synthesized vowel via RTLA, as well as the output pressure (P_{Out}) radiated at the lips, the glottal flow U_g and the glottal opening at the location of mass 1 (solid line), and mass 2 (dashed line). The right column displays the same quantities computed thanks to ESMF. The spectral envelope is computed using a 10-order LPC (*Linear Predictive Coding*) [45]

modified, due to the different models of acoustic losses. It is accepted that RTLA has the advantage of better accounting for the frequency dependence of the acoustic losses and the acoustic radiation. This may also explain the differences between formant bandwidths obtained with RTLA and ESMF.

It is worth noting that slight modifications of the input parameters of the vocal folds suffice to recover similar behaviors of the vocal folds. This consists in multiplying the values of the spring stiffness by a factor, say Q, such that the obtained fundamental frequency matches that obtained with RTLA. This modified version of ESMF is labeled ESMF* in the paper. This is highlighted by Fig. 8, which compares several physical quantities derived from the motion of the vocal folds, obtained with RTLA and ESMF, both with and without modifications of the input parameters, for all 6 vowels. The quantities A_{m_1} and A_{m_2} are the mean absolute values of the amplitude of oscillation of mass 1 and mass 2, respectively. ESMF* corresponds to the values obtained using the ESMF method with modified input parameters of the vocal folds model.

Fig. 8 shows that, although both methods simulate qualitatively similar oscillations of the vocal folds and glottal flow, they are quantitatively different. The main differences lie in the phase shift and in the open quotient: when connected to RTLA, the phase shift between the rear and front parts of the vocal folds is larger than when they are connected to ESMF. There is no significant variation when the values of the mass and stiffness are modified. The open quotient is also significantly different when the vocal folds are connected to RTLA. However, in this case, the modification of the input parameters modifies the open quotient. Also, for both techniques, the ampli-

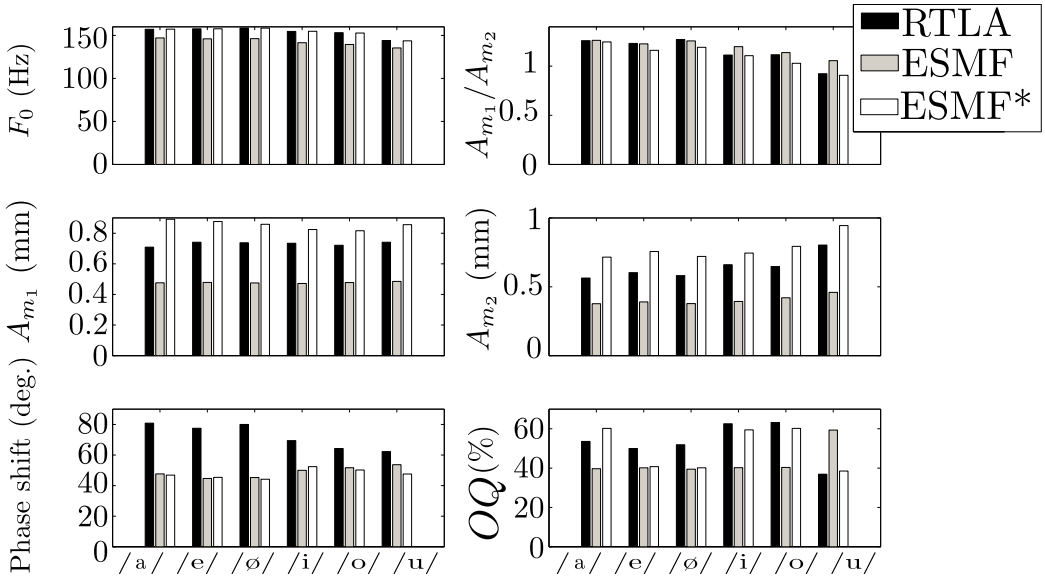


Figure 8: Comparison of quantities derived from the motion of the vocal folds obtained with the different methods. ESMF* corresponds to the quantities obtained with modified mass and stiffness of the vocal folds model.

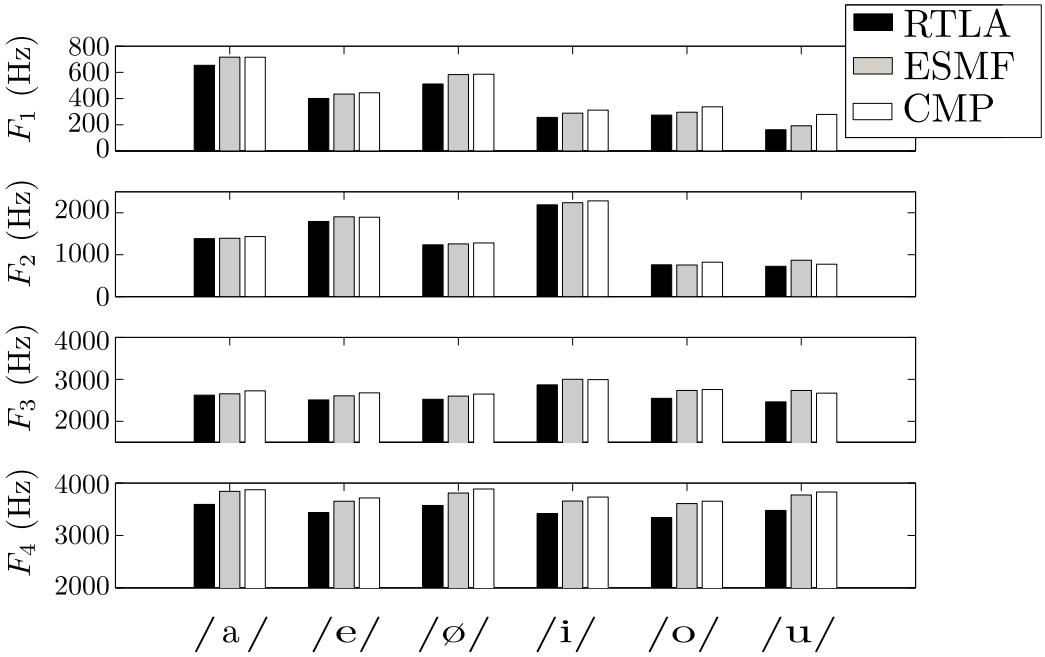


Figure 9: Formant frequency of the first 4 formants obtained with the different methods for 6 French vowels. Results are compared with values obtained with the chain-matrix paradigm (CMP) [27] for the corresponding area functions.

tude ratio of the vocal folds' oscillations decreases for close vowels (/i/ and /u/). Since A_{m_1} does not significantly change, it is mainly due to the rise of the amplitude of mass 2, namely the rear part of the vocal folds, which is directly connected to the vocal tract. Consequently, the presence of a supra-glottal constriction seems to have a strong effect on the downstream part of the vocal fold.

Fig. 9 shows the effect of the methods on the formant frequencies. The values obtained with the different methods are compared with values obtained from the transfer function of the corresponding vocal tracts using the chain-matrix paradigm (CMP) [27]. This frequency-based method is used as

a reference for computing the VT resonance frequencies independently of the glottal source. The formant frequencies obtained with RTLA are globally lower than those obtained with ESMF. In comparison with the resonance frequencies of the vocal tract derived with CMP, EMSF gives closer formant frequencies than RTLA.

To summarize, the presented ESMF has shown its accuracy in computing the acoustic propagation and in accounting for the coupling between the vocal tract and the glottal source when the latter is modeled by self-sustaining models, such as classic two-mass models. The comparison of the method with the classic reflection-type line analog model (RTLA) shows qualitatively similar oscillations of the vocal folds. However, they are quantitatively slightly different: both methods give different fundamental frequencies, amplitudes of oscillations, phase shifts, and open quotients. Such quantitative differences have been previously observed for the two-mass model between predictions of RTLA and mechanical models [15, 46]. Some of these quantities may be modified by adjusting the values of the mass and stiffness of the two-mass model by a factor Q. Besides, the formant frequencies computed from the vowels simulated with ESMF are closer to the resonance frequencies of the vocal tract obtained by CMP. Consequently, since ESMF presents the advantage of easily dealing with dynamic geometries of the vocal tract, including length variations and the connecting of numerous side cavities, it is a complete simulation framework useful to either synthesize natural continuous speech or to qualitatively study the coupling between the vocal tract and the vocal folds in the context of continuous speech.

4.3. Effect of the glottal chink

This section presents a short study about the effect of the glottal chink on the acoustic parameters. It consists in computing the glottal flow and the motion of the vocal folds coupled with a static configuration of the vocal tract and a linearly increasing length of the glottal chink. We chose to represent the effect of the glottal chink on /a/ and /i/. To modify the size of the glottal chink, we modify the length l_{ch} via a ratio $\alpha \in [0,1]$ such that $l_{ch} = \alpha l_t$ (see Fig. 4). The quantities l_t and h_{ab} are kept constant, and by deduction, $l_g = (1 - \alpha)l_t$.

Fig. 10 shows the simulation results. It displays the narrow-band spectrogram, the output pressure radiated at the lips P_{Out}, the total volume velocity through the glottis U_t, the motion of the vocal folds, and the length of the glottal chink l_{ch}. The effect of the glottal chink on these physical quantities is clearly seen in both configurations (/a/ for the left column, and /i/ for the right column). For instance, as expected, the presence of the glottal chink stops U_t from being null, since the glottis is never totally closed. The waveform of U_t is then the superimposition of the glottal flow U_g inside the vibrating part of the vocal folds and an offset corresponding to U_{ch}. The U_g component behaves similarly to the case of a totally closed chink, and U_{ch} increases as l_{ch} increases, i.e. when the glottal chink opens up. The effect is more visible in Fig. 11, which shows details of a few phonatory cycles at two different time instants, corresponding to a glottal configuration without chink, and with a chink of length 0.5 cm. A DC component of the glottal flow waveform appears for the open chink configuration. The value of the DC component for this case is around 30 cm^3/s. The amplitude of the oscillatory

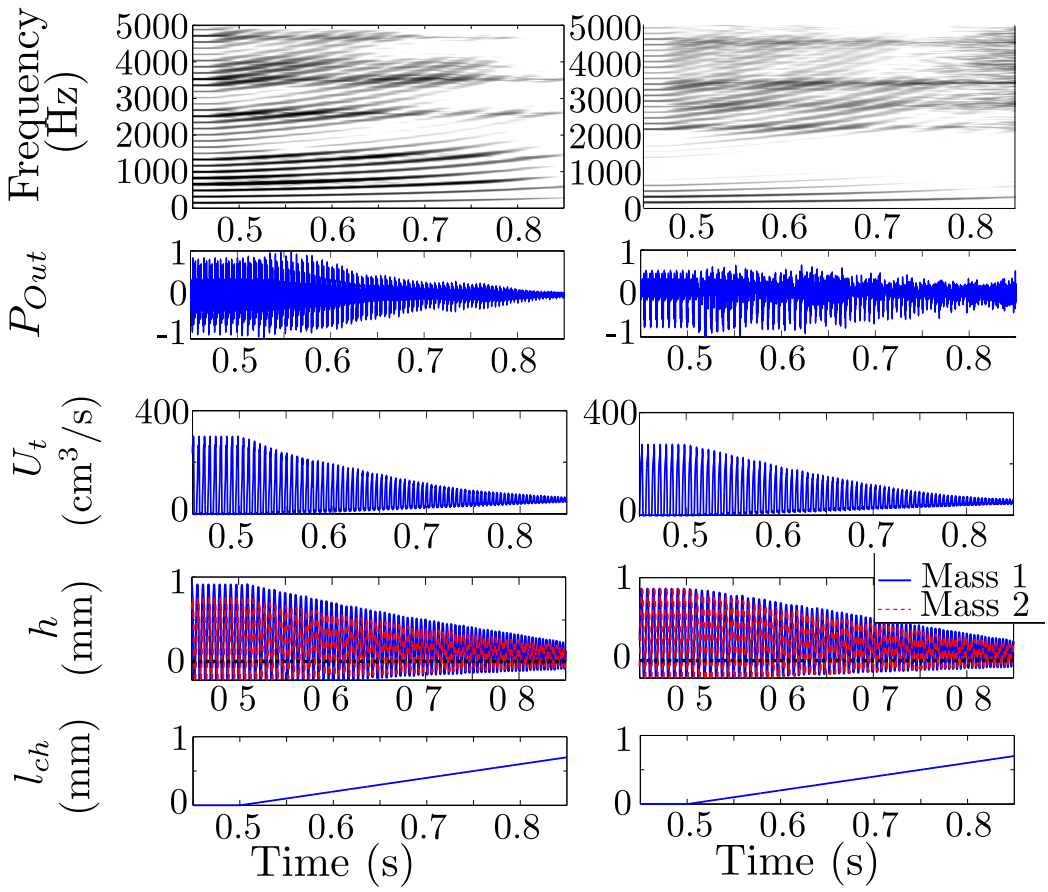


Figure 10: The results of the simulation for two French vowels /a/ (left column), and /i/ (right column). From top to bottom, the left column represents the narrow-band spectrogram, the output pressure, radiated at the lips, the volume velocity inside the glottis ($U_t = U_g + U_{ch}$), the glottal opening at the location of mass 1 (solid line) and mass 2 (dashed line), and the length l_{ch} of the glottal chink, computed for /a/. The right column represents these quantities computed for /i/.

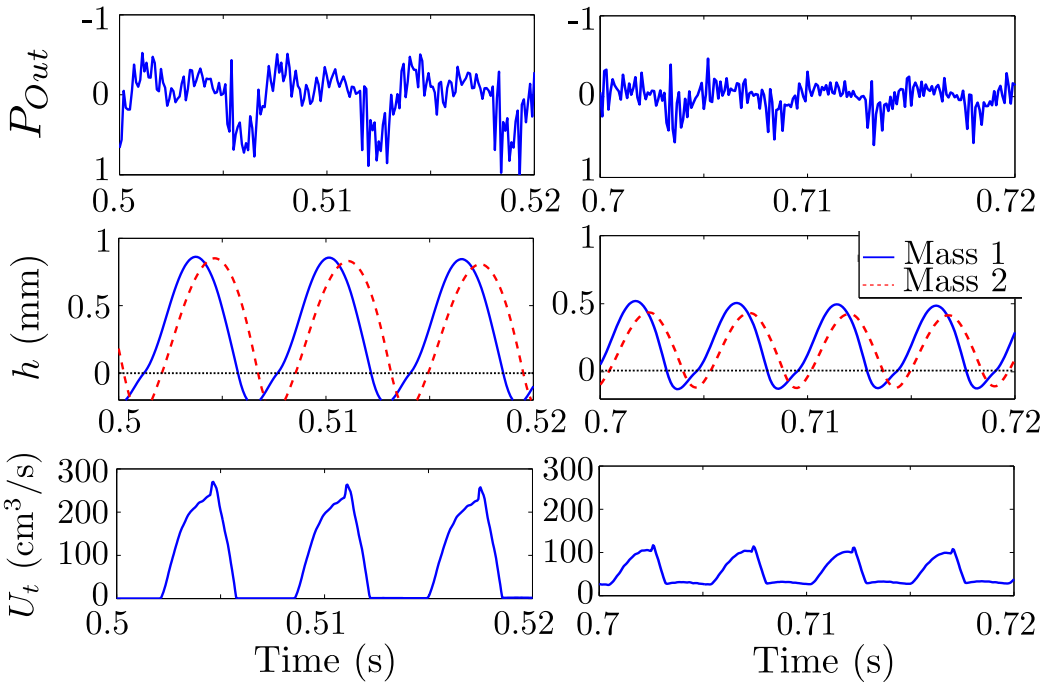


Figure 11: Details of phonatory cycles of the simulated sustained /i/ at two time instants: (left) closed glottal chink, and (right) open glottal chink, $l_{ch} = 0.5$ cm. From top to bottom: P_{Out}, opening height of the vocal folds, and U_t.

quantities (P_{Out}, U_t, and h) are also considerably reduced when the glottal chink is open compared to the closed chink configuration: the maximal opening height of the vocal folds goes approximately from 0.8 mm to 0.5 mm, and the maximal glottal flow value from 250 cm^3/s to 100 cm^3/s.

Due to a smaller length of the vibrating part of the vocal folds when the chink appears, the amplitude of U_g vanishes. As a consequence, the amplitude of motion of the vocal folds decreases. This finally results in a fading P_{Out}. When the length of the chink l_{ch} is larger than l_g, which occurs for $t > 0.75$ s, the U_g component is very weak in relation to the DC component imposed by U_{ch}. For /a/, which is an open vowel, it results in a very weak

output signal. On the other hand, for the close vowel /i/, this results in the generation of frication noise, due to an high airflow passing through the supraglottal constriction. In that case, P_{Out} contains more energy, in the mid and high frequency range, from the frication noise source than from the glottal source. This can be seen in the spectrogram of the simulated /i/ for $t > 0.75$.

Since the length of the vibrating vocal folds drops due to the presence of the glottal chink, this raises the fundamental frequency of the produced utterance. This is evidenced by the trajectory of harmonics in the narrow-band spectrograms of both vowels. The formant frequency pattern seems to be barely modified by the presence of the chink.

The presented simulation clearly shows the effect of a glottal chink on some acoustic parameters. This confirms the interest of such models to thoroughly investigate the acoustic or phonatory phenomena involved in speech production. Modeling the partially closed glottis is also important for the synthesis of breathy and/or pathological voices, and for the realistic synthesis of voiced fricatives. If used together with a realistic glottis-vocal tract coordination model, it may also be useful to investigate the transition between voiced/voiceless sounds, as shown in [21].

5. Conclusions

This paper has presented the theoretical aspects for extending the single-matrix formulation [3] of the vocal tract to a more general and complete tool. The presented framework allows the connection of self-oscillating models of the vocal folds, or the connection of a glottal system made up of

self-oscillating vocal folds and a glottal chink. It can also account for anastomozing waveguides to study the acoustic effect of bilateralization.

The accuracy of the simulation framework to account for bilateral channels has been studied by computing transfer functions of vocal tract configurations derived from a previously published paper [35]. The effects of the bilateralization on the simulated transfer functions quantitatively agrees with the theoretical predictions [35, 43]. Acoustic features of lateral consonants, such as the introduction of an antiresonance due to asymmetric bilateral channels, may thus be reproduced using ESMF.

The connection of the self-oscillating model of the vocal folds has been tested with a 2×2-mass model with smooth contours. In comparison with the widely accepted reflection-type line analog model of acoustic propagation, the connection of the vocal folds to the single-matrix formulation yields to qualitatively similar behaviors of the vocal folds, though it may quantitatively modify their oscillations. However, at this stage, it is not possible to conclude whether these differences are important for quantitative studies. Simulations also revealed that the formant frequencies of the vowels obtained with ESMF are closer to the resonance frequencies of the vocal tract than those of the vowels simulated via the reflection-type line analog model.

The paper also provides simulations with various lengths of the glottal chink. They clearly show the acoustic effects, including the appearance of a DC component in the glottal flow waveform. As the glottal chink becomes larger, this generates a large DC acoustic flow inside the vocal tract, which may lead to the generation of a frication noise in addition to the voiced glottal source if there is a supraglottal constriction in the vocal tract. This feature

is important to simulate voiced fricatives.

The simulation framework is intended to be used for the synthesis of natural continuous speech, *i.e.* natural phrase-level utterances. For instance, it has been used to validate a two-dimensional model of the velum for the copy synthesis of utterances containing nasal phonemes [47]. The transmission line circuit analog model enables the dynamic variations of the vocal tract geometry and its complexity to be accurately taken into account. Consequently, it can easily support realistic geometries and dynamic deformations of the vocal tract, which constitutes its main advantage. Thanks to the contributions of the paper, it is now possible to account for a realistic coupling between the vocal folds and the vocal tract, as well as the possibility of including glottal leakage. This last point allows the different types of phonation to be realistically simulated. Although it has only been briefly discussed in the paper, the developed approach is compatible with the subglottal network proposed in [32]. Combining the subglottal and supraglottal networks would be a step forward in the implementation of a complete and realistic speech synthesizer. This would create a useful tool to study the phenomena involved in speech production. This paper may be considered as a basis to make such investigations. More specifically, it may be used to relate acoustic cues 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.

Acknowledgements

This study is supported by the ANR (*Agence Nationale de la Recherche*) ArtSpeech project. The authors would like to sincerely thank Dr. Shinji Maeda and Dr. Parham Mokhtari, for their useful advice and fruitful discussions.

References

[1] J. L. Kelly, C. C. Lochbaum, Speech synthesis, in: Proceedings of the Fourth International Congress on Acoustics, 1962, pp. 1–4.

[2] S. Maeda, A digital simulation method of the vocal-tract system, Speech Communication 1 (1982) 199–229.

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

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

[5] P. Birkholz, Enhanced area functions for noise source modeling in the vocal tract, in: 10th International Seminar on Speech Production, Köln, 2014, pp. 1–4.

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

[7] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands, 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) (1994) 3416–3431.

[8] B. D. Erath, S. D. Peterson, M. Zañartu, G. R. Wodicka, M. W. Plesniak, A theoretical model of the pressure field arising from asymmetric intraglottal flows applied to a two-mass model of the vocal folds, J. Acoust. Soc. Am. 130(1) (2011) 389–403.

[9] S. R. Moisik, J. H. Esling, Modeling the biomechanical influence of epilaryngeal stricture on the vocal folds: A low-dimensional model of vocal–ventricular fold coupling, Journal of Speech, Language, and Hearing Research 57 (2) (2014) S687–S704.

[10] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, D. Mürbe, Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall, Biomechanics and modeling in mechanobiology 14 (4) (2015) 719–733.

[11] B. H. Story, Vowel acoustics for speaking and singing, Acta Acustica united with Acustica 90 (4) (2004) 629–640.

[12] B. J. Kröger, P. Birkholz, Articulatory synthesis of speech and singing: State of the art and suggestions for future research, in: Multimodal Signals: Cognitive and Algorithmic Issues, Springer, 2009, pp. 306–319.

[13] B. H. Story, I. R. Titze, Voice simulation with a body-cover model of the vocal folds, J. Acoust. Soc. Am. 97(2) (1995) 1249–1260.

[14] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, A. Hirschberg, A symetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prothesis design, Acta Acustica 84 (1998) 1135–1150.

[15] L. Bailly, X. Pelorson, N. Henrich, 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) (2008) 3296–3308.

[16] P. Birkholz, 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.

[17] Y. Laprie, R. Sock, B. Vaxelaire, 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, EDP Sciences, 2014, pp. 1285–1298.

[18] P. Birkholz, B. Kröger, P. Birkholz, A survey of self-oscillating lumped-element models of the vocal folds, Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung (2011) 47–58.

[19] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, S. D. Peterson, A review of lumped-element models of voiced speech, Speech Communication 55 (5) (2013) 667–690.

[20] D. H. Klatt, L. C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers, J. Acoust. Soc. Am. 87(2) (1990) 820–857.

[21] B. Elie, Y. Laprie, A glottal chink model for the synthesis of voiced fricatives, in: ICASSP, 2016.

[22] B. Cranen, J. Schroeter, Modeling a leaky glottis, Journal of Phonetics 23 (1–2) (1995) 165 – 177.

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

[24] R. Wilhelms-Tricarico, A modified two-mass model of the vocal folds with a chink and gradual closure, speech Communication Group Working Papers (1994).

[25] M. Zañartu, G. E. Galindo, B. D. Erath, S. D. Peterson, G. R. Wodicka, 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) (2014) 3262–3271.

[26] 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.

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

[28] M. Zañartu, L. Mongeau, G. R. Wodicka, Influence of acoustic loading on an effective single mass model of the vocal folds, J. Acoust. Soc. Am. 121(2) (2007) 1119–1129.

[29] I. T. Tokuda, J. Horáček, J. G. Švec, H. Herzel, Comparison of biomechanical modeling of register transitions and voice instabilities with excised larynx experiments, J. Acoust. Soc. Am. 122(1) (2007) 519–531.

[30] R. Schwarz, M. Döllinger, T. Wurzbacher, U. Eysholdt, J. Lohscheller, Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model, J. Acoust. Soc. Am. 123(5) (2008) 2717–2732.

[31] C. Vilain, X. Pelorson, C. Fraysse, M. Deverge, A. Hirschberg, J. Willems, Experimental validation of a quasi-steady theory for the flow through the glottis, J. of Sound and Vibration 276(3–5) (2004) 475 – 490.

[32] J. C. Ho, M. Zañartu, G. R. Wodicka, An anatomically based, time-domain acoustic model of the subglottal system for speech production, J. Acoust. Soc. Am. 129(3) (2011) 1531–1547.

[33] S. S. Narayanan, A. A. Alwan, K. Haker, Toward articulatory-acoustic models for liquid approximants based on mri and epg data. part i. the laterals, J. Acoust. Soc. Am. 101(2) (1997) 1064–1077.

[34] Z. Zhang, C. Y. Espy-Wilson, M. Tiede, Acoustic modeling of American English lateral approximants, in: Proceedings of the Eighth English Eurospeech Conference, 2003.

[35] Z. Zhang, C. Y. Espy-Wilson, A vocal-tract model of american english /l/, J. Acoust. Soc. Am. 115(3) (2004) 1274–1280.

[36] K. Stevens, Acoustic Phonetics, MIT Press, Cambrige, MA, 1998.

[37] S. Narayanan, D. Byrd, A. Kaun, Geometry, kinematics, and acoustics of tamil liquid consonants, J. Acoust. Soc. Am. 106(4) (1999) 1993–2007.

[38] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, J. Sturm, DOC-VACIM 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.

[39] Y. Laprie, M. Loosvelt, S. Maeda, E. Sock, 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.

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

[41] D. Sciamarella, C. d'Alessandro, On the acoustic sensitivity of a symmetrical two-mass model of the vocal folds to the variation of control parameters, Acta Acustica united with Acustica 90(4) (2004) 746–761.

[42] L. Bailly, N. Henrich, X. Pelorson, Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modelinga), J. Acoust. Soc. Am. 127(5) (2010) 3212–3222.

[43] A. Prahler, Analysis and synthesis of the American English lateral consonant, Ph.D. thesis, MIT, Cambridge, Massachussets (1998).

[44] P. Meyer, R. Wilhelms, H. W. Strube, A quasiarticulatory speech synthesizer for german language running in real time, J. Acoust. Soc. Am. 86(2) (1989) 523–539.

[45] S. McCandless, An algorithm for automatic formant extraction using linear prediction spectra, IEEE Trans 22 (1974) 135–141.

[46] N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, 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) (2007) 479–490.

[47] Y. Laprie, B. Elie, 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.