SYNTHESIS OF THE VOICE SOURCE USING A PHYSICALLY-INFORMED MODEL OF THE GLOTTIS

Federico Avanzini$^{1,2}$, Carlo Drioli$^1$ and Paavo Alku$^2$

$^1$ Università di Padova, Dip. di Elettronica ed Informatica. Via Gradenigo 6/A, 35131 Padova, Italy.
$^2$ Helsinki University of Technology, Lab. of Acoustics. P.O. Box 3000, Fin-02015 Espoo, Finland
Email: avanzini@dei.unipd.it

Abstract
A physically-informed glottal model is proposed; some physical information is retained in a linear block that accounts for fold mechanics, while non-linear coupling with the airflow is modeled using a regressor-based mapping. The model is used in an identification/resynthesis scheme. Given a real signal, system parameters are estimated via non-linear identification techniques; then the model is used for resynthesizing the signal. With a proper choice of the regressor set the system accurately fits the target waveform and is stable during resynthesis. Physical parameters can be used to change voice quality and speaker identity.

INTRODUCTION
Features of the glottal source signal (i.e. the glottal flow) carry most of the information that characterizes voice quality and speaker identity [1, 2], and accordingly research on source models is becoming increasingly important in speech synthesis. Parametric models fit the glottal signal with piecewise analytical functions, using a small number of parameters. As an example, the Liljencrants and Fant (LF) model [3] characterizes one cycle of the flow derivative using as few as four parameters. The LF model has been successfully used for fitting flow derivatives computed by inverse filtering real utterances [2, 4, 5]. Physical models describe the glottal system in terms of physiological quantities. The Ishizaka and Flanagan (IF) model [6] treats one vocal fold as two coupled mechanical oscillators, driven by the glottal pressure. Physical models capture the basic non-linear mechanisms that initiate self-sustained oscillations, and can simulate subtle features (e.g. interaction with the vocal tract); however they typically involve many parameters (19 in IF) and are not suitable for identification purposes.

The model presented here relies on a hybrid approach. Its structure is similar to that of IF (and other musical non-linear oscillators [7]): the vocal fold is treated as a linear harmonic oscillator, and 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 allows us to exploit some advantages of both parametric and physical approaches. Given a flow signal, weights for the regressors are estimated using non-linear identification techniques in order to fit the glottal waveform. After identification the system is used for resynthesis and can be controlled using its physical parameters.

PHYSICALLY-INFORMED MODEL
We describe the model in the discrete-time domain; a schematic representation and a block diagram are depicted in Fig. 1. Glottis is treated as a lumped (+, +) valve, this notation meaning that the valve tends to open when an overpressure is applied from both upstream and downstream sides (see Fletcher [8]). We assume the valve to be perfectly symmetrical, so that only one fold needs to be modeled. During the open phase each vocal fold is described as a linear second order oscillator, whose transfer function is denoted by $H_r$. Collisions between the two folds are treated as in the IF model [6]: a restoring contact force is added, made of an elastic and a dissipative component. Equivalently, during collisions the resonance $f_r$ of $H_r$ is increased and its quality factor $q_r$ is lowered. Glottal flow is assumed to depend non-linearly on glottal area,
or equivalently on fold displacement. Physical models, such as IF, describe this dependence analytically using very crude simplifications (e.g. quasi-steadiness of the flow). Differently from IF, we choose a black box approach, in which a non-linear regressor-based mapping relates flow to fold displacement. The inputs to the non-linear block are \( x(n) \) and \( x(n-1) \).

The non-linear black box, the vocal tract reflectance and the wave impedance at tract entrance, respectively.

\[
Z_0 u(n) = \begin{cases} 
F(n) = \sum_{i=0}^{M} w_i \psi_i(n) & \text{if } x(n) > 0, \\
0 & \text{if } x(n) \leq 0, 
\end{cases}
\tag{1}
\]

where \( x \) is fold position and \( \psi_i(n) = \psi_i(x(n), x(n-1)) \). Several choices are possible for the regressor set \( \{ \psi_i \} \), here we use a third order polynomial expansion in \( x(n), x(n-1) \). Lastly, the glottal pressure \( p \) that drives the linear oscillator is related to glottal flow through the input impedance \( Z_{in}(z) \) of the vocal tract. Equivalently, wave variables \( p^\pm = (p \pm Z_0 u)/2 \) can be used: then \( p^\pm \) are related to each other through the vocal tract reflectance \( R(z) \).

A final remark concerns the insertion of a fictitious delay element \( z^{-1} \) (see Fig. 1): this is needed in order to make the model computable. More accurate and elegant techniques are available [10] for dealing with such computability problems. However, we claim that in this case the insertion of \( z^{-1} \) does not deteriorate or anyhow affect properties of the system; the reasons for this are explained in the next section.

IDENTIFICATION

We now address the following problem: given a target glottal flow waveform \( Z_0 \bar{u} \) (be it synthetic or inverse filtered), we want to identify the model so that the output \( Z_0 u \) from the non-linear block fits the target as closely as possible. The problem is not trivial, since we have to identify a physical model of the source rather than a parametric model of the signal [2, 4, 5].

We used both synthetic and real flow waveforms; in both cases a single period was chosen and a periodized signal was constructed and used as target. Synthetic signals were produced using the model described in [11], while inverse filtering was computed using an automatic method described in [12]. 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 utterance analyzed in the present study is a sustained /a/ vowel produced in normal phonation by a male speaker. The system is identified in three main steps (see Fig. 2).

- **Step 1**. From the target \( Z_0 \bar{u} \), the corresponding pressure signal \( \bar{p} \) is computed. In order to do that, we arbitrarily choose the reflectance \( R \) to be that of a uniform vocal tract: \( R(z) = z^{-2m_L} Z_{load}(z) \), where \( m_L \) defines the length (in samples) of the tract and \( Z_{load} \) has a low-pass characteristics. We would like to point out that, in doing this, we are not looking for physical insight: all we want is a pressure signal that provides plausible excitation to the vocal fold.

- **Step 2**. The linear block \( H_r \) is driven using \( \bar{p} \) and its output \( \bar{x} \) is computed. Open- and closed-phase resonances for \( H_r \) are chosen interactively, in such a way that the open and closed phase for the output \( \bar{x} \) match those of the target flow \( \bar{u} \) (see Fig. 2).

- **Step 3**. We now have a complete I/O description of the non-linear block. Then, during the open phase weights \( w = [w_0 \ldots w_M] \) for the regressors \( \psi_i \) are identified by choosing a training
Figure 2: (a) Target $Z_0\bar{u}$ (solid gray, computed from real speech by inverse filtering), synthesized pressure $\bar{p}$ (dashed) after Step 1, output $\bar{x}$ from linear block (dotted) after Step 2, output $Z_0u$ from non-linear block (solid black), after Step 3. (b) Derivative of the target flow $d\bar{u}/dt$ (solid gray) and derivative of output from the non-linear block $du/dt$ (solid black).

window with starting time $l$ and length $m$, and defining the training sets

$$T_u = [\bar{u}(l+1), \bar{u}(l+2), \ldots, \bar{u}(l+m)], \quad T_\psi = \begin{bmatrix} \psi_0(l+1) & \cdots & \psi_0(l+m) \\ \vdots & \ddots & \vdots \\ \psi_M(l+1) & \cdots & \psi_M(l+m) \end{bmatrix}.$$ (2)

From Eq. (1), the weights $w$ must solve the system $w \cdot T_\psi = T_u$. The LS solution of such a system is known [13] to be

$$w = T_u \cdot T_\psi^+.$$ (3)

where the symbol $+$ has the meaning of pseudo-inversion. It is now clear that the insertion of $z^{-1}$ does not deteriorate accuracy of the model: given the structure in Fig. 1 and a flow signal, the identification of the non-linear block automatically takes into account the $z^{-1}$ element.

**RESULTS AND DISCUSSION**

Figure 2 shows that the identification procedure allows for accurate fit of the target. The performance is comparable to that obtained from LF-based fits [2, 4, 5]. In particular from Fig. 2(b) one can see that in the closing phase both the width and the amplitude of the negative pulse in the derivative are well approximated. The large bandwidth (11.025 kHz in this case), and the consequent noise in the flow derivative waveform, does not affect the identification. This is because the identification target is the glottal flow itself, rather than its derivative; also, the LS solution given in Eq. (3) is robust with respect to noise. We note that large bandwidths can deteriorate the performance of typical LF identification techniques [4].

After identification we study the behavior of the model in autonomous evolution. The system is seen to reach steady state self-sustained oscillations after a short transient. The steady state waveform coincides with the one obtained from identification. If we then adjust the values for some of the parameters, the system exhibits robustness to such changes. Fig. 3 shows an example where $f_r$ is increased after the system has reached 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. 3; after transition maximum amplitude turns back to its original value.

A major limitation in our identification procedure is that we are not able to guarantee a priori that the identified system is stable during resynthesis. In order to do that different approaches can be used, such as the harmonic balance technique [14] or the imposition of appropriate conditions on the gradient $\nabla F$ [15], and future work shall concentrate on this issue. Moreover, problems are encountered when strong ripples (due to interaction with tract formants) appear on the opening phase of the target signal: these affect heavily the accuracy of identification. A possible solution to this could be to focus accuracy on the closing phase. This portion is the most important to be fitted accurately, since it defines most of the spectral (and perceptual)
features of the signal. One more open problem is concerned with identification of non-periodic signals: in this case a straightforward strategy is to identify system parameters once for each period of analyzed data (as already done with the LF model [2]). But we can also expect that, with a proper choice of the regressors, the system is able to “follow” a non-periodic signal with a single set of parameters. Finally we remark that our identification procedure is far from being automatic: while the weights \( \{ w_i \} \) are estimated automatically from Eq. (3), the linear filter \( H_r \) is adjusted interactively.

Acknowledgments: this research has been partially funded by the Academy of Finland (“Sound Source Modeling” project).

REFERENCES