Chapter 4

Source models for articulatory speech synthesis

Early research in analysis, synthesis and coding of voice has traditionally focused on the vocal tract filter, paying less attention to the source signal. Especially in the last decade, however, more emphasis has been given to the characteristics of the glottal source waveform: the development of a good model for the glottal excitation has been recognized to be a key feature for obtaining high quality speech synthesis, and for characterizing voice quality (e.g. modal voice, vocal fry, breathy voice [6, 32]).

Parametric models fit the glottal signal with piecewise analytical functions, and typically use a small number of parameters. As an example, the Liljencrants-Fant model [44] characterizes one cycle of the flow derivative using as few as four parameters. The Liljencrants-Fant model has been successfully used for fitting flow derivatives computed by inverse filtering real utterances [32, 101, 132].

Physical models describe the glottal system in terms of physiological quantities. The Ishizaka-Flanagan (IF) model [73] is a known example of lumped model of the vocal folds. Physical models capture the basic non-linear mechanisms that initiate self-sustained oscillations in the glottal system, and can simulate features (e.g. interaction with the vocal tract) that are not taken into account by parametric models. However they typically involve a large number of control parameters. As a consequence, the models are not easily controlled, since the model parameters do not map in an intuitive way into perceptual dimensions. Moreover, physical models are less suitable than parametric models for identification purposes.

Section 4.1 reviews the IF model in detail, and discusses the main qualities and drawbacks of lumped models of the glottis. Sections 4.2 and 4.3 describe two different approaches for simplifying the IF model. Both the proposed models rely on a common assumption, i.e. the vocal fold is treated as a single mass. However, the two models differ in the treatment of the non-linear block that accounts for the interaction between vocal folds and glottal airflow.

The physically-informed model presented in Sec. 4.2 describes the non-linear block

---

This chapter is partially based on [11, 41, 42].
using a black-box approach: no physical information is used to describe the interaction, instead it is modeled as a regressor-based mapping that relates the fold displacement to the glottal flow. This approach permits to use the model for identification purposes, i.e. the model parameters can be estimated in order to fit a given target glottal waveform. At the same time, the physical structure of the model can be exploited for controlling its behavior during resynthesis.

The one-delayed-mass model presented in Sec. 4.3 describes the non-linear block using a white-box (physical) approach. The pressure distribution along the glottis is modeled starting from the IF equations. The equations are then simplified by exploiting additional constraints in the vocal fold motion. As a result, the effects due to phase differences between the upper and lower margins of the folds are incorporated in the non-linear equations. Numerical simulations show the behavior of the one-delayed-mass model resembles that of the IF model, while using about half of the parameters and half of the parameters.

Table 3.1 summarizes the main variables and parameters that are used throughout the chapter.

### 4.1 Glottal models

Section 4.1.1 describes the Ishizaka-Flanagan model, where each vocal fold is treated using two lumped masses connected by springs and damping elements. Section 4.1.2 discusses the properties of the IF model and the use of lumped models for the synthesis of glottal signals.

#### 4.1.1 The Ishizaka-Flanagan model

The human vocal folds consist of two opposing ligaments, that form a constriction at the beginning of the trachea. The orifice formed by the two folds is called the glottis. In the production of voiced sounds, the fundamental frequency and the quality of the sounds is to a large extent controlled by the quasi-periodic oscillations of the vocal folds. When sufficient subglottal pressure $p_s$ is provided by the lungs, the folds start to vibrate and act as a pressure controlled valve, that tends to open for positive glottal pressures and to close for negative pressures (according to the notation introduced by Fletcher [48], this behavior corresponds to that of a $(+, +)$ valve). In normal phonation the folds typically collide with each other, and therefore the glottis is completely closed for a certain part of the oscillation. As a consequence, the quasi-periodic waveform describing the glottal airflow is made of a series of positive “puffs” produced when the glottis is open, and is zero when the glottis is closed.

The Ishizaka-Flanagan (IF) model describes each vocal fold using a two-mass approximation. Moreover, it assumes that the folds are bilaterally symmetric, so that only one of the needs to be modeled. As a consequence, the whole model is constructed using two masses, as depicted in Fig. 4.1. The masses are permitted only lateral motion (as in the coordinates $x_1$ and $x_2$ given in Fig. 4.1(b)). Along this lateral direction, the masses are
<table>
<thead>
<tr>
<th>quantity</th>
<th>symbol</th>
<th>unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air shear viscosity</td>
<td>ν = 1.85 \cdot 10^{-5}</td>
<td>[N \cdot s/m^2]</td>
</tr>
<tr>
<td>Vocal tract input area</td>
<td>S = 5 \cdot 10^{-4}</td>
<td>[m²]</td>
</tr>
<tr>
<td>Press. at tract entrance</td>
<td>p(t)</td>
<td>[Pa]</td>
</tr>
<tr>
<td>Press. waves</td>
<td>p(±)(t)</td>
<td>[Pa]</td>
</tr>
<tr>
<td>Glottal flow</td>
<td>u(t)</td>
<td>[m^3/s]</td>
</tr>
<tr>
<td>Vocal fold length</td>
<td>l_g</td>
<td>[m]</td>
</tr>
</tbody>
</table>

**IF model**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocal fold masses</td>
<td>m_k, k = 1, 2</td>
<td>[Kg]</td>
</tr>
<tr>
<td>Stiffness on m_k</td>
<td>k_k(x_k)</td>
<td>[N/m]</td>
</tr>
<tr>
<td>Viscous resist. of m_k</td>
<td>r_k = 0.1 \sqrt{m_k k_k}</td>
<td>[N \cdot s/m]</td>
</tr>
<tr>
<td>Thickness of m_k</td>
<td>d_k</td>
<td>[m]</td>
</tr>
<tr>
<td>Displacement of m_k</td>
<td>x_k(t)</td>
<td>[m]</td>
</tr>
<tr>
<td>Rest position of m_k</td>
<td>x_0_k</td>
<td>[m]</td>
</tr>
<tr>
<td>Glottal area under m_k</td>
<td>A_k(t) = 2x_k(t)l_g</td>
<td>[m²]</td>
</tr>
<tr>
<td>Area under m_k at rest</td>
<td>A_0_k = 2x_0_kl_g</td>
<td>[m²]</td>
</tr>
</tbody>
</table>

**1-mass models**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocal fold mass</td>
<td>m</td>
<td>[Kg]</td>
</tr>
<tr>
<td>Fold stiffness</td>
<td>k</td>
<td>[N/m]</td>
</tr>
<tr>
<td>Fold viscous resist.</td>
<td>r = 0.1 \sqrt{m k}</td>
<td>[N \cdot s/m]</td>
</tr>
<tr>
<td>Fold thickness</td>
<td>d</td>
<td>[m]</td>
</tr>
<tr>
<td>Fold displacement</td>
<td>x(t)</td>
<td>[m]</td>
</tr>
<tr>
<td>Fold rest position</td>
<td>x_0</td>
<td>[m]</td>
</tr>
<tr>
<td>Glottal area</td>
<td>A(t) = 2x(t)l_g</td>
<td>[m²]</td>
</tr>
<tr>
<td>Glottal area at rest</td>
<td>A_0 = 2x_0l_g</td>
<td>[m²]</td>
</tr>
</tbody>
</table>

Table 4.1: Symbols used throughout the chapter.

assumed to behave as simple second-order mechanical oscillators, i.e. they are subject to elastic and dissipative forces. For the accurate simulation of the elastic properties of the fold, the springs are non-linear and \( k_1(x_1) \) and \( k_2(x_2) \) are modeled as quadratic functions of the corresponding displacements. In addition, the two masses are coupled through a third spring \( k_{12} \).

Collisions between the folds are modeled by adding to the equations an additional restoring contact force, which is represented by an equivalent non-linear spring. In other words, when one of the masses \( m_k \) collides (i.e., when the condition \( x_k < 0 \) holds), its stiffness \( k_k(x_k) \) increases. Summarizing, the equations for the mechanical system are
Figure 4.1: Schematic representation of the Ishizaka-Flanagan model: (a) coronal view and (b) superior view.

given by

\[
\begin{align*}
    m_1 \ddot{x}_1(t) + r_1 \dot{x}_1(t) + k_1(x_1(t) - x_{01}) + k_{12}[x_1(t) - x_2(t)] &= l_g d_1 p_{m1}(t), \\
    m_2 \ddot{x}_2(t) + r_2 \dot{x}_2(t) + k_2(x_2(t) - x_{02}) - k_{12}[x_1(t) - x_2(t)] &= l_g d_2 p_{m2}(t),
\end{align*}
\]

(4.1)

where \(l_g d_1\) and \(l_g d_2\) are the driving surfaces of the masses, on which the pressures \(p_{m1}\) and \(p_{m2}\) act. The authors claim that using two masses can account for most of the relevant glottal detail, including the modeling of phase differences in the motion of the lower and upper edges of the folds.

The interaction of the mechanical model with the glottal pressure distribution is derived under the assumption of quasi-steady glottal flow. Ishizaka and Flanagan approximate the pressure distribution inside the glottis as successive discrete steps \(p_{ij}\) at each end \(j\) of each mass \(i\) (see Fig. 4.1(a)).

The first pressure drop \(p_s - p_{11}\) is derived from the Bernoulli law for an ideal fluid in the static regime. According to the authors, the \textit{vena contracta} produced by the abrupt contraction in the cross-sectional area makes the glottal area \(A_1\) under the mass \(m_1\) appears smaller than it is. As a consequence, the drop is greater than in the ideal case. Along the masses, the pressure drops \(p_{11} - p_{12}\) and \(p_{21} - p_{22}\) are governed by viscous losses and are proportional to the air shear viscosity \(\nu\). At the junction between the areas \(A_1\) and \(A_2\), the change in pressure \(p_{12} - p_{21}\) equals the change in kinetic energy per unit volume of the fluid. Finally, the abrupt expansion at the upper edge of the
glottis causes the pressure to recover toward the atmospheric value. An estimate for the pressure recovery is found by Ishizaka and Flanagan by imposing continuity of the flow. Summarizing, the pressure distribution along the glottis is given by the following equations:

\[
\begin{align*}
    p_s - p_{11}(t) &= 0.69 \rho_{\text{air}} \frac{u(t)^2}{A_1(t)^2}, \\
    p_{11}(t) - p_{12}(t) &= 12 \nu d_1 \frac{l_g^2 u(t)}{A_1(t)^3}, \\
    p_{12}(t) - p_{21}(t) &= \frac{1}{2} \rho_{\text{air}} u(t)^2 \left( \frac{1}{A_2(t)^2} - \frac{1}{A_1(t)^2} \right), \\
    p_{21}(t) - p_{22}(t) &= 12 \nu d_2 \frac{l_g^2 u(t)}{A_2(t)^3}, \\
    p_{22}(t) - p(t) &= \frac{1}{2} \rho_{\text{air}} \frac{u(t)^2}{A_2(t)^2} \left[ 2 \frac{A_2(t)}{S} \left( 1 - \frac{A_2(t)}{S} \right) \right].
\end{align*}
\]

The authors also discuss the inclusion of air inertance in the equations, when time-varying conditions are considered.

Given the pressure distribution of Eq. (4.2), the driving pressures \( p_{m1} \) and \( p_{m2} \) acting on the two masses have to be derived. In the IF model, these are simply assumed to be the mean pressures along each mass:

\[
\begin{align*}
    p_{m1}(t) &= \frac{1}{2} [p_{11}(t) + p_{12}(t)] = p_s - 0.69 \rho_{\text{air}} \frac{u(t)^2}{A_1(t)^2} - 6 \nu d_1 \frac{l_g^2 u(t)}{A_1(t)^3}, \\
    p_{m2}(t) &= \frac{1}{2} [p_{21}(t) + p_{22}(t)] = p(t) + \frac{1}{2} \rho_{\text{air}} \frac{u(t)^2}{A_2(t)^2} \left[ 2 \frac{A_2(t)}{S} \left( 1 - \frac{A_2(t)}{S} \right) \right] + 6 \nu d_2 \frac{l_g^2 u(t)}{A_2(t)^3}. 
\end{align*}
\]

In conclusion, the IF model is completely described by Eqs. (4.1) and (4.3). Note that it is structurally very similar to the single reed model described in Sec. 3.1. The two main differences between the systems are that (1) the reed model uses only one mass while the IF model uses two, and (2) the single reed closes for positive pressures while the glottis opens.

### 4.1.2 Properties of lumped glottal models

The model described in the last section was implemented numerically by Ishizaka and Flanagan using the backward Euler method, and coupled to a vocal tract model. From the numerical simulations (run with sampling rates ranging from 10 to 30 [kHz]) the authors studied the dependence of the model behavior on the physical parameters and observed that realistic glottal signals can be obtained. The model can take into account subtle features that are not reproduced by a parametric model. In particular, acoustic
interaction with the vocal tract is considered, thus allowing to develop a full articulatory model [129, 140]. This interaction gives rise to several “natural” effects: occurrence of oscillatory ripples on the glottal flow waveform, as well as a slight influence of the load characteristics on the pitch and the open quotient (i.e., the ratio between the open and closed phases in the glottal waveform).

Many refinements have been proposed to the IF model, in which the vocal folds are treated using a larger number of masses, or the description of the airflow though the glottis is modified. An example is the three-mass model by Story and Titze [131], where one large mass is used for describing the internal parts of the fold (the body), while two smaller masses account for the motion of the external part (the cover). On the other hand, simpler one-mass models are used by many authors in articulatory speech synthesizers (see e.g. [91]), or even in modeling non-human phonatory systems [47, 75]. One advantage of one-mass models lies in their reduced computational loads and better controllability. However, these models are not able to account for phase differences in the vocal fold motion.

Berry and Titze [21] studied the glottal system using a different modeling approach. In their study the vocal folds are modeled as a distributed, three-dimensional system, and is discretized using Finite Element methods. Using this numerical model the authors studied the normal modes of oscillation of the system. Assuming that the fold tissue is nearly incompressible, the first two excited modes in the model were found to be the ones depicted in Fig. 4.2. Interestingly, the two masses in the IF model also have two eigenmodes which are conceptually equivalent to those of the distributed model. The IF mode where the two

![Figure 4.2: First two excited modes in a distributed model of the vocal folds.](image-url)
masses move $\pi$-out of phase corresponds approximately to the one in Fig. 4.2(a), while
the mode with the two masses in phase corresponds to that of Fig. 4.2(b). Berry and
Titze suggest that the success of the IF model in describing the glottal behavior might
be attributed to its ability to capture these two eigenmodes, and therefore facilitate
self-oscillation.

A serious objection to all the models mentioned above has been raised by Villain et al.
[150]. These authors remark that elementary mechanical constraints on the physiological
problem are completely neglected in these models. As an example, in both the lumped
and the distributed approach it is assumed that the elastic structure is fixed to a rigid
wall, which is clearly a very crude approximation since in reality a significant radiation
of surface waves from the throat can be noticed when voiced sounds are produced. The
effect of this radiation may be significant in terms of energy loss in the system. Villain
et al. used an experimental setup where the folds are modeled by thin latex tubes
filled with water. Two mechanical boundary conditions were considered: one where
the volume of the “water” fold is kept at a constant value and a second one where
the internal water pressure remains constant independently on the blowing pressure.
The experiments showed that the behavior of the latex valve is strongly affected by the
mechanical constraints.

Another objection that can be raised to lumped models has to do with the glottal
closure. As an example, in the IF model the glottal areas $A_1$ and $A_2$ are assumed to be
rectangular. As a consequence, closure of the glottis occurs in an abrupt manner and the
flow signals obtained from the model exhibit a sharp corner at the beginning of the closed
phase. Equivalently, the airflow derivative exhibits a narrow negative peak at closure.
This phenomenon affects the spectral tilt of the glottal source, introducing additional
energy at high frequencies. In natural flow signals, a smoother glottal closure is usually
observed. For example, stroboscopic measurements often show zipper-like movements of
the glottal area during the closing phase. The IF model clearly does not take into account
these phenomena.

Finally, when used in speech synthesis applications the IF model suffers from an over-
parametrization: as many as 19 parameters have to be estimated in order to account for
non-linear corrections in the elastic forces, for collisions between the two folds, and other
features. This results in problems in tuning the parameters. Proposed refinements to IF
(such as the three-mass model by Story and Titze [131]) involve an even larger number
of parameters and are hardly controllable and more computationally expensive.

4.2 Non-linear block: identification

The model presented in this section relies on a hybrid approach. Its structure is
similar to that of IF, but not all of the functional blocks are modeled following through
a physical description.

The vocal fold is treated as a lumped mass subject to an elastic restoring force and
to dissipation. The mass is driven by the pressure $p_g$ at glottis. Therefore, in this
approximation the displacement $x$ of the mass is governed by the second-order linear oscillator equation:

$$m \ddot{x}(t) + r \dot{x}(t) + k(x(t) - x_0) = l_g d p_g(t),$$

where $l_g d$ is the driving surface on which the pressure $p_g$ acts.

A non-linear block accounts for interaction with glottal pressure. Unlike IF, however, no physical information is retained in the non-linear block. This is treated as a black box and described by a regressor-based mapping. As such, the model can be said to be physically-informed rather than really physical. This permits to exploit advantages of both the parametric and the physical approach. Namely, given a glottal flow signal the weights for the regressors can be estimated using non-linear identification techniques, in order to fit the waveform.

### 4.2.1 The identification procedure

In the following, the model is described in the discrete-time domain. A schematic representation and a block diagram are depicted in Fig. 4.3. Analogously to the IF model, it is assumed that the valve is perfectly symmetrical, so that only one fold needs to be modeled.

In the discrete-time domain, each vocal fold is described during the open and closed phases as a linear second-order oscillator, whose transfer function is denoted by $H_r$ and given by

$$H_r(z) = \frac{\beta_0}{(1 + \alpha_1 z^{-1} + \alpha_2 z^{-2})}. \quad (4.5)$$

The filter $H_r(z)$ is the digital equivalent of the continuous-time system (4.4). It is completely defined by its gain factor $\beta_0$, its center frequency $\omega_0$, and its quality factor $q_0$. Given the parameters $(\omega_0, q_0)$, the poles $p_d = re^{\pm j\phi}$ of the filter $H_r$ are defined in the
following way [92]:

\[
  r = 1 - \frac{\omega_0}{2q_0}, \quad \cos(\phi) = 2r \frac{\cos(\omega_0)}{1 + r^2}.
\]

The coefficients \(a_2, a_3\) in the denominator of \(H_r\) are then found as \(a_2 = -2r \cos(\phi)\) and \(a_3 = r^2\).

The fold displacement \(x\) is related to the glottal pressure \(p_g\) through the equation \(X(z) = H_r(z)p_g(z)\). Collisions between the two folds are treated using an approach analogous to that adopted in the IF model: during collisions, an elastic restoring contact force is added to the mass. Equivalently, during collisions the resonance \(\omega_0\) of \(H_r\) is increased.

The glottal flow is assumed to depend non-linearly on the glottal area, or equivalently on the fold displacement. A shown in Sec. 4.1, physical models such as IF describe this dependence analytically using very crude simplifications (e.g. quasi-steadiness of the flow). A different viewpoint is taken here: no attempt is made to model the glottal flow behavior in a physical way, and instead the dependence of the glottal flow on the displacement is modeled using a black-box approach. In this approach, a non-linear regressor-based mapping \(F\) relates flow to fold displacement. The inputs to the non-linear block are \(x(n)\) and \(x(n - 1)\):

\[
  Z_0 u(n) = \begin{cases}
    F(n) = \sum_{i=0}^{M} w_i \psi_i(n), & \text{if } x(n) > 0, \\
    0, & \text{if } x(n) \leq 0,
  \end{cases}
\]

(4.6)

where \(Z_0 = \rho_{\text{air}} c / S\) is the wave impedance of the vocal tract, while the regressors are denoted as \(\psi_i(n) = \psi_i(x(n), x(n - 1))\).

Several strategies are possible for the choice of the regressor set \(\{\psi_i\}\). Local models, such as gaussian functions or any other radial basis function, are often used. Radial Basis Function Networks (RBFN) [31] are often used in the field of time series analysis and modeling. Alternatively, a polynomial expansion of the input can be used. This choice leads to a class of so-called NARMAX models [30], known in the fields of system identification and control. Finally, the regressors can be derived on the basis of physical considerations.

In the following, a third order polynomial expansion in \(x(n), x(n - 1)\) is used. One reason for using \(x(n - 1)\) is that the vocal fold velocity \(\dot{x}\) is assumed to contribute to the total flow. This assumption is made in analogy to the single reed model described in chapter 3 (see in particular Eq. (3.5)). Taking into account \(\dot{x}\) corresponds, in the discrete-time domain, to taking into account at least one past value of \(x\). This qualitative discussion justifies the inclusion of the term \(x(n - 1)\) as an input variable to the regressor set \(\{\psi_i\}\).

Interaction with the vocal tract is drastically simplified by neglecting pressure recovery at the vocal tract entrance. In other words, it is assumed that the pressure \(p\) at the vocal tract entrance equals the glottal pressure \(p_g\). Again, the vocal tract model can only be said to be physically-informed, and not physical, since the assumption \(p_g = p\) has
no counterpart in the physical reality. Under this assumption, the glottal pressure $p_g$ is related to the glottal flow through the input impedance $Z_{in}(z)$ of the vocal tract. Equivalently, wave variables $p^\pm = (p_g \pm Z_0 u) / 2$ can be used: then $p^\pm$ are related to each other through the vocal tract reflectance $R(z)$:

$$p_g(n) = z_{in}(n) * u(n), \quad \iff \quad p^-(n) = r(n) * p^+(n).$$

A final remark about Fig. 4.3 concerns the insertion of a fictitious delay element $z^{-1}$; this is needed in order to compute the delay-free feedback path in the numerical model. The K method has been used elsewhere in this thesis for dealing in a more accurate way with such computability problems. However, in this case the insertion of $z^{-1}$ does not deteriorate or anyhow affect the properties of the numerical system. The reasons for this are explained at the end of this section.

The following problem is now addressed: given a target glottal flow waveform $Z_0 u$ (be it a synthetic signal or an inverse-filtered one), the physically-informed model has to be identified so that the output $Z_0 u$ from the non-linear block fits the target as closely as possible. Many works are available in the literature [32, 101, 132] where this problem has been studied using parametric models of the glottal source (e.g. the Liljencrants-Fant model). In the case under consideration here, a physical model of the source has to be identified rather than a parametric model of the signal, which makes the problem not trivial.

The identification procedure described in the following holds for both synthetic and real flow waveforms. In both cases a single period is chosen and a periodicized signal is constructed and used as target. The example utterance analyzed in this section is a sustained /a/ vowel produced in normal phonation by a male speaker. The inverse-filtered flow signals has been computed using an automatic method developed by Alku and described in [5]. This method estimates the glottal flow directly from the acoustic speech pressure waveform using a two-stage structure, where LP analysis is used as a computational tool.

The system is identified in three main steps.

**Step 1** From the target $Z_0 u$, the corresponding glottal pressure signal $\tilde{p}_g$ is computed using Eq. (4.7). In order to do that, the reflectance $R(z)$ of the vocal tract has to be known. In the following $R$ is arbitrarily chosen to be that of a uniform vocal tract, and is implemented using waveguide modeling. Therefore, $R(z) = z^{-2 m_L} R_{load}(z)$, where $m_L$ defines the length (in samples) of the tract and $R_{load}$ accounts for the reflection at mouth and has a low-pass characteristics. Note that the above defined $R(z)$ is formally identical to the one used for the idealized clarinet bore in Sec. 3.2 (see in particular Eq. (3.14)).

**Step 2** The linear block $H_r(z)$ is driven using the synthesized target pressure $\tilde{p}_g$, and the output $\tilde{x}$ is computed. At this stage, resonances for $H_r$ in the open phase and the closed phase are chosen interactively, in such a way that the open and closed phase for the target fold displacement $\tilde{x}$ match those of the target flow $\tilde{u}$ (in other
words, $\bar{x} = 0 \Leftrightarrow \bar{u} = 0$, see the example in Fig. 4.4). The value $q_0 = 10$ is chosen for the quality factor. This value is deduced from the parameters used by Ishizaka and Flanagan in [73].

**Step 3** A complete input-output description of the non-linear block is available at this point, where the inputs are $\bar{x}(n), \bar{x}(n-1)$ and the output is the target $\bar{u}(n)$. Therefore, during the open phase the weights $w = [w_0 \ldots w_M]$ for the regressor set $\{\psi\}_i$ can be identified using the non-linear identification technique described in the following.

The assumption made in **Step 1** is questionable, since for real utterances the vocal tract cross-section is obviously not uniform. However, in the physically-informed approach adopted here a pressure signal that provides plausible rather than realistic excitation to the vocal fold is needed. Figure 4.4(a) summarizes the three steps of the identification procedure. Figure 4.4(b) shows the time-derivatives of the target flow $\bar{u}$ and the identified flow $u$.

The identification of the weights $w$ in **Step 3** is carried out as follows. A training window with starting time $l$ and length $N$ is chosen. Inside the window, two training sets $T_u$ and $T_\psi$ are defined as

$$T_u = [\bar{u}(l+1), \bar{u}(l+2), \ldots, \bar{u}(l+N)] ,$$

$$T_\psi = \begin{bmatrix} \psi_0(l+1) & \cdots & \psi_0(l+N) \\ \vdots & \ddots & \vdots \\ \psi_M(l+1) & \cdots & \psi_M(l+N) \end{bmatrix} .$$

Using these definitions and Eq. (4.6), it is seen that the weights $w$ must solve the system
\( \mathbf{w} \cdot \mathbf{T}_\psi = \mathbf{T}_u \). The least-squares solution of such a system is known [83] to be

\[
\mathbf{w} = \mathbf{T}_u \cdot \mathbf{T}_\psi^+,
\]

(4.9)

where the symbol + has the meaning of pseudo-inversion.

It is now clear that the insertion of \( z^{-1} \) does not deteriorate the accuracy of the model: given the structure in Fig. 4.3 and a flow signal, the identification of the non-linear block automatically takes into account the \( z^{-1} \) element.

### 4.2.2 Results and applications

The identification procedure described in the last section has been tested using both real and synthetic signals. The performance is generally good and comparable to that obtained using signal-based models [32, 101, 132]. In particular, from the example in Fig. 4.4(b) one can see that in the closing phase both the width and the amplitude of the negative pulse in the flow derivative are well approximated. This portion is the most important to be fitted accurately, since it defines most of the spectral (and perceptual) features of the glottal source signal.

Moreover, large bandwidths and the consequent noise in the flow derivative waveform are seen not to affect the identification procedure. For the signal used in Fig. 4.4, the bandwidth is 11.025 [kHz], and considerable noise can be noticed in the flow derivative signal. However, the proposed identification procedure uses the glottal flow –rather than its derivative– as the target signal, while signal-based models (such as Liljencrants-Fant) typically try to fit the derivative of the glottal flow. It has been verified that large bandwidths can deteriorate the performance of identification techniques based on the Liljencrants-Fant model [101].

However, problems are encountered when strong ripples (due to interaction with tract formants) appear on the opening phase of the target signal: these can deteriorate the accuracy of identification drastically. This problem can be solved by focusing the accuracy of the identification on the portions of the signal (such as the closing phase) that are known to be the most important, and to loosen the accuracy requirements for those features (such as the ripples in the opening phase) that are known to be perceptually less relevant. One possible strategy for achieving this goal is to pre-process the target signal by applying time-warping, such that the relevant portions are magnified while the rest of the signal is time-compressed. However, at present no attempt has been made to implement this strategy.

The target signal to be identified is obtained through periodicization of one cycle of the flow waveform. Therefore one open problem is concerned with identification of non-periodic signals. A straightforward strategy is to identify the system parameters once for each period of analyzed data (as already done by Childers and Ahn [32] with the Liljencrants-Fant model). Alternatively, a single set \( \mathbf{w} \) of weights may be used while adjusting only the filter parameters \( \omega_0 \) and \( q_0 \) in order to account for changes in pitch and amplitude. However, one further problem is that the identification procedure is far from being automatic: while the weights \( \{w_i\} \) are estimated automatically from Eq. (4.9),
the linear filter $H_r$ is adjusted interactively. A technique for automatic estimation of the filter parameters has to be found in order to extend the identification procedure to the non-periodic case. The remaining of this section addresses the use of the proposed model for speech synthesis purposes and for voice quality assessment.

The physically-informed model can be used after identification for resynthesizing the target glottal flow signals. When studying the system behavior in autonomous evolution, it is seen that it reaches steady state self-sustained oscillations after a short transient. The steady state waveform coincides with the one obtained from identification. If the values of some of the model parameters are then adjusted, the system exhibits robustness to such changes. Figure 4.5 shows an example where $\omega_0$ is increased after the system has reached the steady state: as a consequence the pitch of the signal increases correspondingly, while the flow shape is preserved. An increase in amplitude during pitch transition can also be noticed from Fig. 4.5. After the transition, the maximum amplitude turns back to its original value.

This example shows that the physically-informed parameters of the model can be used to perform transformations on the resynthesized signal (as an example, pitch shift or vibrato effects are obtained by adjusting $\omega_0$). A major limitation is that the identification procedure is not able to guarantee a priori that the identified system is stable during resynthesis. Different approaches can be used in order to guarantee stability, such as the harmonic balance technique [64] or the imposition of additional constraints on the gradient $\nabla F$ [40].

A second potential application of the model is in the analysis and assessment of voice quality, including the detection and classification of voice pathologies. Voice quality assessment is traditionally based on subjective perceptual rating, which is still in current studies considered the only reasonable way to classify certain types of voice disorders.
The objective assessment of voice, however, is still an open problem that calls for reliable analysis tools and algorithms. Many researchers [33, 60] use a set of parameters derived from analysis of the acoustic signal (i.e., the radiated pressure): examples of such parameters are time-domain measures such as jitter (pitch variation in successive oscillation periods) and shimmer (magnitude variation in successive oscillation periods), or frequency-domain parameters such as the spectral slope and the spectral flatness of the inverse-filtered signal. The main advantage of these techniques lies in their non-invasive nature.

However, all of the above mentioned parameters depend exclusively on the signal features, and no assumptions are made on the physiology of the source. As a consequence, signal-based analysis retains a mixed information in which the contributions of the vocal tract and the glottis are undistinguishable. Using a physically-informed model permits to localize the observation at the non-linear excitation mechanism, limiting the influence of other elements. The values of the identified weights $w$ provide information on the glottal flow waveform, while analysis of their variations in time can be used to study the stability of the waveform. Early results on voice quality assessment using the proposed physically-informed model have been presented in [42].

4.3 Non-linear block: modified interaction

In Sec. 4.2, the IF model has been simplified by treating the vocal fold as a single mass, and by describing the interaction with the glottal flow using a black-box approach. The model presented in this section follows a different strategy. The vocal fold is still treated a single mass, but the interaction with the airflow is modeled through a physical description. The effect of the second mass of the IF model is taken into account by introducing a delay $t_0$ in the mass position. The glottal airflow is assumed to depend on this “delayed mass”, and the non-linear aerodynamics equations are consequently modified.

Results from the simulations presented in Sec. 4.3.2 show that the model behaves qualitatively as IF, while using only one degree of freedom (one mass) instead of two. As a consequence, less than half of the IF parameters are needed. Among them, the delay $t_0$ gives control on the airflow skewness, which is known to be a perceptually relevant feature [32]. Having a small set of meaningful control parameters, the proposed physical model can be “competitive” with parametric ones, such as Liljencrants-Fant.

4.3.1 A one-delayed-mass-model

As shown in Sec. 4.1.1, Ishizaka and Flanagan describe the pressure drops $p_{ij}$ along the vocal folds according to Eq. (4.2). Therefore the positions $x_1$ and $x_2$ of both masses are needed in order to compute the pressure drops and the resulting airflow. The “one-delayed-mass model” presented here avoids the use of a second mass by exploiting additional information on the system.
• The IF model has two eigenmodes: the one with two masses in phase and the one with two masses $\pi$-out of phase. As already remarked in Sec. 4.1.2, these modes correspond roughly to the first two excited modes observed by Berry and Titze [21] using a distributed model of the vocal folds (see Fig. 4.3). Berry and Titze found that the two eigenfrequencies are very closely spaced. As a consequence, 1 : 1 mode locking occurs during self-oscillation.

• In a recent paper, de Vries et al. [36] used a similar distributed model for estimating “correct” values for the IF parameters. Such values are found by requiring the behavior of the IF model to resemble as closely as possible that of the distributed model. Their results differ significantly with the values originally stated by Ishizaka and Flanagan. In particular the parameter values for the two masses are found in [36] to be much more symmetrical: the ratio between $m_1$ and $m_2$ is close to one (while it is close to five in the IF parameters), and the same holds for the spring constants, damping factors and geometrical parameters ($d_1$ and $d_2$ in Fig. 4.1 are found to be the same).

Using this additional information, the IF model can be consistently simplified using the following assumptions.

a1. The masses $m_{1,2}$ are taken to be equal, together with their thickness $d_{1,2}$ and their spring constants and damping factors.

a2. The two masses are taken to move with constant phase difference, because of mode locking; this means that the area $A_2(t)$ under the second mass follows the first on $A_1(t)$ with a constant phase difference.
The assumption $a2$ can be restated as:

$$A_2(t) = A_1(t - t_0), \quad (4.10)$$

where $t_0$ represents the delay (in seconds) between the motion of the upper and lower edges of the fold. As a consequence, the pressure distribution $p_{ij}$ in Eqs. (4.2) can be written as

$$
\begin{align*}
  p_s - p_{11}(t) &= 0.69 \rho_{\text{air}} \frac{u(t)^2}{A(t)^2}, \\
  p_{11}(t) - p_{12}(t) &= 12 \nu_1 \frac{l_2^2 u(t)}{A(t)^3}, \\
  p_{12}(t) - p_{21}(t) &= \frac{1}{2} \rho_{\text{air}} u(t)^2 \left( \frac{1}{A(t-t_0)^2} - \frac{1}{A(t)^2} \right), \\
  p_{21}(t) - p_{22}(t) &= 12 \nu_2 \frac{l_2^2 u(t)}{A(t-t_0)^3}, \\
  p_{22}(t) - p(t) &= \frac{1}{2} \rho_{\text{air}} \frac{u(t)^2}{A(t-t_0)^2} \left[ \frac{2A(t-t_0)}{S} \left( 1 - \frac{A(t-t_0)}{S} \right) \right], \\
\end{align*}
\quad (4.11)$$

where $A(t) \equiv A_1(t)$. In Eqs. (4.11), the pressure drops are non-linear functions of the area $A$ and the same area delayed by $t_0$. This suggests that only one degree of freedom is needed in the model.

Therefore, in the following the fold is described using a single mass and is treated as a linear second-order oscillator. Similarly to IF, the driving pressure acting on the fold is chosen to be the mean pressure $p_m$ at the glottis:

$$m \ddot{x}(t) + r \dot{x}(t) + k(x(t) - x_0) = l_g dp_m(t), \quad (4.12)$$

where $l_g d$ is the driving surface on which the pressure drop acts. From the assumption $a1$, the mass $m$ is given by $m = 2m_1$. Explicit expressions for the driving pressure $p_m$ and the pressure at vocal tract entrance $p$ are derived from Eq. (4.11), and depend only on the variables $(A(t), A(t-t_0), u(t))$:

$$
\begin{align*}
  p_m(t) &= p_m(A(t), A(t-t_0), u(t)), \\
  p(t) &= p(A(t), A(t-t_0), u(t)). \\
\end{align*} \quad (4.13)$$

In this way, the effects due to phase differences between the upper and lower margins of the folds are incorporated in the non-linear equations (4.13), and are controlled by the delay $t_0$. One last equation relates the glottal flow to the pressure $p$ at the vocal tract entrance:

$$u(t) = z_{in}(t) \ast p(t), \quad (4.14)$$

where the load impedance $z_{in}$ is the input impedance of the vocal tract.
Equations (4.12),(4.13),(4.14) describe the one-delayed-mass model. From Eq. (4.12), it is seen to be a one-mass model, but the dependence on the delayed area $A_1(t-t_0)$ in Eq. (4.13) results in a modified non-linear block. A graphic representation of the model, as opposed to IF, is depicted in Fig. 4.6. As shown in the following section, by introducing the parameter $t_0$ it is possible to preserve the main features of a two-mass model while using a single degree of freedom.

### 4.3.2 Numerical simulations

The linear differential Eq. (4.12) is discretized using the bilinear transform, i.e. each occurrence of $s$ in the Laplace-transformed Eq. (4.12) is substituted with the mapping (see Sec. 2.3.1)

$$s = 2F_s \frac{1 - z^{-1}}{1 + z^{-1}}.$$  

(4.15)

It is easily seen that the resulting numerical system resembles the general structure described in Sec. 2.3.2 and depicted in Fig. 2.11. Therefore, a delay-free path is generated between the linear and the non-linear blocks of the system:

$$\begin{align*}
  w(n) &= \tilde{w}(n) + C p(n), \\
  x(n) &= \tilde{x}(n) + K p(n), \\
  p(n) &= f(\tilde{x}(n) + K p(n)),
\end{align*}$$  

(4.16)

where the variables are given by

$$\begin{align*}
  w(n) &= \begin{bmatrix} x(n) \\
                      \dot{x}(n) \end{bmatrix}, \\
  u(n) &= \begin{bmatrix} x_0 \\
                       p_s \\
                       p_0^{-1}(n) \end{bmatrix}, \\
  x(n) &= \begin{bmatrix} u(n) \\
                       x(n) \\
                       x(n-n_0) \end{bmatrix}, \\
  p(n) &= \begin{bmatrix} p_0(n) \\
                          p(n) \end{bmatrix},
\end{align*}$$

and where the numerical delay $n_0$ is defined as $n_0 = t_0 F_s$. The vectors $\tilde{w}(n)$ and $\tilde{x}(n)$ have no instantaneous dependence on the pressures $p(n)$, and are therefore computable at each step. The non-linear mapping $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is simply Eq. (4.13) restated in vector notation.

The K method is used in order to compute the delay-free path. The mapping $f$ is sheared and the non-linear equation in system (4.16) is turned into a new-one:

$$p(n) = f(\tilde{x}(n) + K p(n)) \quad \text{K method} \quad p(n) = h(\tilde{x}(n)).$$  

(4.17)

The new mapping $h(\tilde{x})$ is computed iteratively at sample rate, according to the discussion at the end of Sec. 2.3.2. The numerical implementation described above is used to study the properties of the one-delayed-mass model.

The dependence on $A_1(t-t_0)$ in Eq. (4.13) results in an additional delay loop in the system. This is a potential source of instability [7]. Due to the non-linear nature of
the system, analytical conditions for stability are not easily found. Therefore, stability properties of the system have to be investigated experimentally. This can be done in two steps: first, simulations are run at a very high sampling rates \( F_s = 200 \text{ [kHz]} \), and these simulations are taken as a reference for the behavior of the continuous-time system. In a second step, simulations are run at standard sampling rates \( F_s = 11.025, 22.05 \text{ [kHz]} \) and compared with the reference.

The results for the case \( F_s = 22.05 \text{ [kHz]} \) are plotted in Fig. 4.7: these show that the numerical delay \( n_0 = t_0F_s \) affects the system’s stability. With very small delays \( (t_0 < 2 \cdot 10^{-4} \text{ [s]}, \text{i.e. } n_0 < 4 \text{ at } F_s = 22.05 \text{ [kHz]} \) the system is unstable, as seen from Fig. 4.7(a): the first few cycles in the oscillation show the increasing error with respect to the reference. In the following cycles, this trend continues until a steady state far from the reference is reached. In general, results from the simulations showed that for values \( n_0 > 4 \) the system appears to be stable. Figure 4.7(b) shows that with a delay \( t_0 = 2 \cdot 10^{-4} \text{ [s] } \text{i.e. } n_0 = 4 \text{ at } F_s = 22.05 \text{ [kHz]} \) is stable. Analogous results are found for \( F_s = 11.025 \text{ [kHz]} \).

A second experimental study with numerical simulations examines the effect of the delay \( t_0 \) in shaping the glottal waveform. Figure 4.8 shows the areas \( A_1(t), A_1(t - t_0) \) and the airflow \( u(t) \) for two different values of \( t_0 \) (for clarity, the signals are normalized in the figure). From these plots, it can be clearly seen that the skewness of the airflow is controlled by the delay \( t_0 \). A quantitative measure of the flow skewness is given by the speed quotient. This parameter is defined as the ratio between the opening phase \( (\dot{u}(t) > 0) \) and the closing phase \( (\dot{u}(t) < 0) \). The speed quotient is known to have perceptual relevance in characterizing different voice qualities: for instance, analysis on real signals by Childers and Ahn [32] show that the speed quotient ranges from about 1.6 to 3 when the voice quality changes from breathy voice to vocal fry and finally to modal voice.

In order to investigate quantitatively the influence of the delay \( t_0 \) on the signal pa-
Figure 4.8: \textit{Dependence of the flow skewness on the time delay $t_0$. Simulations are run at $F_s = 22.05$ [kHz].}

rameters (such as pitch, open quotient, speed quotient, maximum amplitude), automatic analysis of numerical simulations has to be developed. The following results are obtained by analyzing 0.3 [s] long flow signals, where the values of $t_0$ range from 0.1 to 1.9 [ms].

Figure 4.9(a) shows the dependence of the speed quotient on $t_0$: it is seen that, in the range under consideration, the speed quotient is approximately a linear function of $t_0$. By appropriately choosing $t_0$, one can range from very low up to extremely high values of the speed quotient.

Figure 4.9(b) shows another interesting feature of the system: the maximum amplitude for $u$ exhibits a peak around $t_0 = 8 \cdot 10^{-4}$ [s]. This suggests the existence of an optimum delay $t_0$ that maximizes the aerodynamic input power (defined as mean subglottal pressure times mean glottal flow). The aerodynamic input power is in turn related to the glottal efficiency, usually defined as the ratio of radiated acoustic power to aerodynamic power (i.e., the power delivered to the vocal system by the lungs). Further analysis is needed in order to assess the precise influence of $t_0$ on the glottal efficiency.

The main advantages of the proposed model are its simple structure and its low number of control parameters. On the one hand, only one degree of freedom is needed, instead of two [73] or more [131] usually assumed in higher-dimensional lumped models of the vocal folds. On the other hand, the dependence on $t_0$ in Eq. (4.13) results in realistic glottal flow waveforms, that are not obtained with usual one-mass models [91]. In particular, the results given in this section show that $t_0$ provides control on the airflow skewness. The model is therefore a reasonable trade-off between accuracy of the description and simplicity of the structure.

Interaction of the model with vocal tract loads has not yet been investigated in detail. Preliminary results have been obtained by coupling the one-delayed-mass model with a uniform tube model, implemented as a digital waveguide line. The waveguide implementation is structurally identical to that used in chapter 2 for an ideal clarinet bore (see Eq. (3.14)). The simulations show the occurrence of ripples in the airflow signal, mainly due to interaction with the first resonance of the tract. Moreover, automatic analysis reveals a slight dependence of pitch on the vocal tract characteristics (length of the tube,
low-pass characteristics of the reflection filter). Further efforts have to be devoted to this issue, in order to discuss applications of the proposed glottal model in articulatory speech synthesis.

One drawback of the model concerns closure: as the glottal area is assumed to be rectangular, closure of the glottis occurs abruptly, while in natural flow signals the closure is usually smoother due to, for example, zipper-like movements of the glottal area (see the discussion in Sec 4.1.2. Further studies must therefore concentrate on how to integrate such features into the model.

Summary

The use of lumped physical models for the synthesis of glottal flow signals has been discussed. Section 4.1 has reviewed the existing literature, focusing in particular on the Ishizaka-Flanagan model. The remaining of the chapter has presented two different attempts to simplify the IF model while preserving the physical description of the synthesis algorithms.

The physically-informed model of Sec. 4.2 can be used for identification, starting from real signals. The performance of the model has been shown to be comparable to that obtained from identification schemes that use the Liljencrants-Fant model. The proposed identification scheme is robust with respect to large bandwidths and noise in the target waveforms. After identification, the model can be used to resynthesize the target signals and to modify the voice quality by adjusting the physical parameters. A second application that has been discussed concerns the use of the model for voice quality assessment. Further improvements to the model must point towards a more rigorous study of the stability properties of the identified system, as well as complete automatization of the identification procedure.

The one-delayed-mass model proposed in Sec. 4.3 resembles closely the IF model,
but uses only one degree of freedom instead of two, and about half of the variables and the parameters. Analysis of the numerical simulations has shown that the model is a reasonable trade-off between accuracy of the description and simplicity of the structure. Due to its low computational costs it can be suitable for real-time applications. Further studies must investigate nature of the interaction between the glottal model and vocal tract loads. Another interesting direction for future research is concerned with the modeling of zipper-like movements of the vocal folds, and in general more accurate modeling of the closing phase.