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

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 vocal folds with 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 chinks, to be connected to the vocal tract. It also integrates the case of a local division of the main air path into two lateral channels, as it may occur during the production of lateral approximants. Extensions are detailed by a reformulation of the acoustic conditions at the glottis, and at the upstream connection of bilateral channels. Numerical simulations are provided to validate the simulation framework. The introduction of a zero due to the presence of bilateral channels is confirmed by the simulations. The position of the zero 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 single-matrix formulation or to the classic reflection type line analog model. Finally, the acoustic

*Corresponding author. Tel: +33 383593036

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

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 [1–4] and/or the self-sustaining oscillations of the vocal folds [5–7]. In comparison with finite element based methods [8], which require a huge amount of time, their low computation time make them interesting for continuous speech synthesis.

The acoustic models use the assumption of a one-dimensional wave propagation, generally planar wave, along a set of acoustic tubes. The dimensions of the elementary tubes (or *tubelets*) approximate the geometry of the vocal tract. In regards with the typical dimensions of the human vocal tract, these models are valid up to frequencies around 5 kHz. This limit is acceptable for the simulation of speech production, for which most of the information lies under 5 kHz.

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 gestures and their temporal coordination, that of the anatomic characteristics of the human vocal tract, and of the in-

teractions between the vocal folds and the vocal tract. In order to enable studying speech production 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...) 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 brings out 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.

RTLA computes the forward and backward pressure waves using the impedance discontinuities at the tube junctions. It has the advantage of accurately modeling the acoustic losses and the acoustic radiation by applying discrete filters to the scattering equations, and thus accounting for their frequency dependence. However, there are constraints on the dimensions of the tubelets that model the vocal tract in regards with the chosen simulation frequency F_s: the latter must be set so that the acoustic wave travels a distance equal to the length of each tubelet, say l. This yields the constraint $F_s = c_s/l$, where c_s is the sound celerity. 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. RTLA is widely used to study the self-sustained motion of the vocal folds [9–11] coupled with simplified acoustic resonators, but the aforementioned limitation makes its use in continuous speech synthesis difficult when dealing with realistic geometries of the vocal tract [4]. Note that using RTLA with time-varying length of the vocal tract may be possible by changing the sampling frequency at each time step, and by accurately resampling the simulated utterance, as proposed in [12]. However, this technique does still not overcome the limitation of an evenly sampled vocal tract, which may be problematic when dealing with complex geometries of the vocal system, *e.g.* when numerous side cavities are connected considered.

On the other hand, many continuous speech synthesizers use TLCA [2, 5, 13, 14]. It is based on the electric-acoustic analogy: the vocal tract acoustics is seen as a lumped electric circuit. The acoustic and aerodynamic elements of each tubelet are modeled by circuit elements. Unlike RTLA, it can easily deal with length variations since the dimensions of the acoustic tubelets are independent. However, this analogy does not allow the frequency dependence of the acoustic losses and the acoustic radiation to be accurately taken into account. Another difficulty, originally encountered by TLCA users, was the connection with more than one side branch. Indeed, the simulation method by Maeda [2] cannot include the simultaneous connection of several side branches. This issue has been overcome recently by Mokhtari *et al.* [3] via the reformulation of the equations governing the acoustic propagation into a

system matrix. This formulation, called *Single-Matrix Formulation* (SMF), supports the connection of any number of side branches. Consequently, it is a useful tool to study the acoustic effects of the numerous side cavities (piriform fossae, sublingual cavity, paranasal sinuses...) in the context of continuous speech synthesis. Yet, SMF, as presented in [3], presents some limitations: 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 approximants, is not discussed.

Starting from the single-matrix formulation presented in [3], this paper details the theory and the methodology for extending the SMF by overcoming the aforementioned limitations. 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 model enabling a partial glottal closure.

The main aspects of the simulation framework, called *Extended Single-Matrix Formulation* (ESMF), are outlined by the organization of the paper. 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 vocal folds. Then, Sec. 3 details the mathematical formulations for introducing the case of anastomozing waveguides into the single-matrix formulation, as well as the math-

ematical formulations to connect self-oscillating models of vocal folds and a glottal chink to the single-matrix formulation. Finally, numerical simulations present, in Sec. 4, the accuracy of the extended single-matrix formulation to deal with the new features.

2. Theoretical background

This section describes 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, enabling the simulation of the 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 are 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*) for computing the acoustic propagation inside it. The length and the cross-sectional areas of the tubelets are such that they approximate the vocal tract geometry. The single-matrix formulation [3] uses the *transmission line circuit analog* approach, which is preferred to the *reflection type line analog* model. This is motivated by the fact that the latter does not include the possibility to easily deal with length variations of the vocal tract. It is consequently hardly suitable for continuous speech synthesis with realistic dynamic vocal tract geometries. In transmission line circuit analog models,

each tubelet is modeled by lumped circuit elements. Fig. 1 shows the chosen lumped circuit elements of a single tube section and Tab. 1 lists the lumped circuit elements of the acoustic-electric analogy.

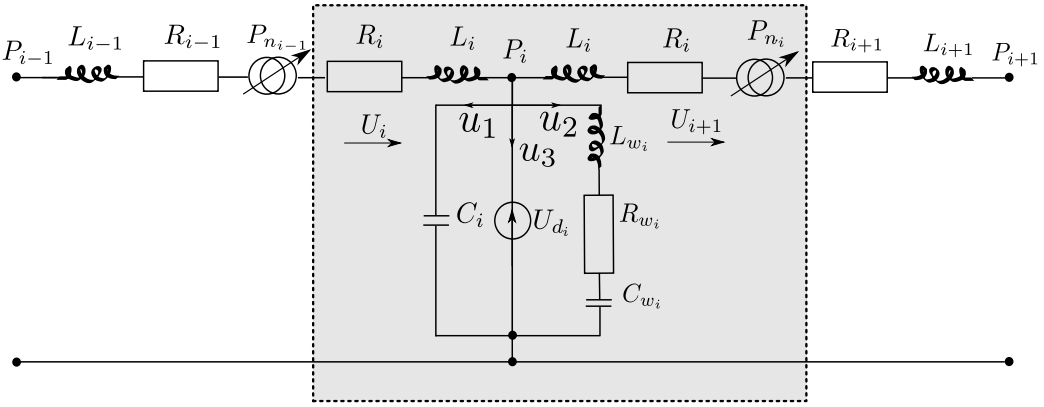


Figure 1: Lumped circuit element of an acoustic tube. The acoustic-electric analogy is detailed in Tab. 1.

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 [13], 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, indexes follow the air flow direction. For instance, considering the vocal tract, index 1 denotes the tubelet connected to the glottis, and index N denotes the tubelet corresponding to the lip termination.

The 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} [15–17], that is active when the air flow is considered as turbulent, namely when the Reynolds number Re is above a certain threshold Re_c.

This paper follows the formula provided in [15] to compute the internal

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}	Fricative noise source

resistance R_{n_i}:

$$R_{n_i} = \kappa\rho\frac{U_{DC}}{a_{i-1}^2} + 8\pi\mu\frac{l_{i-1}}{a_{i-1}^2}, \tag{1}$$

where the term κ denotes a scaling coefficient set to 1.42, as used in [5, 15], and U_{DC} is the low frequency component of the air flow [15]. Finally, the noise level P_{n_i} is

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

where ξ is an arbitrarily adjustable real constant used to control the noise level, and w is a Gaussian white noise to which a first-order lowpass and third-order highpass filters have been applied [15]. The choice of the thresholds Re_c for the study is 2700, and $\xi = 10^{-6}$, as suggested by Sondhi and

Schroeter [16]. The frication noise source is usually located at the next point downstream the supraglottal constriction [15–17], hence the spatial lag $i-1$ in Eqs. (1) and (2).

2.2. Equations

Considering the lumped circuit elements displayed in Fig. 1, the time-domain simulation consists in solving the following equations at each time point t

$$\begin{aligned} P_{i-1} - P_i &= \frac{\partial}{\partial t}\left[\left(L_{i-1} + L_i\right) U_i\right] + \left(R_{i-1} + R_i + R_{n_{i-1}}\right) U_i + P_{n_{i-1}} \\ U_i - U_{i+1} &= u_1 + u_2 + u_3, \end{aligned} \tag{3}$$

where

$$u_1 = \frac{\partial}{\partial t}\left[C_i P_i\right], \tag{4}$$

$$u_3 = -U_{d_i}, \tag{5}$$

$$P_i = \frac{\partial}{\partial t}\left[L_{w_i} u_2\right] + R_{w_i} u_2 + \int_0^t \frac{u_2}{C_{w_i}} dt. \tag{6}$$

The paper follows the discrete representation of Eq. (3) introduced by Maeda [2]. The derivative and integrative terms are defined in Appendix A.

The discretization of the above equations yields to the following set of linear equations at a discrete instant n:

$$\left\{ \begin{array}{lcl} F_1(n) & = & Z_1(n)U_1(n) + b_1(n)U_2(n) \\ F_i(n) & = & b_{i-1}(n)U_{i-1}(n) + Z_i(n)U_i(n) + b_i(n)U_{i+1}(n) \quad \text{for } 2 \leq i \leq N \\ F_{N+1}(n) & = & b_N(n)U_N(n) + Z_{N+1}(n)U_{N+1}(n) \end{array} \right. , \tag{7}$$

where, N is the number of tubelets modeling the considered tract,

$$
\begin{aligned}
F_i &= -b_{i-1} S_{(\Upsilon,U_d)_{i-1}} + b_i S_{(\Upsilon,U_d)_i} + P_{n_i} - \Phi_i, && \text{(8a)}\\
Z_i &= -b_{i-1} - b_i - \frac{2}{T}(L_{i-1} + L_i) - R_{i-1} - R_i - R_{n_{i-1}}, && \text{(8b)}\\
b_i &= \left[2C_i/T + G_{w_i}\right]^{-1}, && \text{(8c)}\\
G_{w_i} &= \left[2L_{w_i}/T + R_{w_i} + C_{w_i}(T/2\right]^{-1}, && \text{(8d)}\\
S_{(\Upsilon,U_d)_i} &= \Upsilon_i + U_{d_i}, && \text{(8e)}
\end{aligned}
$$

and where T is the sampling time period, Φ_i and Υ_i are terms included for time derivation and time integration respectively. Their expressions are detailed in Appendix A. One may refer to [2] for detailed steps leading to Eq. (7). For the sake of clarity, the term (n), denoting the temporal step, is not displayed for the rest of the paper. Similarly, $\Upsilon(n-1)$ and $\Phi(n-1)$ are replaced by Υ and Φ.

Finally, the pressure P_i in a tubelet is

$$P_i = b_i \left[U_i - U_{i+1} + S_{(\Upsilon,U_d)_i}\right]. \tag{9}$$

Boundary conditions should be defined for $i = 1$ and $i = N + 1$.

Boundary conditions at the glottis

The original single-matrix formulation of the vocal tract [3] follows the paper from Maeda [2] and defines the boundary conditions as follows

$$P_{sub} - P_1 = \frac{2}{T}\left[L_g + L_1\right] U_1 + \left[R_v + R_1\right] U_1 - \Phi_1, \tag{10}$$

where, according to Eq. (9),

$$P_1 = b_1 \left[U_1 - U_2 + S_{(\Upsilon,U_d)_1}\right], \tag{11}$$

and where U_1 is the volume velocity of the air flow passing through the glottis.

The substitution of Eq. (11) into Eq. (10) yields

$$F_1 = Z_1 U_1 + b_1 U_2, \tag{12a}$$

where

$$\begin{aligned} Z_1 &= -b_1 - \frac{2}{T} L_1 - R_1, & \text{(12b)} \\ F_1 &= b_1 S_{(\Upsilon, U_d)_1} - \Phi_1. & \text{(12c)} \end{aligned}$$

Boundary condition at the lips

The other boundary condition describes the radiation at the lips. The network is then terminated by a radiation impedance, which has been chosen similar to the one described by Flanagan [18]. It consists of a parallel circuit approximation, where a conductance G_{Rad} and a susceptance S_{Rad} are branched in parallel. The boundary condition at an open termination, namely the lips or the nostrils, writes

$$u(x_N, t) = \int_0^t S_{Rad}(t) p(x_N, t) dt + G_{Rad}(t) p(x_N, t), \tag{13a}$$

where

$$\begin{aligned} S_{Rad} &= \frac{3\pi\sqrt{\pi a(x_N, t)}}{8\rho}, & \text{(13b)} \\ G_{Rad} &= \frac{9\pi a(x_N, t)}{128\rho c}. & \text{(13c)} \end{aligned}$$

Following the discrete representation by Maeda [2], the discretization of Eqs. (13) and their integration into the system of equations defined by Eqs. (7), yields

$$F_{N+1} = b_N U_N + Z_{N+1} U_{N+1}, \tag{14a}$$

where

$$Z_{N+1} = -b_N - b_{N+1} - \frac{2}{T}L_N - R_N, \tag{14b}$$

$$b_{N+1} = [S_{Rad} + G_{Rad}]^{-1}. \tag{14c}$$

The acoustic propagation is computed by estimating the values of the volume velocity U_i inside each tubelet $i = 1, \ldots, N+1$. The previous system of equations form 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

$$\mathbf{f} = \mathbf{Z}\mathbf{u}, \tag{15}$$

where $\mathbf{f} \in \mathbb{R}^{(N+1)} = [F_1, \ldots, F_{N+1}]^T$, $\mathbf{Z} \in \mathbb{R}^{(N+1)\times(N+1)}$ is a tridiagonal matrix containing impedance and loss terms associated to each tubelet, and $\mathbf{u} \in \mathbb{R}^{N+1} = [U_1, \ldots, U_{N+1}]^T$ is the vector containing the volume velocities inside each tubelet.

2.3. Single-matrix formulation of the vocal tract

The single-matrix formulation, as defined by Mokhtari *et al.* [3], is a reformulation of the transmission line circuit analog model of the vocal tract seen as a waveguide network into a single matrix. Each waveguide represents a side cavity, modeled by a parallel side branch in the analog lumped circuit.

In a waveguide network, the main oropharyngeal tract, 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 tract. In the particular case where several waveguides are connected to the same parent at the same location, these waveguides are called *twins*. This is the case for the piriform fossae, for instance.

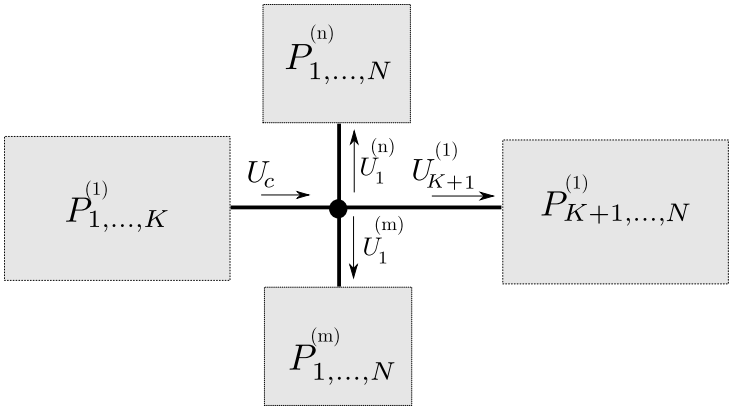


Figure 2: Junctions between twin waveguides and their parent.

Fig. 2 shows the waveguide configuration in the case of 2 twin waveguides, denoted by the symbols m and n as exponent, connected to the root node. Let

$$U_C = U^{(1)}_{K+1} + U^{(m)}_1 + U^{(n)}_1$$

be the volume velocity upstream the connection, located at the K^{th} tubelet of the parent. Then, introducing U_C in Eq. (9) at the junction K yields

$$P^{(1)}_K = b^{(1)}_K \left[U^{(1)}_K - U^{(1)}_{K+1} - U^{(m)}_1 - U^{(n)}_1 + S^{(1)}_{(\Upsilon,U_d)_K} \right]. \tag{16}$$

Applying Eq. (3) at the junction between the upstream part and each of the 3 downstream parts yields

$$
\begin{aligned}
P_K^{(1)} - P_{K+1}^{(1)} &= \frac{\partial}{\partial t}\left[\left(L_K^{(1)} + L_{K+1}^{(1)}\right) U_{K+1}^{(1)}\right] + P_{n_K}^{(1)} + \left[R_K^{(1)} + R_{K+1}^{(1)} + R_{n_K}^{(1)}\right] U_{K+1}^{(1)} \\
&\quad + \frac{\partial}{\partial t}\left[L_K^{(1)}\left(U_1^{(m)} + U_1^{(n)}\right)\right] + \left(R_{n_K}^{(1)} + R_K^{(1)}\right)\left(U_1^{(m)} + U_1^{(n)}\right), \\
P_K^{(1)} - P_1^{(m)} &= \frac{\partial}{\partial t}\left[\left(L_K^{(1)} + L_1^{(m)}\right) U_{K+1}^{(1)} + L_1^{(m)} U_1^{(m)}\right] + P_{n_K}^{(1)} + \left[R_K^{(1)} + R_1^{(m)} + R_{n_K}^{(1)}\right] U_1^{(m)} \\
&\quad + \left[R_{n_K}^{(1)} + R_K^{(1)}\right] U_{K+1}^{(1)}, \\
P_K^{(1)} - P_1^{(n)} &= \frac{\partial}{\partial t}\left[\left(L_K^{(1)} + L_1^{(n)}\right) U_{K+1}^{(1)} + L_1^{(n)} U_1^{(n)}\right] + P_{n_K}^{(1)} + \left[R_K^{(1)} + R_1^{(n)} + R_{n_K}^{(1)}\right] U_1^{(n)} \\
&\quad + \left[R_{n_K}^{(1)} + R_K^{(1)}\right] U_{K+1}^{(1)}.
\end{aligned}
\tag{17}
$$

After discretization, introduction of Eq. (16) into Eq. (17), and reorganization, so that constant terms are at the left side, the junction conditions write

$$
\begin{aligned}
F_K^{(1)} &= b_{K-1}^{(1)} U_{K-1}^{(1)} + Z_K^{(1)} U_K^{(1)} + b_K^{(1)} U_{K+1}^{(1)} + b_K^{(1)} U_1^{(m)} + b_K^{(1)} U_1^{(n)}, \\
F_{K+1}^{(1)} &= b_K^{(1)} U_K^{(1)} + Z_{K+1}^{(1)} U_{K+1}^{(1)} + b_{K+1}^{(1)} U_{K+2}^{(1)} + Z_C^{(1,m)} U_1^{(m)} + Z_C^{(1,n)} U_1^{(n)}, \\
F_1^{(m)} &= Z_1^{(m)} U_1^{(m)} + b_1^{(m)} U_2^{(m)} + b_K^{(1)} U_K^{(1)} + Z_C^{(1,m)} U_{K+1}^{(1)} + Z_C^{(1,m)} U_1^{(n)}, \\
F_1^{(n)} &= Z_1^{(n)} U_1^{(n)} + b_1^{(n)} U_2^{(n)} + b_K^{(1)} U_K^{(1)} + Z_C^{(1,n)} U_{K+1}^{(1)} + Z_C^{(1,n)} U_1^{(m)},
\end{aligned}
\tag{18}
$$

where

$$
\begin{aligned}
F_{K+1}^{(1)} &= -b_K^{(1)} S_{(\Upsilon,U_d)_K}^{(1)} + b_{K+1}^{(1)} S_{(\Upsilon,U_d)_{K+1}}^{(1)} + P_{n_{K+1}}^{(1)} - \Phi_{K+1}^{(1)}, \\
F_1^{(m)} &= b_1^{(m)} S_{(\Upsilon,U_d)_1}^{(m)} - b_K^{(1)} S_{(\Upsilon,U_d)_K}^{(1)} + P_{n_1}^{(m)} - \Phi_1^{(m)}, \\
F_1^{(n)} &= b_1^{(n)} S_{(\Upsilon,U_d)_1}^{(n)} - b_K^{(1)} S_{(\Upsilon,U_d)_K}^{(1)} + P_{n_1}^{(n)} - \Phi_1^{(n)}, \\
Z_C^{(1,m)} &= -\left[b_K^{(1)} + R_K^{(1)} + R_{n_K}^{(1)} + \frac{2}{T} L_K^{(1)}\right] = Z_C^{(1,n)}.
\end{aligned}
\tag{19}
$$

Note that, since $Z_C^{(1,m)}$ depends only on terms derived from the oropharyngeal tract, $Z_C^{(1,m)} = Z_C^{(1,n)}$.

As shown by Mokhtari *et al.* [3], the simultaneous resolution of the equations driving the acoustic propagation along the different waveguides can be performed via the concatenation of all systems, hence

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(m)} \\ \mathbf{f}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(m)T} & \mathbf{C}_1^{(n)T} \\ \mathbf{C}_1^{(m)} & \mathbf{Z}^{(m)} & \mathbf{C}_m^{(n)T} \\ \mathbf{C}_1^{(n)} & \mathbf{C}_m^{(n)} & \mathbf{Z}^{(n)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(m)} \\ \mathbf{u}^{(n)} \end{bmatrix}, \tag{20}$$

where

$$\begin{aligned} \mathbf{u}^{(m)} &= \left[U_1^{(m)}, U_2^{(m)}, \ldots, U_{N+1}^{(m)}\right]^T, \\ \mathbf{Z}^{(m)} &= \begin{bmatrix} Z_1^{(m)} & b_1^{(m)} & 0 & \\ b_1^{(m)} & Z_2^{(m)} & b_2^{(m)} & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & b_N^{(m)} & Z_{N+1}^{(m)} \end{bmatrix}. \end{aligned}$$

The matrix $\mathbf{C}_1^{(m)}$ and $\mathbf{C}_1^{(n)}$ are coupling matrices accounting for the junction between the parent, (1) in this case, and its child, (m) or (n). They are sparse matrices: after Eq. (18), the non-zero elements are $c_{1,K} = b_K^{(1)}$ and $c_{1,K+1} = Z_C^{(1,m)}$. The coupling matrix $\mathbf{C}_m^{(n)}$ is a sparse matrix whose the sole non-zero element is $c_{1,1} = Z_C^{(1,m)}$. The symbol T as an exponent denotes the transpose matrix, it should not be confused with the sampling time period.

In the case of the connection of a single side branch, e.g. the nasal tract, the system in Eq. (20) is modified by deleting the last raw of submatrices and the last column of submatrices in the block matrix containing the linear coefficients.

2.4. 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 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 [6, 10], but other models could be considered, like one mass model [19], three mass model [20], or even models dealing with more masses [21]. 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 [5]. The model presented in this section considers recent improvements : contours are smooth, allowing a mobile separation point [6, 10], and it adds corrective terms to take into account the viscous losses and the unsteady flow effects [11, 22].

The geometry of the glottal constriction used in this paper is illustrated in Fig. 3.

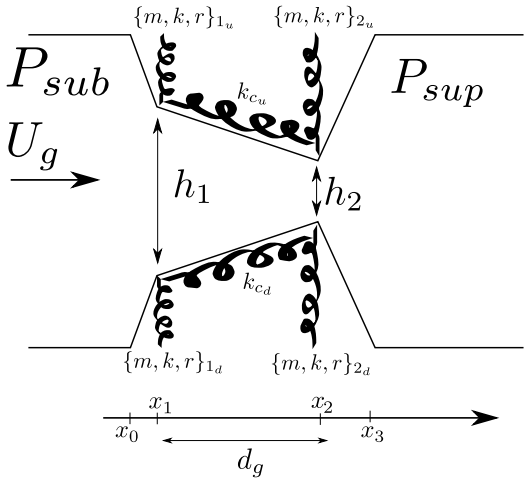


Figure 3: Geometry of the glottal constriction, as introduced by Lous *et al.* [10]. Indexes u and d stand for *upper* and *down* respectively.

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

$$\begin{aligned} P(x,t) &= P_{sub} + Be(x,t) + Po(x,t) + In(x,t) \qquad x < x_s \\ P(x,t) &= P_{sup} \qquad\qquad\qquad\qquad\qquad\qquad\quad x > x_s, \end{aligned} \tag{21}$$

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

$$\begin{aligned} Be(x,t) &= -\frac{\rho U_g^2(t)}{2l_g^2}\left[\frac{1}{h^2(x,t)} - \frac{1}{h^2(x_0,t)}\right], \\ Po(x,t) &= -\frac{12\mu U_g(t)}{l_g}\int_{x_0}^{x}\frac{dx}{h^3(x,t)}, \\ In(x,t) &= -\frac{\rho}{l_g}\frac{\partial}{\partial t}\left[U_g(t)\int_{x_0}^{x}\frac{dx}{h(x,t)}\right], \end{aligned} \tag{22}$$

where $h(x)$ is the glottal opening along the x coordinates, 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 with the glottal constriction geometry:

- if $1.2h_1 > h_2$, $x_s = x_2$
- if $1.2h_1 < h_2$, x_s is such that $h_s = h(x_s) = 1.2h_1$

The value 1.2 is an *ad-hoc* criterion, as used in [6, 10]. 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. (23)

$$P_{sub}(t) - P_{sup}(t) = R_b(t)U_g^2(t) + R_v(t)U_g(t) + \frac{\partial}{\partial t}\left[L_g(t)U_g(t)\right], \tag{23}$$

where

$$\begin{aligned} R_b(t) &= \frac{\rho}{2l_g^2}\left[\frac{1}{h^2(x_s,t)} - \frac{1}{h^2(x_0,t)}\right], \\ R_v(t) &= \frac{12\mu}{l_g}\int_{x_0}^{x_s}\frac{dx}{h^3(x,t)}, \\ L_g(t) &= \frac{\rho}{l_g}\left[\int_{x_0}^{x_s}\frac{dx}{h(x,t)}\right], \end{aligned} \tag{24}$$

From the determination of the glottal flow U_g, Eq. (21) 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

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{R}\dot{\mathbf{y}} + \mathbf{K}\mathbf{y} = \mathbf{F}, \tag{25}$$

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, $\mathbf{K}$ writes

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

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 it could be a phonatory articulator that should be considered for time-domain continuous speech synthesis, especially in order to simulate natural prosody, it is not 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.* [23].

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 some lateral approximants [24–26]. The acoustic effect is not fully apprehended, mainly because of the lack of acoustic models, and also because of the lack of relevant articulatory data. Indeed, lateral consonants are usually modeled by a

single lateral channel, and additionally a supralingual cavity [27, 28]. Zhang *et al.* [25, 26] studied the effect of bilateralization with a frequency based method to compute the transfer function of bilateral vocal tracts. However, at the best of our knowledge, there is still no existing model of bilateral vocal tract to be used in the context of time-domain continuous speech synthesis. The original single-matrix formulation of the vocal tract [3] does not discuss the case, but this section shows that it can be integrated to it.

Fig. 4 shows the waveguide connections in the case of bilateralization. It includes a secondary waveguide that is connected to the oral branch at two points.

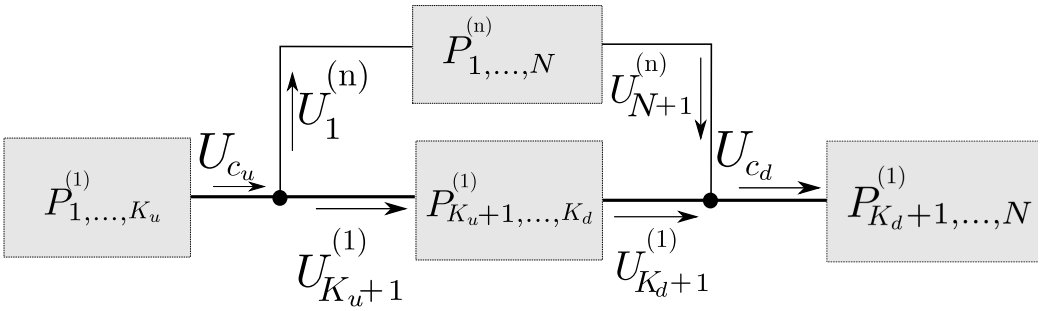


Figure 4: Equivalent diagram of anastomozing waveguides.

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. 4 is waveguide (1), while (n) is its anabranch.

Equations describing the conditions at the upstream connection K_u are similar to a general parent-child relationship, described in Eqs. (16) to (19).

In that case, since there is no waveguide (m), $U_1^{(m)} = 0$ and equations referring to (m) at the left side should not be considered. Hence

$$\begin{aligned} F_{K_u}^{(1)} &= b_{K_u-1}^{(1)} U_{K_u-1}^{(1)} + Z_{K_u}^{(1)} U_{K_u}^{(1)} + b_{K_u}^{(1)} U_{K_u+1}^{(1)} + b_{K_u}^{(1)} U_1^{(n)}, \\ F_{K_u+1}^{(1)} &= b_{K_u}^{(1)} U_{K_u}^{(1)} + Z_{K_u+1}^{(1)} U_{K_u+1}^{(1)} + b_{K_u+1}^{(1)} U_{K_u+2}^{(1)} + Z_C^{(1,n)} U_1^{(n)}, \\ F_1^{(n)} &= Z_1^{(n)} U_1^{(n)} + b_1^{(n)} U_2^{(n)} + b_{K_u}^{(1)} U_{K_u}^{(1)} + Z_C^{(1,n)} U_{K_u+1}^{(1)}. \end{aligned} \tag{26}$$

At the downstream connection K_d, the merged volume velocity $U_{c_d} = U_{N+1}^{(n)} + U_{K_d+1}^{(1)}$, is introduced into Eq. (9)

$$P_{K_d+1}^{(1)} = b_{K_d+1}^{(1)} \left[U_{K_d+1}^{(1)} + U_{N+1}^{(n)} - U_{K_d+2}^{(1)} + S_{(\Upsilon,U_d)_{K_d+1}}^{(1)} \right]. \tag{27}$$

Applying Eq. (3) at the downstream junction yields

$$\begin{aligned} P_{K_d}^{(1)} - P_{K_d+1}^{(1)} &= \frac{\partial}{\partial t} \left[\left(L_{K_d}^{(1)} + L_{K_d+1}^{(1)} \right) U_{K_d+1}^{(1)} + L_{K_d+1}^{(1)} U_{N+1}^{(n)} \right] + P_{n_{K_d}}^{(1)} + \left[R_{K_d}^{(1)} + R_{K_d+1}^{(1)} + R_{n_{K_d}}^{(1)} \right] U_{K_d+1}^{(1)} \\ &\quad + R_{K_d+1}^{(1)} U_{N+1}^{(n)}, \\ P_N^{(n)} - P_{K_d+1}^{(1)} &= \frac{\partial}{\partial t} \left[\left(L_N^{(n)} + L_{K_d+1}^{(1)} \right) U_{N+1}^{(n)} + L_{K_d+1}^{(1)} U_{K_d+1}^{(1)} \right] + P_{n_N}^{(n)} + \left[R_N^{(n)} + R_{K_d+1}^{(1)} + R_{n_N}^{(n)} \right] U_{N+1}^{(n)} \\ &\quad + R_{K_d+1}^{(1)} U_{K_d+1}^{(1)}. \end{aligned} \tag{28}$$

The substitution of Eq. (27) into Eqs. (28) yields

$$\begin{aligned} F_{K_d+1}^{(1)} &= b_{K_d}^{(1)} U_{K_d}^{(1)} + Z_{K_d+1}^{(1)} U_{K_d+1}^{(1)} + b_{K_d+1}^{(1)} U_{K_d+2}^{(1)} + Z_{C+1}^{(1,n)} U_{N+1}^{(n)}, \\ F_{K_d+2}^{(1)} &= b_{K_d+1}^{(1)} U_{K_d+1}^{(1)} + Z_{K_d+2}^{(1)} U_{K_d+2}^{(1)} + b_{K_d+2}^{(1)} U_{K_d+3}^{(1)} + b_{K_d+1}^{(1)} U_{N+1}^{(n)}, \\ F_{N+1}^{(n)} &= b_N^{(n)} U_N(n) + Z_{N+1}^{(n)} U_{N+1}^{(n)} + b_{K_d+1}^{(1)} U_{K_d+2}^{(1)} + Z_{C+1}^{(1,n)} U_{K_d+1}^{(1)}, \end{aligned} \tag{29}$$

where

$$
\begin{aligned}
F_{N+1}^{(n)} &= -b_N^{(n)} S_{(\Upsilon,U_d)_N}^{(n)} + b_{K+1}^{(1)} S_{(\Upsilon,U_d)_{K+1}}^{(1)} + P_{n_N}^{(n)} - \Phi_{N+1}^{(n)}, \\
Z_{C+1}^{(1,n)} &= -\left[b_{K+1}^{(1)} + R_{K+1}^{(1)} + \frac{2}{T} L_{K+1}^{(1)} \right], \\
Z_{N+1}^{(n)} &= -b_N^{(n)} - b_{K+1}^{(1)} - \frac{2}{T}\left(L_N^{(n)} + L_{K+1}^{(1)} \right) - R_N^{(n)} - R_{K+1}^{(1)} - R_{n_N}^{(n)}.
\end{aligned}
\tag{30}
$$

Note that in that case, since the end of the waveguide (n) is directly connected to the main oral tract, there is no radiation at the end, hence the particular formula in Eq. (30) for computing $Z_{N+1}^{(n)}$.

The matrix formulation is then

$$
\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(n)T} \\ \mathbf{C}_1^{(n)} & \mathbf{Z}^{(n)} \end{bmatrix} . \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(n)} \end{bmatrix}, \tag{31}
$$

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 utterance with good quality, but it cannot be used to study the acoustic coupling between the vocal folds and the vocal tract. In [23], the authors connect the original two-mass model by Ishazaka and Flanagan [5] to the single-matrix formulation, in order to simulate the self-oscillating motion of the vocal folds.

However, the quadratic term in Eq. (23) is not taken into account in [23]. 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. (23) should be integrated to the single-matrix formulation. Once U_g is known, pressure forces are derived from Eq. (21), and the mass positions are computed following Eq. (25).

Introducing Eq. (23) into the single-matrix formulation of Eq. (41) requires the first line of the system to be modified into a quadratic equation

$$F_1 = Z_1 U_1 + b_1 U_2 + R_b U_1^2. \tag{32}$$

The matrix form is then

$$\mathbf{f} = \mathbf{Z}\mathbf{u}_Z + \mathbf{Q}\mathbf{u}_Q, \tag{33}$$

where $\mathbf{Q}$ is a square matrix the same size as $\mathbf{Z}$ having only one non-zero element, that is $Q_{(1,1)} = R_b$, 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. (33) 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. (33) by $\mathbf{Z}^{-1}$ yields

$$\mathbf{Z}^{-1}\mathbf{f} = \mathbf{I}\mathbf{u}_Z + \mathbf{Z}^{-1}\mathbf{Q}\mathbf{u}_Q, \tag{34}$$

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. (33) can be rearranged

$$\tilde{\mathbf{f}} = \tilde{\mathbf{Z}}\tilde{\mathbf{u}}_Z, \tag{35}$$

where $\tilde{\mathbf{f}} \in \mathbb{R}^N = [f_2 - b_1U_1, f_3, \ldots, f_n, \ldots, f_{N+1}]^T$, $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times N}$ is the matrix $\mathbf{Z}$ to 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. (35) 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 vocal folds presented in the previous section considers the latter 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 occurs in the glottal flow waveform. The latter is thus never null.

Parametric models of the glottal chink have been proposed by Cranen and Schroeter [29, 30]. 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 (see Fig. 5 a), 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. More recently, Wilhelms-Tricarico [31] proposed a modification of the classic two-mass model by Ishizaka and Flanagan [5] 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.

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. 5. Unlike in [31] U_g and U_{ch} separate upstream the glottal contraction, and mix downstream the glottal expansion. These assumptions follow the model detailed in [30]. Our formulation assumes that the glottis partial closure is due to a partial abduction of the vocal folds and the chink is linked to the mobile part of the glottis, as shown in Fig. 5. Since the vocal fold oscillation is assumed to be small in relation to the partial abduction, the parallel branch assumption can be considered as valid, *i.e.* the geometry of the glottal chink should not be disturbed by the vocal folds oscillations.

The glottal chink area is $a_{ch} = l_{ch} h_{ab}$, where l_{ch} is the length of the glottal

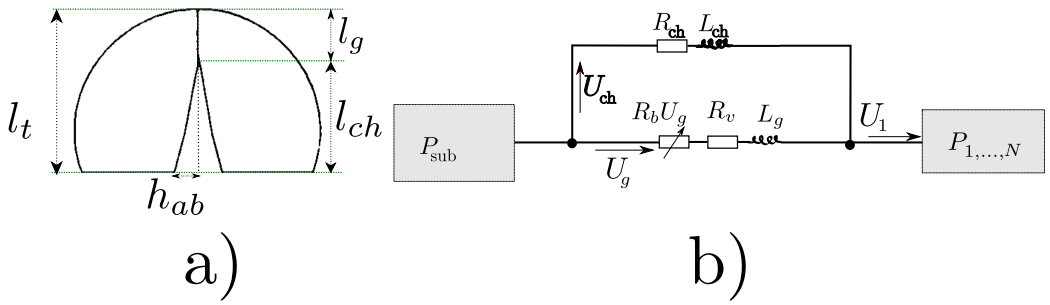


Figure 5: a) View of the partially closed glottis, extracted from Cranen and Schroeter [29]. In this model, the partial closure is due to a partial abduction of the vocal folds. l_g is the length of the vibrating part of the vocal folds, l_{ch} is the length of the glottal chink and $l_t = l_g + l_{ch}$ 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.

chink and h_{ab} is the abduction of the vocal folds. Note that in the original model in [30], the glottal chink may be extended with a constant opening area, due to the inter-arytenoid portion of the glottis. It is not taken into account in the presented model, but its 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 associated to U_g. Indeed, as seen in Fig. 5 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. (34). Indeed, applying Eq. (10) to both glottal branches yields

$$\begin{aligned} P_1 - P_{sub} &= R_b U_g^2 + [R_v + R_1]\, U_g + \frac{\partial}{\partial t}\,[L_g + L_1]\, U_g \\ &\quad + R_1 U_{ch} + \frac{\partial}{\partial t} L_1 U_{ch} \qquad (36) \\ P_1 - P_{sub} &= [R_{ch} + R_1]\, U_{ch} + \frac{\partial}{\partial t}\,[L_{ch} + L_1]\, U_{ch} \\ &\quad + R_1 U_g + \frac{\partial}{\partial t} L_1 U_g \qquad (37) \\ P_1 &= b_1 \left[U_g + U_{ch} - U_2 + S_{(\Upsilon, U_d)_1}\right]. \qquad (38) \end{aligned}$$

This yields to the following boundary conditions at the glottis

$$\begin{aligned} F_1 &= R_b U_g^2 + Z_1 U_g + b_1 U_2 + Z_C^{(1,ch)} U_{ch}, \\ F_2 &= b_1 U_g + Z_2 U_2 + b_2 U_3 + b_1 U_{ch}, \qquad (39) \\ F_{ch} &= Z_{ch} U_{ch} + Z_C^{(1,ch)} U_g + b_1 U_2, \end{aligned}$$

where

$$Z_{(1,ch)} = b_1 + R_1 + \frac{2}{T} L_1,$$

and

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

Introducing Eqs. (39) in the single matrix formulation yields

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ F_{ch} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(ch)T} \\ \mathbf{C}_1^{(ch)} & Z_{ch} \end{bmatrix} . \begin{bmatrix} \mathbf{u}^{(1)} \\ U_{ch} \end{bmatrix} + R_b U_g^2, \qquad (40)$$

where $\mathbf{C}_1^{(ch)}$ is the matrix accounting for the coupling the glottal chink and the vocal tract. It is a sparse raw 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 of the vocal tract writes

$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \vdots \\ \mathbf{f}^{(\mathcal{N})} \\ F_{ch} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}_1^{(2)T} & \dots & \mathbf{C}_1^{(\mathcal{N})T} & \mathbf{C}_1^{(ch)T} \\ \mathbf{C}_1^{(2)} & \mathbf{Z}^{(2)} & & \mathbf{C}_2^{(\mathcal{N})T} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{C}_1^{(\mathcal{N})} & \mathbf{C}_2^{(\mathcal{N})} & & \mathbf{Z}^{(\mathcal{N})} & \mathbf{0} \\ \mathbf{C}_1^{(ch)} & \mathbf{0} & \dots & \mathbf{0} & Z_{ch} \end{bmatrix} . \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \vdots \\ \mathbf{u}^{(\mathcal{N})} \\ U_{ch} \end{bmatrix} + \mathbf{Q}\mathbf{u}_Q \quad (41)$$

where $\mathcal{N}$ is the number of waveguides modeling the whole vocal tract, and where $\mathbf{C}_m^{(n)}$ is a matrix accounting for the relationship between the m^{th} and the n^{th} waveguide. There are several possibilities:

1. the branches are not directly connected, hence $\mathbf{C}_m^{(n)} = \mathbf{0}$,
2. (n) is a child of (m), connected at the K^{th} tubelet of (m). Then $\mathbf{C}_m^{(n)}$ is a sparse matrix whose non-zero elements are $c_{1,K} = b_K^{(m)}$ and $c_{1,K+1} = Z_C^{(m,n)}$,
3. (m) and (n) are twins, both connected to waveguide (p). Then $\mathbf{C}_m^{(n)}$ is a sparse matrix whose the sole non-zero element is $c_{1,1} = Z_C^{(p,n)}$,
4. (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).
5. (n) is the glottal chink. In that case, it is connected to (1) and $\mathbf{C}_1^{(ch)} = [Z_{(1,ch)}, b_1, 0, \dots, 0]$.

Finally, the steps to use the extended singe-matrix formulation for speech synthesis are:

I Initialize the parameters of the vocal tract and the vocal folds

- area functions of each waveguide
- coupling relationships between the waveguides
- initial values of the vocal folds model and dimension of the glottal chink

II Update the system

- compute $\mathbf{Z}$ and $\mathbf{f}$ for each waveguide using Eqs. (7), (8), (12), and (14). Compute R_b from Eq. (24)
- compute $\mathbf{C}$ for each relationship

III Solve the system at each temporal step

- find U_g by solving the first line of Eq. (34). If the vocal folds collide, $U_g = 0$
- solve the remaining system from Eq. (35)
- compute P_{Out} as the time first difference of the sum of the volume velocities radiated by the waveguides with an open termination
- compute pressure forces via Eq. (21) and motion of the vocal folds by Eq. (25)

IV Repeat steps 2 and 3 until the end of the simulation

4. Numerical simulations

In order to validate the simulation framework presented in the previous sections, this section provides examples of synthesized speech in static configurations. Area functions are derived from X-ray films comprising short French sentences [32]. They are obtained by dividing the vocal tract shape in tubelets perpendicular to the vocal tract centerline, determined via a specified algorithm [33], and then applying α β transformations to recover the area [34]. Chosen parameters for the vocal folds model are typical values found in the literature (see Tab. 2).

In the following numerical simulations, the vocal folds are supposed symmetric, that is the upper vocal fold has the same mechanical parameters than the lower vocal fold.

Note that when a waveguide is closed at its end, *e.g.* the piriform fossae, a termination area set to 0 may make the numerical computation to break because of infinite or NaN values. Setting the termination area to a very small value, *e.g.* the order of magnitude of the unit roundoff, is sufficient to efficiently approximate a closed termination. A similar technique can be used in the case when sudden modifications of the global structure of the network occur, *e.g.* 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.

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	$k/2$

4.1. Anastomozing waveguides

The acoustics of lateral consonants has been previously studied mainly by using single lateral channel models [27, 28]. 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.* [25, 26], using the same area functions. Fig. 6 shows the single-tube model used in the simulation in [26]. The area functions of the different parts are derived from MRI [25], and are averaged over the length of each

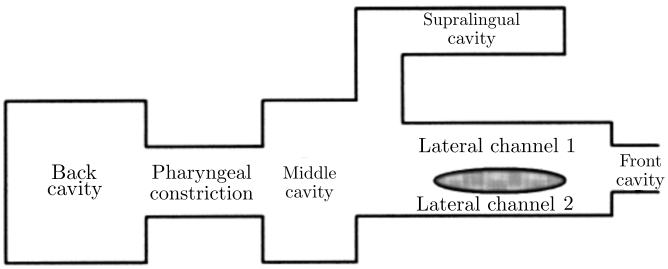


Figure 6: Single-tube model used for the simulation of bilaterals. Figure adapted from [26].

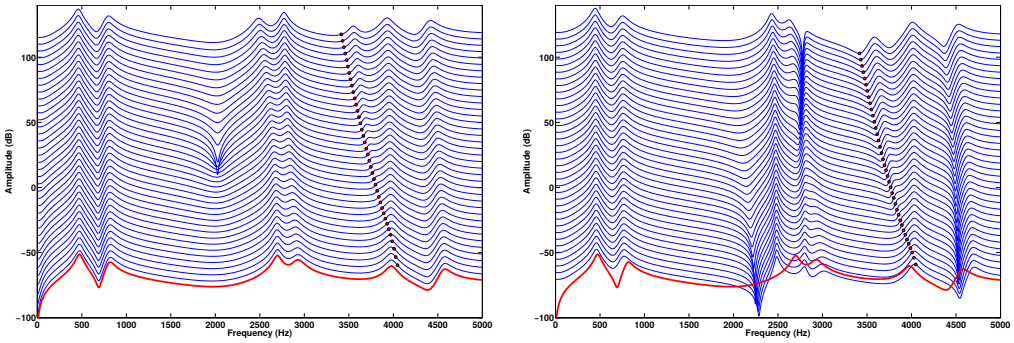
section of the single-tube model (see Fig. 6). Dimensions can be found in the original paper [25]. When the supralingual cavity is taken into account, the latter is connected to the main oral tract at the same location as the lateral channel. Consequently, the supralingual cavity is a twin waveguide 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 configurations of bilateralization. They are defined by an asymmetric factor

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

where l_{c_2} and l_{c_1} denote the length of both lateral channels. Simulations include two main configurations: with and without the consideration of the supralingual cavity.

Fig. 7 shows transfer functions for several configurations obtained with



(a) Without the supralingual cavity (b) With the supralingual cavity

Figure 7: 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. From bottom to top are plotted the transfer functions of the vocal tract with bilateralization, 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 position of zeros introduced by the bilateralization are denoted by circle marks.

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 [26] 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 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 [35], hence a drop of the frequency as the length increases. The theoretical zero frequencies are represented in Fig. 7 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 with 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 [26], 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

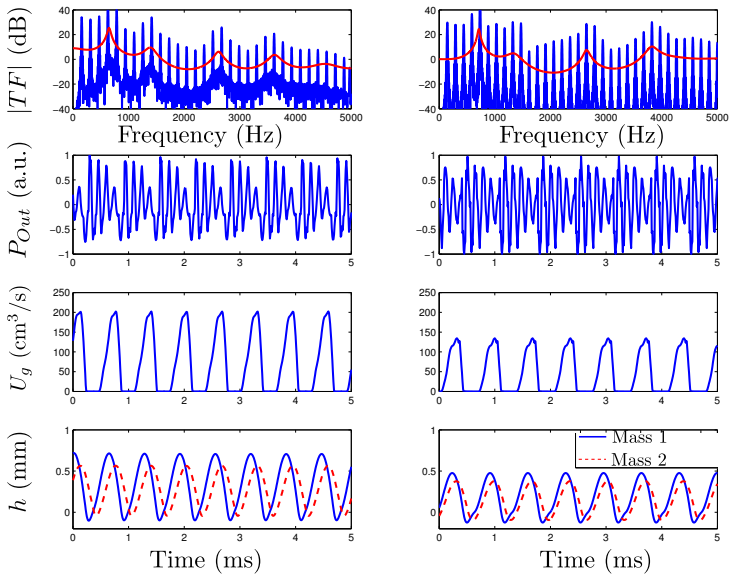
supralingual cavity.

The agreement between the global effects of bilateralization and the supralingual cavity on the vocal tract acoustic response functions obtained with the extended single-matrix and those obtained with the independent frequency based technique used in [26] 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 approximants.

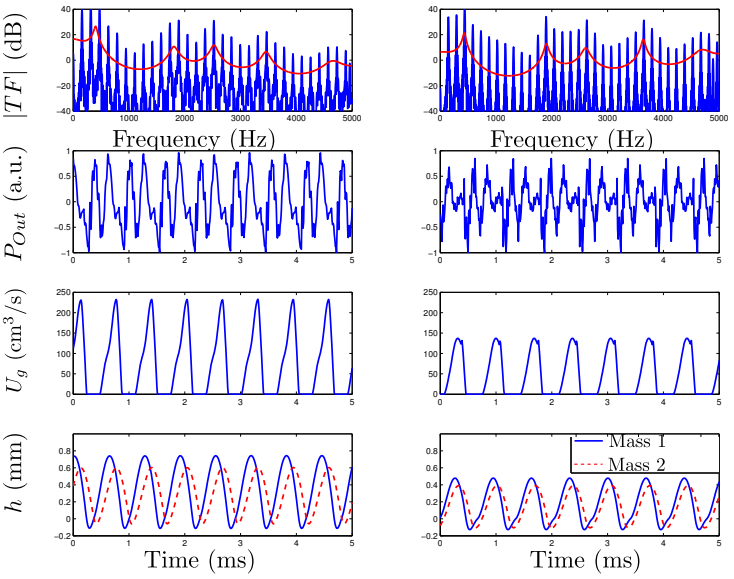
4.2. Effect of the acoustic model on the motion of the vocal folds

In this section, numerical simulations are used to study the validity of the extended single-matrix formulation of the vocal tract to be connected with a self-sustaining model of vocal folds. The method is compared with the concurrent approach using *reflection type line analog* models [1, 36]. This approach is widely used and validated when connected with self-oscillating models of vocal folds, including the 2×2-mass model with smooth contours used in this paper [10, 11, 22]. Thus, simulations consist in computing the vocal folds motion via the model described in Sec. 2.4 connected to several configurations of the vocal tract, using two models of acoustic propagation: the extended single-matrix formulation (ESMF), and the reflection type line analog model (RTLA). The configurations of the vocal tract are area functions corresponding to 6 French vowels: /a/, /e/, /ø/, /i/, /o/, and /u/.

Fig. 8 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



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



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

Figure 8: 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 the reflection type line analog (RTLA) model, 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 columns displays the same quantities computed thanks to the extended single-matrix formulation (ESMF). The spectral envelope is computed by a 10-order LPC (*Linear Predictive Coding*) [37]

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 the 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. 9). This yields to a lower amplitude of the glottal flow. Such differences are somewhat expected since the coupling with the acoustic propagation may be modified, due to the different models of acoustic losses. It is accepted that the RTLA method has the advantage to better accounting for the frequency dependance of the acoustic losses and the acoustic radiation. This may also explain the different formant bandwidth between those obtained with the RTLA method and those obtained with the ESMF method.

It is worth noting that slight modifications of the input parameters of the vocal folds suffice to recover similar behaviors of the vocal folds. It consists in multiplying the values of the stiffness by a factor, say Q, such that the obtained fundamental frequency matches that obtained with RTLA. This is highlighted by Fig. 9, 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

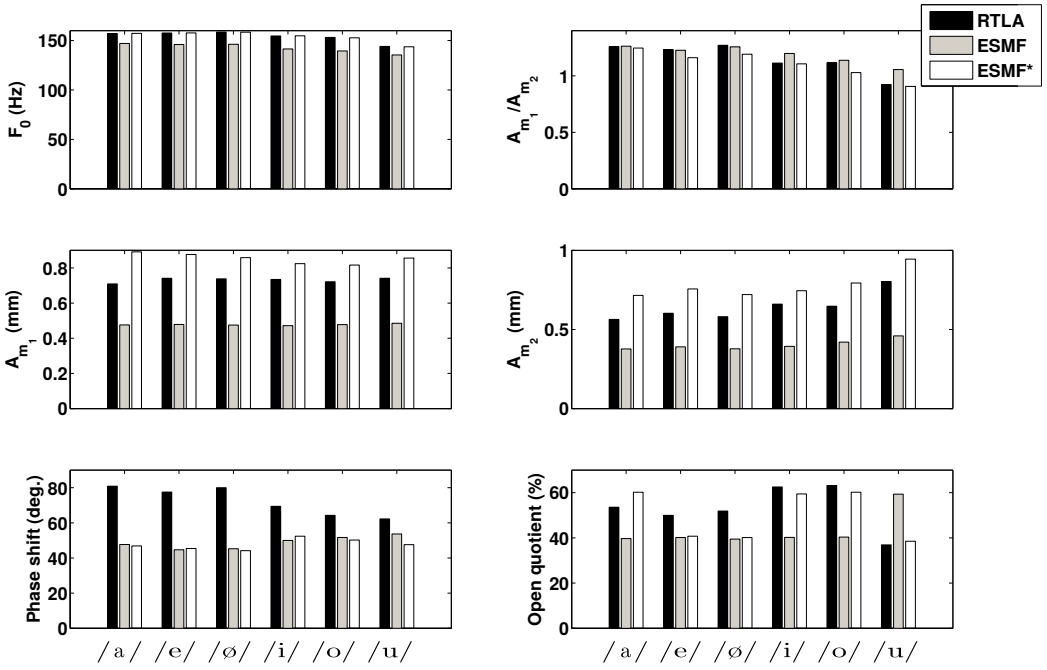


Figure 9: Comparison of quantities derived from the motion of the vocal folds obtained with the different methods. ESMF* corresponds to the quantities obtained with modified mass and stiffness of the vocal folds model.

the vocal folds model.

Fig. 9 shows that, although both methods simulate qualitatively similar oscillations of the vocal folds and glottal flow, they are quantitatively different. Main differences lie in the phase shift and in the open quotient: when connected to RTLA, the phase shift between the rear and front part 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

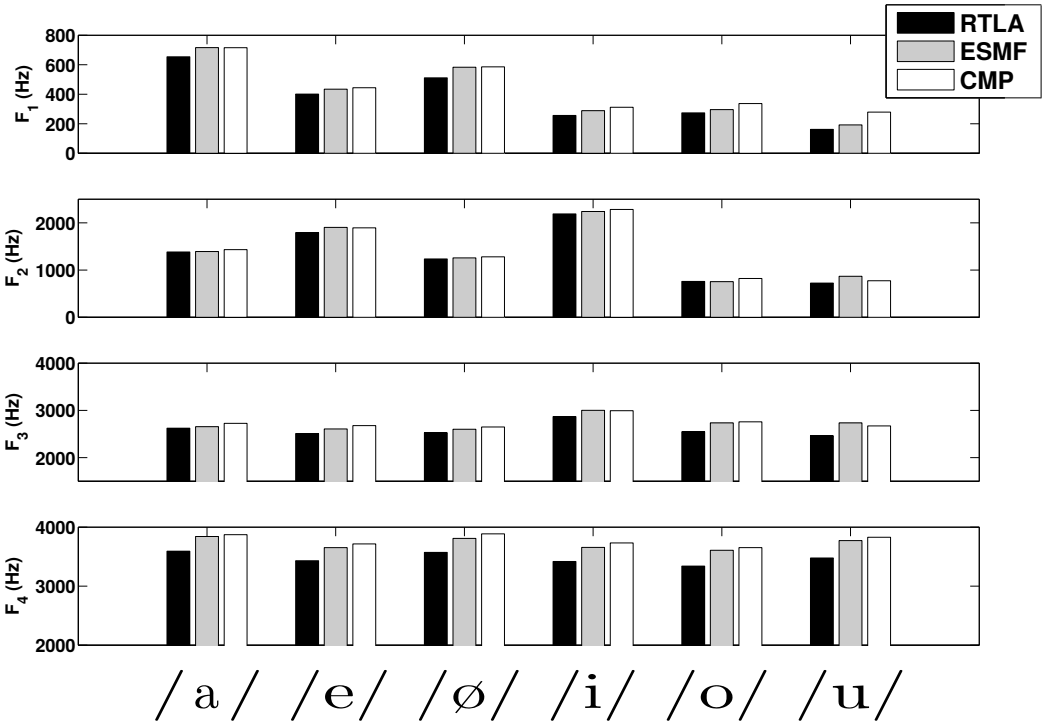


Figure 10: Formant frequency of the first 4 formants obtained with the different methods for 6 French vowels. Results are compared with values obtained with the chain-matrix paradigm (CMP) [16] for the corresponding area functions.

parameters modifies the open quotient. Also, for both techniques, the amplitude 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 supraglottal constriction seems to have a strong effect on the downstream part of the vocal folds.

Fig. 10 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) [16]. The formant frequency 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 method gives closer formant frequencies than RTLA.

To summarize, the presented ESMF has shown its accuracy to compute the acoustic propagation and to account 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 [11, 38]. 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 with the frequency domain based chain-matrix paradigm [16]. Consequently, since ESMF presents the advantage of easily deal 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 qual-

itatively 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. 5). The quantities l_t and h_{ab} are kept constant, and by deduction, $l_g = (1 - \alpha) l_t$.

Fig. 11 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, the fundamental frequency, and the length of the glottal chink l_{ch}. The effect of the glottal chink on these physical quantities are 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 refrains U_t to be 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.

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

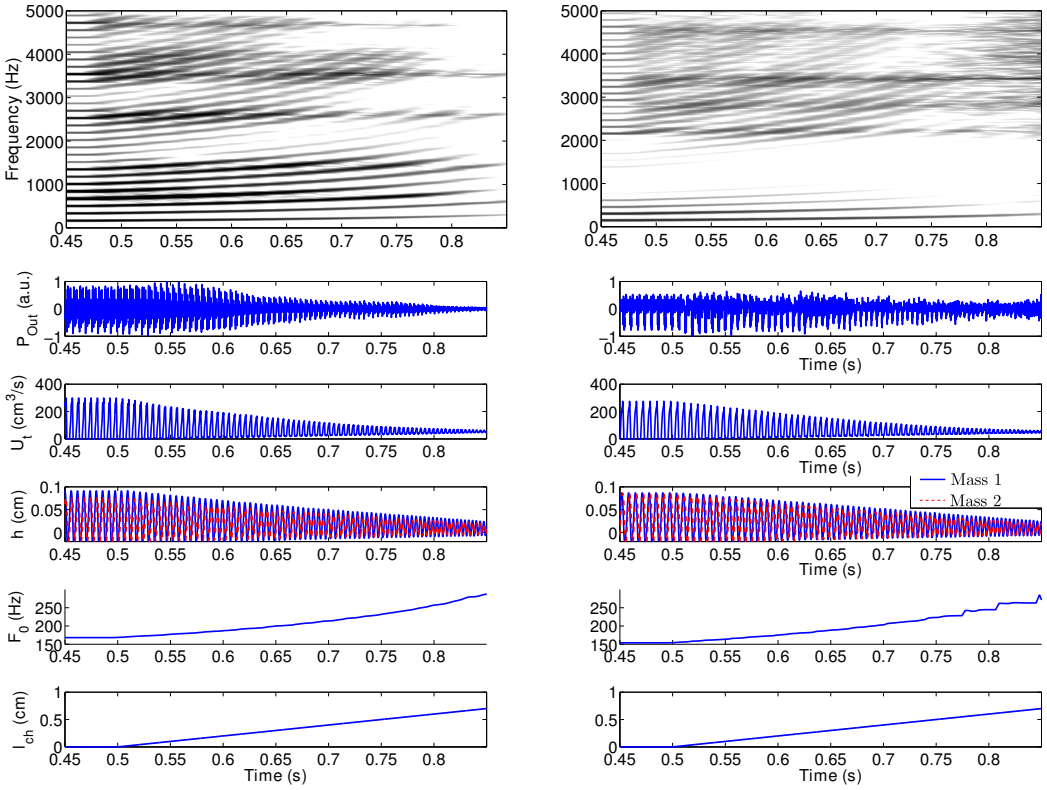


Figure 11: Results of the simulation for two French vowels /a/ (left column), and /i/ (right column). From top bottom, the left column represents the narrow-band spectrogram, the output pressure, radiated at the lips, the volume velocity inside the glottis ($U_t = U_g + U_{ch}$), the glottal opening at the location of mass 1 (solid line) and mass 2 (dashed line), the fundamental frequency, and the length l_{ch} of the glottal chink, computed for the /a/. The right column represents these quantities computed for /i/.

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 important air flow 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 pattern seems to be barely modified by the presence of the chink.

The presented simulation clearly shows the effect of the glottal chink on some acoustic parameters. This confirms the interest of such models for thoroughly investigate the acoustic or phonatory phenomena involved in speech production. Modeling the partially closed glottis is also important for 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.

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 have been studied by computing transfer functions of vocal tract models derived from a previous study [26]. The effects of the bilateralization on the vocal tract acoustic response functions are in agreement with those observed with an independent frequency based technique. Indeed, the bilateralization with asymmetric lateral channels introduces a pole/zero pair at frequencies around 4 kHz. This frequency drops as the total length of the lateral channels increases. This agrees with theoretical predictions [26, 35] which relate the pole/zero frequencies to the first resonance frequency of a tube having a length equal to the sum of the lengths of both lateral channels. The consideration of the supralingual cavity connected to the lateral channels at the same point to the main oral tract introduces a zero around 2.7 kHz, which corresponds to the resonance of the first quarter-wavelength of the supralingual cavity.

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. For instance, the amplitude of oscillation is smaller when the vocal folds are connected to the single-matrix formulation. These modifications are potentially correctable by modifying the mass and stiffness of the two-mass model of the vocal folds with a linear factor. Since these values are arbitrary and do not directly correspond to physiological data, this is sufficient for continuous speech synthesis. Besides, the reflection type line analog model is known to accurately predict the behavior of the vocal folds qualitatively but fails to predict it quantitatively. Therefore, at this stage, it is not possible to conclude whether these differences are important for quantitative studies. The simulations also revealed that the formant frequencies of the vowels simulated by the extended single-matrix formulation are closer to the resonance frequencies of the vocal tract than those of the vowels simulated by the reflection type line analog model.

The paper also provides simulations with various lengths of the glottal chink. They clearly show its acoustic effect: the simultaneous shortening of the vibrating part of the vocal folds together with the partial closure of the glottis leads to a drop of the glottal flow amplitude and to the amplitude of the vocal folds motion, and consequently to a drop of the acoustic level radiated at lips. The partial closure of the glottis also leads to glottal leakage, *i.e.* the total glottal flow is never null. As the glottal chink becomes large, this generates a large DC component U_{DC} 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, even for vowels, as shown by the simulation of /i/. A realistic model of the glottal partial

closure is therefore important for the synthesis of voiced fricatives.

The simulation framework is also 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 utterance containing nasal phonemes [39]. 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 to include glottal leakage. This last point allows the different types of phonation to be realistically simulated.

This constitutes a useful tool to study the phenomena involved in speech production. This paper constitutes 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

The authors would like to sincerely acknowledge Dr. Shinji Maeda and Dr. Parham Mokhtari, for their useful advises and fruitful discussions.

Appendix A. Discrete representation of the vocal tract

This appendix clarifies the derivative and integrative terms that are present in the acoustic propagation equations. They are defined as in [2].

In this section, unless specified, the terms included in the following equations are expressed at instant $n-1$, except for Φ, Υ, and I, which are expressed at instant $n-2$, if they are at the right side. In Eq. (8a), the terms $\Phi_i^{(m)}$ of the m^{th} waveguide are defined by

$$\begin{cases} \Phi_1^{(1)} = \frac{4}{T}\left[L_1^{(1)} + Lg\right] U_g + \frac{4}{T} L_1^{(1)} U_{ch} - \Phi_1^{(1)}, \\ \Phi_1^{(m)} = \frac{4}{T}\left[L_1^{(m)} + L_K^{(n)}\right] U_1^{(m)} + \frac{4}{T} L_K^{(n)} \left[U_{K+1}^{(n)} + \sum_{l=1}^{\mathcal{L}} U_1^{(l)}\right] - \Phi_1^{(m)}, \quad m \neq 1 \\ \Phi_i^{(m)} = \frac{4}{T}\left[L_{i-1}^{(m)} + L_i^{(m)}\right] U_i^{(m)}(n-1) - \Phi_i^{(m)}, \quad 2 \leq i \leq M \\ \Phi_{K+1}^{(m)} = \frac{4}{T}\left[L_K^{(m)} + L_{K+1}^{(m)}\right] U_{K+1}^{(m)} + \frac{4}{T} \sum_{j=1}^{\mathcal{J}} L_K^{(m)} U_1^{(j)} - \Phi_{K+1}^{(m)}, \\ \Phi_{K+1}^{(m)} = \frac{4}{T}\left[L_K^{(m)} + L_{K+1}^{(m)}\right] U_{K+1}^{(m)} + \frac{4}{T} L_{K+1}^{(m)} U_{N+1}^{(n)} - \Phi_{K+1}^{(m)}, \text{ if } n \text{ is an anabranch of } m \\ \Phi_{N+1}^{(m)} = \frac{4}{T} L_N^{(m)} U_{N+1}^{(m)} - \Phi_{N+1}^{(m)}, \text{ if } m \text{ is not an anabranch} \\ \Phi_{N+1}^{(m)} = \frac{4}{T}\left[\left(L_N^{(m)} + L_{K+1}^{(n)}\right) U_{N+1}^{(m)} + L_{K+1}^{(n)} U_{K+1}^{(n)}\right] - \Phi_{N+1}^{(m)}, \text{ if } m \text{ is an anabranch of } n, \\ \Phi_1^{(ch)} = \frac{4}{T}\left[L_1^{(1)} + L_{ch}\right] U_{ch} + \frac{4}{T} L_1^{(1)} U_g - \Phi_1^{(ch)} \end{cases} \tag{A.1}$$

where $.^{(1)}$ denotes the main oral tract, and $.^{(n)}$ denotes the parent waveguide, namely the waveguide to which the m^{th} waveguide ($m > 1$) is connected. K is the position of the T-junction on the parent waveguide, $U_1^{(j)}$, with $j = 1, \ldots, \mathcal{J}$, are the input volume velocities of the $\mathcal{J}$ twin children of m, that are connected to (m) at the same point, and $U_1^{(l)}$, with $l = 1, \ldots, \mathcal{L}$, are the input volume velocities of the $\mathcal{L}$ twin waveguides of m.

The terms Υ_i of Eq. (8a) are defined by

$$\begin{cases} \Upsilon_i(n-1) = \Phi_{C_i}(n-1) - G_{w_i}(n-1)[\Phi_{Lw_i}(n-1) - I_{Cw_i}(n-1)], \quad i \leq M \\ \Upsilon_{M+1}(n-1) = -2S_{rad}(n-1)\sqrt{a_{M+1}(n-1)}P_{M+1}(n-1) + \Upsilon_{M+1}(n-2), \end{cases} \tag{A.2}$$

where

$$\Phi_{C_i} = \frac{4}{T}C_iU_i - \Phi_{C_i}, \tag{A.3}$$

$$\Phi_{Lw_i} = \frac{4}{T}L_{w_i}u_2 - \Phi_{Lw_i}, \tag{A.4}$$

$$I_{Cw_i} = TC_{w_i}u_2 + I_{Cw_i} \tag{A.5}$$

Eq; (A.2) to Eq. (A.5) are similar for every waveguide of the network.

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] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] H. Y. Wu, P. Badin, Y. M. Cheng, B. Guérin, Simulation du conduit vocal : réalisation de la variation continue de longueur dans un modèle

Kelly-Lochbaum . Effet de l'échantillonnage spatial de la fonction d'aire (Simulation of the vocal tract: Realization of the continuous variation of length in a Kelly-Lochbaum model. Effect of the area function spatial sampling), in: Bulletin du laboratoire de la Communication Parlée, 1987, pp. 1–27.

[13] 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.

[14] 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.

[15] 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.

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

[17] P. Birkholz, Enhanced area functions for noise source modeling in the vo-

cal tract, in: 10th International Seminar on Speech Production, Cologne, 2014, pp. 1–4.

[18] J. L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Edition, Springer-Verlag, berlin, 1972.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

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

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

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

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

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

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.