Variational Integration for Speech Signal Processing

Max A. Little[1] (littlem@maths.ox.ac.uk), **Irene M. Moroz**[1], **Patrick E. McSharry**[1,2], **Stephen J. Roberts**[2]
[1]Mathematical Institute, Oxford University, UK, [2]Engineering Science, Oxford University, UK

1 Introduction

Voice production is generally modelled as a two-component dynamical process composed of the vocal folds and vocal tract. Figure 1 shows a diagram of the arrangement of the vocal fold and vocal tract inside the head and neck. The vocal tract is comprised of the pharyngeal, oral and nasal cavities. It is usually modelled as a linear acoustic resonator, and the vocal folds as a nonlinear dynamical system comprising masses, viscoelastic damping and forcing due to lung pressure. However, it is generally the case that systems used for speech transmission, analysis or compression do not utilise an explicit, dynamical model of the vocal folds. They take several different approaches (Kleijn & Paliwal 1995), including: (a) *waveform coders* with no model of speech production, (b) *source coders* which use a vocal tract model and a simple characterisation of the vocal fold behaviour, i.e. whether it is periodic or noise-like, and (c) *hybrid methods* with a vocal tract model and selection of a representation of the vocal fold behaviour that minimises overall waveform error.

In this paper we introduce a method for modelling the dynamical behaviour of the vocal folds in speech processing. This method is based around a discrete dynamical model that is suitable for direct fitting to the vocal fold signal. Thus parameters that represent the biomechanical behaviour of the vocal folds can be identified. These parameters, together with the initial conditions and the model residual are an exact but smaller representation, in the information-theoretic sense, of the vocal fold dynamics. This representation could then, for example, form the basis for a low bit-rate source coder.

2 Continuous-Time Vocal Tract and Fold Models

The vocal tract is generally modelled as an acoustic resonator with varying cross-sectional area (Markel & Gray 1976), for which the governing linear equation is the second-order *Webster's horn equation*, here written in terms of acoustic flow rate $U(y,t)$:

$$\frac{\partial^2}{\partial y^2}U\left(y,t\right)-\frac{1}{S\left(y\right)}\frac{\partial}{\partial y}S\left(y\right)\frac{\partial}{\partial y}U\left(y,t\right)-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}U\left(y,t\right)=0, \tag{2.1}$$

where y is the position along the axis of the vocal tract tube, $S(y)$ is the cross-sectional area of the tube and $U(y,t)$ is the air flow rate.

There are a variety of models of vocal fold behaviour, ranging from complete, continuum models of vocal fold tissue and air flow to "lumped" models which make various simplifying assumptions to reduce the dynamical order of the system (for example the two-mass model of (Steinecke & Herzel 1995)). The vocal fold tissue is modelled as a mass with an internal viscoelastic spring. We consider only low-dimensional models, specifically a simplified, two-dimensional system that exhibits self-sustained oscillation, mimicking the energy balance between aerodynamic forcing and viscoelastic damping. The elasticity of the vocal folds is different when the folds are closed (in collision), and when open. Thus the potential energy due to a discontinuous version of Hooke's law, is:

$$V\left(x\right)=\frac{1}{2}\left[k_1\Phi\left(x\right)+k_2\Phi\left(-x\right)\right]x^2, \tag{2.2}$$

where the Heaviside step function Φ is used to differentiate between non-collision and collision conditions, and k_1, k_2 are the spring constants during collision and non-collision conditions, respectively. When x is negative, the folds are in compression, and there is no airflow. The kinetic energy is $T\left(\dot{x}\right)=\frac{1}{2}m\dot{x}^2$, and the *Lagrangian* of the model is $L(x,\dot{x})=T(\dot{x})-V(x)$. The airflow rate U is proportional to the width of the vocal fold opening:

$$U\left(x\right)=A\Phi\left(x\right)x, \tag{2.3}$$

where A is an amplitude parameter that is proportional to the static lung pressure, and the flow rate is zero when the vocal folds are closed. The *Hamiltonian* of the model represents the amount of energy in the system:

$$H\left(x,\dot{x}\right)=\dot{x}\frac{\partial}{\partial\dot{x}}L\left(x,\dot{x}\right)-L\left(x,\dot{x}\right)=\frac{1}{2}m\dot{x}^2+\frac{1}{2}\left(k_1\Phi\left(x\right)+k_2\Phi\left(-x\right)\right)x. \tag{2.4}$$

Using Hamilton's *variational principle* we arrive at the *Euler-Lagrange* equation:

$$\frac{\partial}{\partial t}\left(\frac{\partial}{\partial\dot{x}}L\left(x,\dot{x}\right)\right)-\frac{\partial}{\partial x}L\left(x,\dot{x}\right)=0 \tag{2.5}$$

from which we can obtain the equations of motion for the vocal fold motion:

$$m\ddot{x}-\left(k_1\Phi\left(x\right)+k_2\Phi\left(-x\right)\right)x=0. \tag{2.6}$$

3 The Combined Speech Model

By taking the Fourier transform of equation (2.1), we can find a relationship between the pressure at the lips to the flow rate at the exit of the vocal folds in terms of a radian frequency ω. We model the acoustic radiation effect at the lips as a piston in an infinite plane baffle (Markel & Gray 1976), and assume that there is no acoustic feedback of the vocal tract upon the vocal folds. Hence the vocal folds, placed at the bottom of the vocal tract, forces the air column in the vocal tract into oscillation at specific resonances. The acoustic pressure at the lips $P(\omega)$ can then be written as:

$$P(\omega)=Z(\omega)U(\omega)H(\omega), \tag{3.1}$$

where Z is the acoustic radiation impedance of the lip opening, H is the transfer function of the vocal tract, and U is the flow rate at the exit of the vocal folds, all expressed in terms of the radian frequency of oscillation ω. In speech recordings we generally have access to a discrete-time pressure signal $p(n)$, where $n \in \mathbb{Z}$ is the discrete time index. We wish to estimate the discrete-time vocal fold flow rate signal $u(n)$.

4 Quasi-Linear Prediction

We define a discrete-time, *quasi-linear model* as an inhomogeneous, nonlinear difference equation of the form:

$$v(n+1) = \sum_{i=0}^{Q-1} a_i f_i \left(v(n), v(n-1), \ldots v(n-D+1)\right) + e(n) \tag{4.1}$$

where D is *state-space dimension* of the model, $v(n-j), j = 0, 1, \ldots D-1$ are the state variables, a_i are the Q parameters, and $e(n)$ is the inhomogeneous driving component considered as an external input. The model is quasi-linear because the f_i are arbitrary functions of the state-variables only, i.e. they contain no arbitrary parameters.

A discrete-time signal $s(n)$ can be *time-delay embedded* (H Kantz, T Schreiber 1997) in a D-dimensional state-space, here with embedding time delay $\tau = 1$. Then we can use the model of equation (4.1) to *predict* $s(n+1)$ given the previous D samples of the signal:

$$s\left(n+1\right) = \sum_{i=0}^{Q-1} a_i f_i \left(s\left(n\right), s\left(n-1\right), \ldots s\left(n-D+1\right)\right) + e\left(n\right). \tag{4.2}$$

The term $e(n)$ is now considered as a "residual" or error. While the model functions f_i together with the parameters a_i, the initial conditions $s(0), s(1), \ldots s(D-1)$ and the residual $e(n)$ are sufficient to reconstruct the original signal $s(n)$ *exactly*, if the model is a good representation of the structure in the signal, the residual can be very small in magnitude. We can therefore obtain a *compressed representation* of the signal in the information theoretic sense by finding the values of a_i that make the residual as small as possible. If we assume that $e(n)$ is a Gaussian, zero-mean, stationary stochastic process, we are lead to the least-squares, maximum-likelihood solution, resulting in the following simultaneous system to be solved for each a_i:

$$\Phi_{0j} = \sum_{k=0}^{Q-1} a_k \Phi_{jk}, \quad j = 0, 1 \ldots Q-1, \tag{4.3}$$

where $\Phi_{jk} = \sum_{n=0}^{N-1} f_j f_k$ for a signal of length N, and $f_0 = s(n+1)$. It is possible to reduce Q by the sequential removal of functions whilst monitoring some measure of the quality of the fit that penalises over-complex models (H Kantz, T Schreiber 1997).

5 Discrete-Time Vocal Tract Model

By assuming that the vocal tract can be modelled as a concatenation of lossless, acoustic tubes, where wave energy is preserved at the junctions between the tubes, we can discretise equation (2.1) as a *digital waveguide network* (Bilbao 2004). This can be shown to be equivalent to a *lattice* implementation of a general, infinite impulse response, linear digital filter. Thus, for the task of estimating the resonances of the vocal tract, we can utilise a simplification of the model formulation of equation (4.1):

$$f_i = v\left(n-i\right), \quad i = 0, 1 \ldots Q-1 \tag{5.1}$$

which has the same state-space dimension as number of parameters, i.e. $D = Q$.

6 Discrete-Time Variational Integrator for Vocal Fold Model

Given a continuous time model of the vocal folds, we must produce a discrete version such that the parameters can be identified from the signal. However, a simple finite difference discretisation (such as a Runge-Kutta method) will introduce *artificial energy loss*. If instead the vocal fold model is discretised using a *varational integrator* (Marsden & West 2001), the long-time energy behaviour of the system is preserved. We perturb the position of two discrete, neighbouring points in time and find the minimum of the *Hamiltonian action functional*:

$$\frac{\partial}{\partial \varepsilon} F\left[L_d\left(x_\varepsilon\left(n\right), x_\varepsilon\left(n+1\right)\right)\right] = \frac{\partial}{\partial \varepsilon} \sum_{n=0}^{N-1} L_d\left(x_\varepsilon\left(n\right), x_\varepsilon\left(n+1\right)\right) = 0 \tag{6.1}$$

when the small perturbation parameter, ϵ, tends to zero. Here L_d is the *discrete Lagrangian* defined at the points $x(n)$ and $x(n+1)$ (rather than defined by position and velocity). We then arrive at the *discrete Euler-Lagrange* equation (Marsden & West 2001):

$$\frac{\partial}{\partial x\left(n\right)} L_d\left(x\left(n\right), x\left(n+1\right)\right) + \frac{\partial}{\partial x\left(n\right)} L_d\left(x\left(n-1\right), x\left(n\right)\right) = 0. \tag{6.2}$$

We now replace the velocity by a typical finite difference approximation $\dot{x}\left(t\right) \approx \left[x\left(n+1\right) - x\left(n\right)\right]/\Delta t$ so that the Lagrangian is discretised for $n = 0, 1 \ldots N-1$ with a sample time interval of Δt:

$$L_d\left(x\left(n\right), \frac{x\left(n+1\right) - x\left(n\right)}{\Delta t}\right) = \frac{1}{2} m \left(\frac{x\left(n+1\right) - x\left(n\right)}{\Delta t}\right)^2 - \frac{1}{2}\left[k_1 \Phi\left(x\left(n\right)\right) + k_2 \Phi\left(-x\left(n\right)\right)\right] x\left(n\right)^2 \tag{6.3}$$

Using equation (6.2), we arrive at the discrete Newtonian equations of motion for $n = 0, 1 \ldots N-1$:

$$x(n+1) - 2x(n) + x(n-1) + \frac{\Delta t^2}{m}\left[k_1\Phi(x(n)) + k_2\Phi(-x(n))\right]x(n) = 0 \tag{6.4}$$

which is a quasi-linear model whose most likely parameters can be estimated using the formalism of equation (4.3). Furthermore, since the model is *conservative*, there is a corresponding *discrete Hamiltonian* which is equivalent to the energy in the system:

$$HE(n) = \frac{1}{2}m\left(\frac{x(n+1) - x(n)}{\Delta t}\right)^2 + \frac{1}{2}\left[k_1\Phi(x(n)) + k_2\Phi(-x(n))\right]x(n)^2 \tag{6.5}$$

7 Relationship to the Teager Energy Operator

The expression for discrete energy (6.5) has similarities with that of the discrete *Teager energy operator* (Kaiser 1990):

$$HE(n) \propto x(n)^2 - x(n+1)\,x(n-1)\,. \tag{7.1}$$

For example, for a simple harmonic oscillator, the discrete energy expression, derived using the framework above is:

$$HE(n) \propto x(n)^2 - x(n)\,x(n+1) + \frac{m}{2k\Delta t^2}x(n+1)^2\,. \tag{7.2}$$

Both expressions are local, nonlinear estimates of the total energy in the signal. They differ in that whereas for the above model we derived an expression for conserved energy from the discrete Lagrangian, the Teager energy operator is derived by inserting the exact solution to the continuous, linear model of simple harmonic oscillation directly into the corresponding Hamiltonian expression, and, using certain trigonometric identities and some suitable approximations, solving for the three unknown parameters using three adjacent samples from the time series. However, we require there to be an exact, explicit solution to the system. In most nonlinear problems, exact solutions do not exist. However, using the variational framework derived above, we can obtain arbitrarily accurate estimators of the nonlinear *first integrals* such as energy without the existence of explicit solutions.

8 Linear Signal Factorisation

Having discretised the vocal tract and vocal fold models, we can apply the z-transform to obtain a discrete version of the source-filter equation (3.1):

$$P(z) = Z(z)\,U(z)\,H(z)\,. \tag{8.1}$$

If we assume that the pressure signal $P(z)$ can be factored in the z-domain into a product of linear part $L(z)$ and stochastic part $E(z)$, then we can use the linear prediction formulation of equation (4.3):

$$P(z) = E(z) \times L(z) = E(z) \times G(z)\,Z(z)\,H(z)\,, \tag{8.2}$$

where $G(z)$ is the linearity of the vocal fold oscillation:

$$U(z) = E(z)\,G(z)\,. \tag{8.3}$$

Hence, given the transfer function $G(z)$, we can obtain from the residual $e(n)$ of the linear prediction process, an estimate for $u(n)$ by convolving $e(n)$ with the impulse response of $G(z)$. In this study $G(z)$ is taken to be a second-order, low-pass filter. We should note that the assumption that the driving signal $E(z)$ is Gaussian, required for the least-squares formulation of the linear prediction, is not exactly satisfied in reality. However, it is accurate enough to give reasonable results.

9 Nonlinear Prediction of Vocal Fold Dynamics

Given an estimate $\hat{u}(n)$ for the vocal fold flow rate signal, we can now use quasi-linear prediction to fit the parameters of the vocal fold model. However, there are certain details that need to be resolved.

Firstly, when the vocal folds are open ($x(n) > 0$) or closed ($x(n) < 0$), the model behaves as a simple harmonic oscillator, whose frequency depends only upon the mass, stiffness and time interval parameters m, k_1 (or k_2) and Δt. However, the amplitude of oscillation depends upon the initial conditions, which are required for perfect reconstruction of the signal. Thus, in order to estimate the amplitude parameter A in equation (2.3), we must assume that the initial conditions are normalised.

Secondly, there is a model parameter ambiguity. The ratios $\Delta t^2 k_1/m$ and $\Delta t^2 k_2/m$ are the single parameters that are estimated from the signal $u(n)$. We know Δt and so we must find k_1, k_2 and m. This can be accomplished by the use of the energy expression $HE(n)$: we know that energy is conserved in the system. It is then only necessary to find the best fit values of k_1, k_2 and m such that $HE(n)$ is as constant as possible.

Thirdly, we can only estimate the parameters when the vocal folds are open: when the folds are closed the signal is zero. Also, the signal $u(n)$ will have all DC components removed, due to unavoidable filtering effects of the recording process. Hence the closed regions always appear in the negative part of the signal, and we must find where the derivative of the flow rate signal is very small and the signal is negative.

Figures 2(a)-(d) show all the intermediate stages in fitting the nonlinear vocal fold model. The pressure signal $p(n)$ is of a vowel sound /aa/, extracted from the TIMIT database (Fisher, Doddington & Goudie-Marshall 1986). In the final stage, figure 2(d), we can see that the energy $HE(n)$ applied to the vocal fold model signal $x(n)$ is indeed constant as expected. However, applying $HE(n)$ to $\hat{u}(n)$, we can track the disparity in energy between the model and the actual signal, and so we can see how external energy is added to or removed from the system over the duration of one cycle.

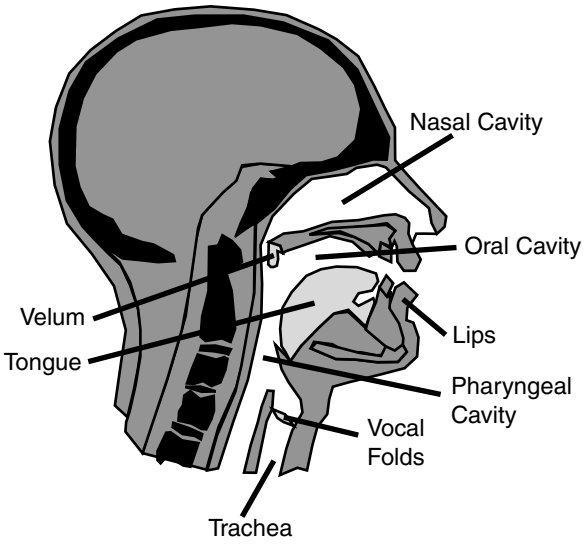

Figure 1: Arrangement of vocal organs inside the head and neck.

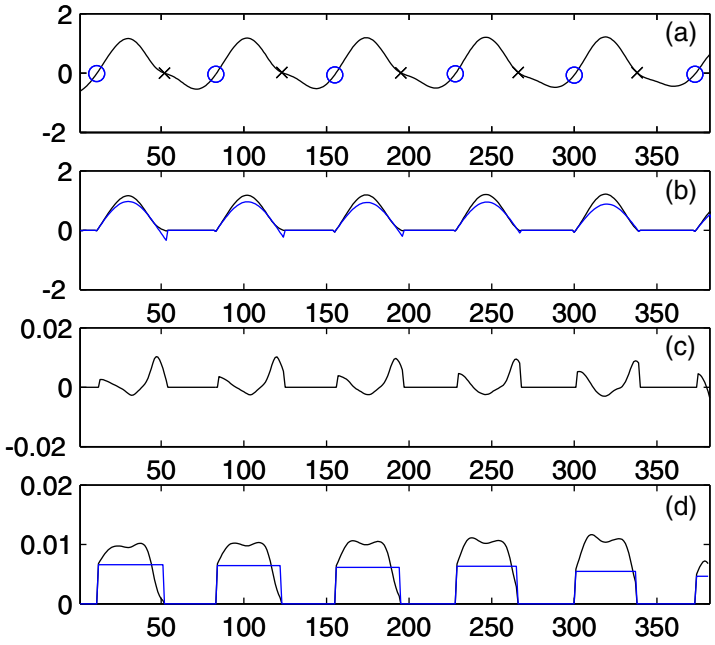

Figure 2: (a) Vocal fold flow rate signal $u(n)$ with open events (blue circles) and close events (black crosses). (b) Open phase flow rate signal only (black) with output signal $x(n)$ from the best fit model (blue). (c) Residual signal $e(n)$. (d) Energy estimate $HE(n)$ for vocal fold flow rate signal (black), energy estimate $HE(n)$ for best fit model $x(n)$ (blue).

10 Conclusions and Future Work

In this paper we have shown a complete, discrete model of voice production that comprises both vocal tract and vocal folds. The continuous vocal tract model was integrated using a waveguide analogy, and hence converted to a standard linear filter. The vocal fold model was integrated using a variational integrator that created a model faithfully reproducing the energy conservation property of the model. We then commented on the similarity between this and the Teager energy operator. We have shown how to find the parameters of the model using the explicit optimisation method of quasi-linear prediction applied in two stages: once to estimate the parameters of the vocal tract resonance, and once to find the parameters of vocal fold oscillation. It was shown that the estimated vocal fold flow rate signal $u(n)$ can be segmented into open and closed vocal fold regions, and parameters fitted to the open regions. Using these parameters and a discrete energy expression, an estimate for the energy in the system was obtained. Finally, we point out that in most cases the model for the vocal tract signal provides a reasonable fit to the data. Therefore, it should be possible in principle to avoid the transmission of the signal residual $e(n)$. Thus we could obtain a very low bit rate speech coder.

For future work, it might be necessary to increase the complexity of the vocal fold model to handle dissipative dynamics as well as forcing due to turbulent noise.

11 Acknowledgements

Max Little would like to thank the Engineering and Physical Sciences Research Council (EPSRC), UK, for the doctoral training grant.

References

Bilbao, S. (2004), *Wave and Scattering Methods for Numerical Simulation*, Wiley.

Fisher, W., Doddington, G. & Goudie-Marshall, K. (1986), 'The DARPA Speech Recognition Research Database: Specifications and Status', *Proceedings of the DARPA Workshop on Speech Recognition* pp. 93–99.

H Kantz, T Schreiber (1997), *Nonlinear Time Series Analysis*, Cambridge University Press.

Kaiser, J. (1990), On a Simple Algorithm to Calculate the 'Energy' of a Signal, *in* 'Proceedings of the IEEE ICASSP', Vol. 1, pp. 381–384.

Kleijn, K. & Paliwal, K. (1995), *Speech Coding and Synthesis*, Elsevier Science, Amsterdam.

Markel, J. D. & Gray, A. H. (1976), *Linear Prediction of Speech*, Springer-Verlag, Berlin.

Marsden, J. & West, M. (2001), 'Discrete Mechanics and Variational Integrators', *Acta Numerica* pp. 357–514.

Steinecke, I. & Herzel, H. (1995), 'Bifurcations in an Asymmetric Vocal-Fold Model', *J. Acoust. Soc. Am.* **97**(3), 1874–1884.