MODEL-BASED SYNTHESIS AND TRANSFORMATION OF VOICED SOUNDS

Carlo Drioli

Dipartimento di Elettronica e Informatica
Università degli Studi di Padova
adrian@dei.unipd.it
www.dei.unipd.it/˜adrian

Federico Avanzini

Dipartimento di Elettronica e Informatica
Università degli Studi di Padova
avanzini@dei.unipd.it
www.dei.unipd.it/˜avanzini

ABSTRACT

In this work a glottal model loosely based on the Ishizaka and Flanagan model is proposed, where the number of parameters is drastically reduced. First, the glottal excitation waveform is estimated, together with the vocal tract filter parameters, using inverse filtering techniques. Then the estimated waveform is used in order to identify the nonlinear glottal model, represented by a closed-loop configuration of two blocks: a second order resonant filter, tuned with respect to the signal pitch, and a regressor-based functional, whose coefficients are estimated via nonlinear identification techniques. The results show that an accurate identification of real data can be achieved with less than 10 regressors of the nonlinear functional, and that an intuitive control of fundamental features, such as pitch and intensity, is allowed by acting on the physically informed parameters of the model.

1. INTRODUCTION

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 glottal excitation has been recognized to be a key feature for obtaining high quality synthesis of voice, and for characterizing *voice quality* (e.g. modal voice, vocal fry, breathy voice) [1].

Among the different glottal models, both *parametric* and *physical* models are found. A very well known parametric model is the *LF model* (see, for example, [2]): this characterizes one cycle of the derivative glottal wave by using four parameters. The model has been proved to be very flexible, and able to reproduce a variety of voice qualities [3]. Among the physical models, the first one has been developed by Ishizaka and Flanagan [4]. This accurately reproduces subtle effects that are not taken into account by a parametric model (e.g. interaction with the vocal tract), but as a counterpart cannot be efficiently used for identification and coding, as a large number of physical parameters has to be estimated.

In this work we propose a glottal model loosely based on the Ishizaka and Flanagan model, where the number of parameters is drastically reduced. The model is exploited in an identification scheme: first, the glottal excitation waveform is estimated, together with the vocal tract filter parameters, using inverse filtering techniques. Then the estimated waveform is used in order to identify the nonlinear glottal model, represented by a closed-loop configuration of two blocks: a second order resonant filter, tuned with respect to the signal pitch, and a regressor-based functional, whose coefficients are estimated via nonlinear identification techniques. Results show that the system accurately identifies the estimated waveform, and can be used in order to obtain a good naturalness of the resynthesized voiced sounds. Moreover, the physical parameters of the model can be used to change voice quality without affecting voice identity: for instance, by changing the resonance frequency of the second order oscillator in the model, the pitch of the resynthesized sound can be controlled.

Physically informed models of the vocal emission can be implied in a wide range of applications. Among these are speech synthesis and transformation, voice quality enhancement, speaker identification, voice pathology classification, speech coding and transmission. In particular, the use of a physical model that learns the individual characteristics of a given speaker (voice identity) can be finalized to improve natural speech synthesis. In speech coding, the enhancement of predictive coders by a nonlinear prediction scheme of glottal pulse waveform could significantly improve the speech compression rates.

Sec. 2 introduces the analysis-synthesis model used for the vocal tract and the glottal source; Sec. 3 describes the identification procedure used for estimating the nonlinear regressors in the glottal model. Finally, results from the model are shown in Sec. 4.

2. ANALYSIS-SYNTHESIS MODEL

2.1. Vocal tract and glottal flow estimation

Voiced speech is produced by excitation of the vocal tract system with the quasi-periodic vibrations of the vocal folds at the glottis (voice source). In many speech analysis and modeling approaches, the accurate separation of the voice source from the vocal tract effects is a fundamental task for a correct determination of voice source features, such as the glottal flow waveform, the glottis opening and closing instants, etc. The most common technique relies on a linear prediction coding scheme (LPC), that estimates the vocal tract filter, and on the determination of the source signal (or *residual*) by inverse filtering of the speech pressure signal at mouth. However, since the vocal tract characteristics change within a pitch period because of the opening and closing of the glottis, the determination of the poles of the vocal tract system is often carried on by a covariance LP analysis restricted to the closed glottis period [5]. This method requires a first estimate of the closing glottis instants (CGI), for example the peaks in the residual error of a pitch-asynchronous autocorrelation LP analysis. Then, a pitch-synchronous covariance LP analysis, starting at time instant CGI+1 and limited to the closed-phase, is used to estimate the all-poles vocal tract filter. This filter should model the formant structure of the speech signal, but often the resulting polynomial exhibits poles in excess, that do not contribute to any formant. For this reason, an improved all-poles filter is derived by solving the roots of the original LP polynomial and by discarding the roots corresponding to resonance frequencies below 250 Hz, and the ones with bandwidth above 500 Hz. The inverse of this improved filter is used in turn to obtain the derivative of the desired glottal volume velocity waveform (note that the lip radiation effect can be modeled with a differentiator filter). The estimate of the desired glottal volume velocity waveform can thus be obtained by integrating the residual waveform. Usually, since the first CGI estimate is not always accurate, a small number of covariance LP analysis centered around the CGI estimate is performed, and the best result is selected on the basis of the residual characteristics [6].

The procedure described above is used to build a time-varying model of the vocal tract and to extract the glottal volume velocity waveform from a voiced pressure waveform. Fig. 1 shows a typical glottal flow waveform obtained with this method. The remaining of the section will focus on the development of a suitable glottis model to represent this excitation signal.

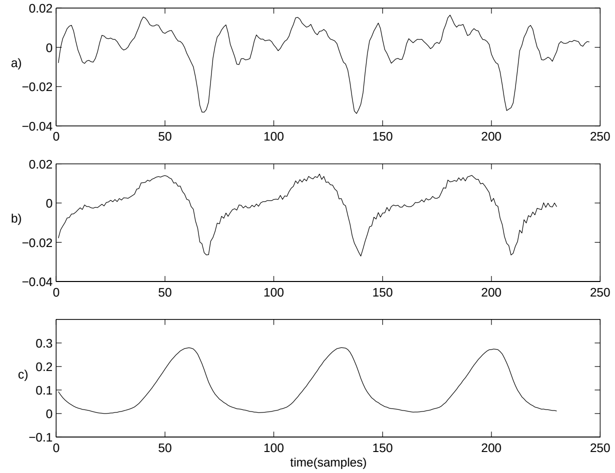

Figure 1: *Inverse filtering procedure for glottal flow waveform estimation. Plot a): speech waveform (pressure at lips); plot b): differentiated glottal flow, derived from pressure waveform by inverse filtering; plot c): glottal flow*

2.2. The glottis model

The glottis model here proposed is loosely based on the well known model by Ishizaka and Flanagan (IF in the following), well described in [4]. The IF model is made of two distinct *functional blocks*. The first one is a quasi-linear block with memory, representing the mechanics of the vocal fold: it is made of two weakly nonlinear oscillators, coupled with each other and driven by the pressure drop at the glottis. The resonance frequencies of the oscillators determine some significant features of the glottal signal, such as the pitch and the OP/CP (Open Phase-Closed Phase ratio). The second block is highly nonlinear and models the fluid dynamics at the glottis: the flow u_g is assumed to depend on the lung pressure p_s and on the vocal fold displacement x, and the analytical expression for the nonlinearity is derived from general theory of fluid dynamics. The two blocks are therefore coupled in a feedback loop.

Many refinements have been proposed to this model (see for instance [7], where an 8-mass model for the vocal fold is used). Here we choose to follow a different direction, in which the IF model is taken as a reference for developing a *physically informed* model rather than a true physical model: the decomposition in two coupled functional blocks is kept as in IF, but internal structures are drastically simplified. For the first block we take the simplest oscillating system, i.e. a second order filter with transfer function

$$H_{res}(z) = \beta_0/(1 + \alpha_1 z^{-1} + \alpha_2 z^{-2}).$$

Therefore it is completely described by its resonance frequency ω_0 and its 3-dB bandwidth $\Delta\omega$. In this way we simplify the IF model in the sense that we describe the vocal fold as a single mass-spring system with damping; the output $x_1(k)$ of H_{res} is related to vocal fold displacement, and the difference $(x_1(k) - x_1(k-1))$ is related to vocal fold velocity. The second block is a nonlinear instantaneous map f that has p_s and the state of H_{res} as inputs and returns the flow u_g; this is taken as a driving signal for the oscillator H_{res}. Therefore the feedback from map f models the interaction between fluid dynamics at the glottis and vocal fold mechanics. The final structure of the model is the one depicted in Fig. 2

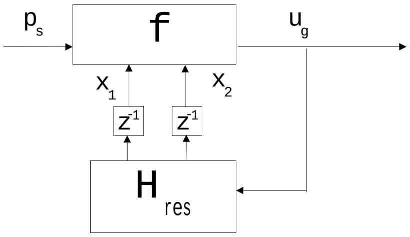

Figure 2: *Physically informed model of the glottis. Here, H_{res} is a second-order resonant filter, tuned on the pitch of the voiced signal, $f(x_1, x_2, p_s)$ is a nonlinear function, u_g is the glottal flow signal, x_1 and x_2 are respectively the output and the first state variable of the second-order filter, and p_s is the lung pressure.*

3. NONLINEAR IDENTIFICATION

3.1. The set of regressors

The nonlinear map f is modeled with a regressor-based functional of the form

$$f(x_1, x_2, p_s) = w_0 + \sum_{i=1}^{M} w_i \psi_i(x_1, x_2, p_s) \qquad (1)$$

where the w_i are weights to be identified, and $\psi_i(x_1, x_2, p_s)$ are the regressors of the input data. The choice of the regressors can be made in several ways. Often local models, such as gaussian functions or any other radial basis function, are used. This approach leads to a model called *Radial Basis Function Network* (RBFN) [8], used in the field of time series analysis and modeling. The use of a polynomial expansion of the input leads to a class of NARMAX models [9], known in the fields of system identification and control. Alternatively, the regressors can be derived on the basis of physical considerations. We will follow here this last approach, and derive the definition of the regressors set from the (simplified) relation between the sub-glottal pressure p_s, the average glottal volume velocity u_g and the cross-sectional area of the glottis A_g in a single-mass glottis model [10]:

$$p_s = \frac{k\rho}{2A_g^2}u_g^2 + \frac{12\eta l^2 d}{A_g^3}u_g + Z_0 u_g \qquad (2)$$

where ρ and η are the density and viscosity coefficient of air, l and d are, respectively, the length and thickness of glottis, k is an empirical constant, and Z_0 is the input impedance of the vocal tract. In the coupling with the mechanical system modeled with a mass and a spring, the cross-sectional area is computed as $A_g = 2d \cdot x_1$, where x_1 is the vocal fold displacement. The solution of Eq. (2) is of the form

$$u_g = \frac{-b}{2} \pm \frac{\sqrt{b^2 - 4c}}{2} \qquad (3)$$

with

$$b = \frac{(12\eta l^2 d - Z_0)(8d^2 x_1^2 Z_0^2)}{8d^3 x_1^3 Z_0 k\rho} \qquad (4)$$

and

$$c = \frac{-2p_s 4d^2 x_1^2 Z_0^2}{k\rho} \qquad (5)$$

As for single-reed models for wind instruments [11], an additional contribute to the total flow is given by vocal fold velocity; this can be expressed as $S_r \cdot \dot{x}_1$, where S_r is an effective surface. As already mentioned in Sec. **??**, taking into account $\dot{x}_1$ corresponds, in discrete time, to taking into account one past value of x_1; therefore, in Eq. (3) we add the term $x_1 \cdot x_2$, where $x_2(k) = x_1(k-1)/\beta_0$ is the first state variable of H_{res}. The emphasis on this term is justified by the fact that it has been experimentally observed to noticeably improve the identification process. Qualitatively, this term can be explained assuming that the effective surface S_r is not constant during the vocal folds motion, but is instead proportional to the displacement x_1.

We now want to write u_g in terms of the functional (1), that is $u_g = f(x_1, x_2, p_s)$. We use a Taylor expansion up to the second order to write u_g as a sum of regressors of x_1 and $x_2 = \dot{x}_1$. We then select the terms giving better results in the identification process, that are:

$$\begin{aligned} \psi_1(x_1,x_2,p_s) &= x_1 & \psi_5(x_1,x_2,p_s) &= 1/x_1 \\ \psi_2(x_1,x_2,p_s) &= x_1^2 & \psi_6(x_1,x_2,p_s) &= 1/x_1^3 \\ \psi_3(x_1,x_2,p_s) &= x_1^3 & \psi_7(x_1,x_2,p_s) &= x_1^3 p_s \\ \psi_4(x_1,x_2,p_s) &= x_1^5 & \psi_8(x_1,x_2,p_s) &= x_1 x_2 \end{aligned} \qquad (6)$$

3.2. Identification procedure

The identification procedure is finalized at computing the right set of coefficients w_i in order to let the system reproduce a desired glottal flow signal (the *target* signal). We will show here the method with respect to a periodic steady-state signal frame $u_g(k)$, of length M, generated as the repetition of a selected period of the desired waveform. This training sequence is especially suited to find the coefficients w_i that lead to a system able to maintain a stable oscillatory motion with a waveform period of the desired shape.

The identification procedure relies on two main steps. First, the resonant filter H_{res} is estimated; to this end, the resonance frequency ω_0 is chosen in order to match the Open Phase frequency (this is qualitatively in agreement to what happens in the IF model). The bandwidth $\Delta\omega$ is chosen so that the quality factor Q of H_{res} matches a reference value Q_0 deduced from IF parameters: this is found to be $Q_0 = 10$.

Once H_{res} has been estimated, the second step is the determination of the input and output temporal sequences of the nonlinear block f. From Fig. 2, it can be seen that the output sequence is just $u_g(k), k = n+1, \ldots, n+M$, whereas $x_1(k) = h_{res} * u_g(k-1)$, $k = n+1, \ldots, n+M$, with n and M being respectively the starting time and the length of the training data window. It can be seen form Fig. 3 that n is taken large enough to skip the transient of the filter H_{res}, and M is chosen large enough to have a few periods in the training set. Finally, if the implementation of the second order IIR filter is a *canonic direct form*, then the second input sequence is $x_2(k) = x_1(k-1)/\beta_0$.

Let us now define $\psi_i(k) = \psi_i(x_1(k), x_2(k), p_s(k))$, i.e. $\psi_i(k)$ is the i-th regressor at the discrete time k. It is now straightforward to build the training data sets $\mathbf{T}_{u_g}$ and $\mathbf{T}_x$ from the input and output sequences:

$$\mathbf{T}_{u_g} = [u_g(n+1), u_g(n+2), \ldots, u_g(n+M)] \qquad (7)$$

$$\mathbf{T}_x = \begin{bmatrix} \psi_1(n) & \cdots & \psi_1(n+M-1) \\ \vdots & \ddots & \vdots \\ \psi_8(n) & \cdots & \psi_8(n+M-1) \end{bmatrix}. \qquad (8)$$

The identification of the set of coefficients $\mathbf{w} = [w_0, w_1, \ldots, w_7]$, requires the solution of the matrix system

$$\mathbf{w}\begin{bmatrix} \mathbf{1} \\ \mathbf{T}_x \end{bmatrix} = \mathbf{T}_{u_g}, \qquad (9)$$

where $\mathbf{1} = [1, \ldots, 1]$ is a row vector of length M. The LS solution of problem (9) is known to be

$$\mathbf{w} = \mathbf{T}_{u_g}\begin{bmatrix} \mathbf{1} \\ \mathbf{T}_x \end{bmatrix}^+ \qquad (10)$$

where the symbol $+$ has the meaning of *pseudo-inversion* of a matrix. In Fig. 3 the periodic waveform, the training sequences, and the result of the identification are shown.

It has to be noticed that in order to let the model reproduce different behaviors, such as transition sequences, amplitude-varying sequences, etc., the model has to be trained with target data sets that present the desired behavior.

4. RESULTS AND DISCUSSION

Due to its limited number of parameters and regressors, and to its structural simplicity, the proposed model can be efficiently used for identification of target waveforms.

First of all, good results are found in the *open loop* configuration: in this configuration the linear block H_{res} is forced by the target signal u_g and the identified glottal waveform is observed at the nonlinear block output, as shown in Fig. 3. It can be seen

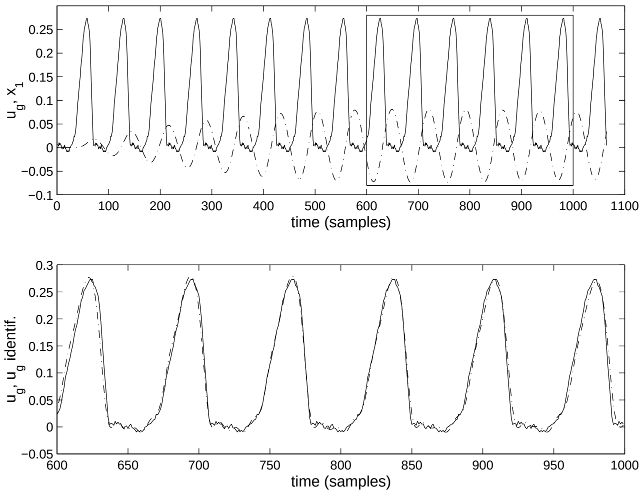

Figure 3: *Identification procedure. Upper plot: target periodic glottal flow (solid line), output of the resonant filter from the target periodic glottal flow (dashed line), and the training data window. Lower plot: identified glottal waveform (dashed line) and target training waveform (solid line).*

that eight polynomial regressors (see Eq. (6)) allow to achieve an accurate reconstruction of the target glottal flow waveform.

Even more interesting results are obtained in the *closed loop* configuration, i.e. the configuration shown in Fig. 2; in this case the feedback loop between the two functional blocks is closed after the oscillation has reached steady state, and then the system evolves autonomously. In this configuration the system appears to be stable and to maintain the target waveform, and can therefore be used for resynthesis of the analyzed voiced sounds. The behavior of the model in the closed loop configuration is shown in Fig. 4 (from sample 600 to sample 1500).

The physical information contained in the model can be exploited in order to introduce modifications in the resynthesized sounds. An example is given by the resonance frequency ω_0 of H_{res}. As already mentioned, in the IF model it is related to the pitch of the glottal excitation signal; time simulations give evidence that this characteristic is preserved in our model. The second row of Fig. 4 shows the behavior of the system when ω_0 is gradually increased from its initial value. It can be seen that the pitch increases correspondingly, and that the glottal signal shape is preserved with good accuracy.

5. REFERENCES

[1] D.G. Childers and C.K. Lee, “Vocal quality factors: analysis, synthesis, and perception,” *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, November 1991.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker recognition,” *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 5, pp. 569–586, September 1999.

[3] D. G. Childers and C. Ahn, “Modeling the Glottal Volume-Velocity Waveform for Three Voice Types,” *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 505–519, Jan. 1995.

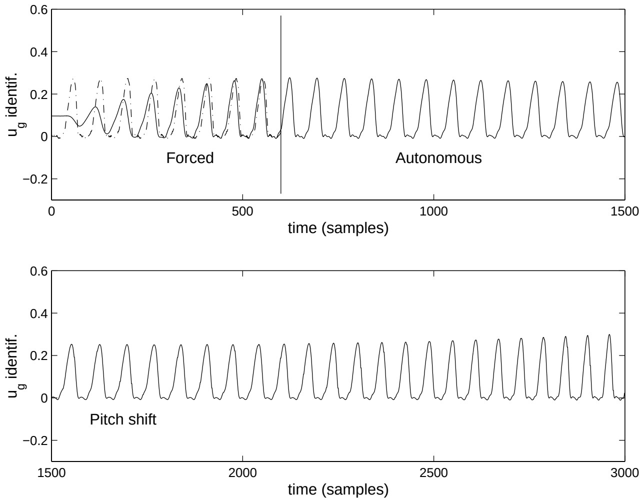

Figure 4: *Pitch transformation: synthesized flow (solid line) and target flow (dashed line). Up to sample* 600, *the system is forced to reach the steady state by feeding the resonant filter* H_{res} *with the target* u_g *(open loop configuration). From sample* 601 *to sample* 1500, *the system evolves autonomously (closed loop configuration). From sample* 1501 *to sample* 3000, *the resonance frequency* ω_0 *of the filter* H_{res} *gradually rises from* $2\pi \cdot 150$ *rad/s to* $2\pi \cdot 200$ *rad/s.*

[4] K. Ishizaka and J. L. Flanagan, “Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords,” *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.

[5] D.Y. Wong, J.D. Markel, and A. H. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, August 1979.

[6] R. Smits and B. Yegnanarayana, “Determination of instants of significant excitation in speech using group delay function,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 5, pp. 325–333, September 1995.

[7] M. Kob, N. Alhäuser, U. Reiter, “Time-domain model of the singing voice,” *Proc. of DAFx99 Workshop*, pp.143–146, Norway, Dec. 1999.

[8] S. Chen, C. F. N Cowan, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis functions networks,” *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 302–309, March 1991.

[9] S. Chen and S. A. Billings, “Representation of non-linear systems: Narmax model,” *Int. J. of Control*, vol. 49, no. 3, pp. 1013–1032, 1989.

[10] X. Rodet, “One and two mass models oscillations for voice and instruments,” *Proc. Int. Computer Music Conf.*, Canada, Sept. 1995.

[11] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, “On the oscillation of musical instruments,” *J. Acoust. Soc. Am.*, vol. 74, no. 5, pp. 1325–1345, 1983.