A Simple, Quasi-linear, Discrete Model of Vocal Fold Dynamics

Max Little[1,⋆], Patrick McSharry[2], Irene Moroz[1], and Stephen Roberts[3]

[1] Applied Dynamical Systems Research Group,
Oxford Centre for Industrial and Applied Mathematics, Oxford University, UK
littlem@maths.ox.ac.uk
http://www.maths.ox.ac.uk/ads
[2] Oxford Centre for Industrial and Applied Mathematics, Oxford University, UK
[3] Pattern Analysis Research Group, Engineering Science, Oxford University, UK

Abstract. In current speech technology, linear prediction dominates. The linear vocal tract model is well justified biomechanically, and linear prediction is a simple and well understood signal processing task. However, it has been established that, in voiced sounds, the vocal folds exhibit a high degree of nonlinearity. Hence there exists the need for an approach to modelling the behaviour of the vocal folds. This paper presents a simple, nonlinear, biophysical vocal fold model. A complementary discrete model is derived that reflects accurately the energy dynamics in the continuous model. This model can be implemented easily on standard digital signal processing hardware, and it is formulated in such a way that a simple form of nonlinear prediction can be carried out on vocal fold signals. This model could be of utility in many speech technological applications where low computational complexity synthesis and analysis of vocal fold dynamics is required.

1 Introduction

The linear signal processing of speech is a well developed science, having a long history of association with the science of *linear acoustics*. Referring to Fig. 1, the use of linear acoustics is well justified biophysically, since a realistic representation of the vocal organs is obtained by assuming that the influence of the vocal tract is that of an *acoustic tube* that acts as a *linear resonator*, amplifying or attenuating the harmonic components of the vocal folds during voiced sounds. This resonator can be represented in discrete-time as a *digital filter* [1].

Access to biophysical speech parameters enables certain technology such as communications (e.g. wireless mobile telephone systems), clinical, therapeutic and creative manipulation for multimedia. For example, *linear prediction* [2] can be used to find vocal tract parameters: thus much effort has been directed towards the application of this particular analysis tool to speech signals. The results of such work are an efficient set of techniques for linear prediction of speech

⋆ Supported by doctoral training grant provided by the Engineering and Physical Research Council, UK.

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 348–356, 2005.

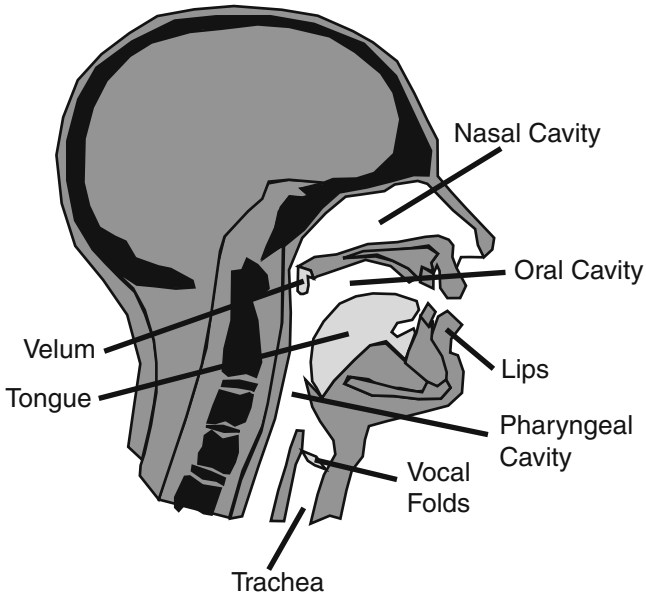

Fig. 1. Arrangement of vocal organs inside the head and neck

that now form a fundamental part of current technology [3]. However, empirical and numerical investigation and modelling of the vocal folds has revealed a high degree of *nonlinearity*, therefore the use of the same linear tools is inappropriate. Hence there exists a need for a similar approach to modelling the behaviour of the vocal fold oscillation.

For typical technological applications, where computational power is at a premium *lumped* models are to be preferred over full, continuum mechanical models because they capture only the important, relevant dynamical effects. Progress on such lumped models has been steady [4], and there exist a range of models of varying complexity (see, for example, the two mass models of [5] and [6]).

This paper presents a simple, practical, continuous model of vocal fold behaviour, and an appropriate discrete counterpart, as described in [7]. The model has only five parameters and can be integrated using only simple computational operations of existing digital signal processing hardware. The discrete counterpart is derived using a specialised integration technique that replicates the long-term energy properties of the continuous model, thus alleviating problems of numerical discretisation error. Furthermore, the model is *quasi-linear* and thus forms a natural extension to linear prediction, for which efficient, parametric identification techniques have already been developed. It exhibits nonlinear oscillation due to the existence of a *stable limit cycle* that captures the energy balancing inherent to typical stable intonations in continuous speech. The model also captures the observed asymmetry of flow rate output behaviour that

is important to the timbral character of individual speakers. It is also only two-dimensional and thus yields readily to straightforward analysis techniques for planar dynamical systems.

There are two main applications of the model. The first is to *synthesise* speech signals. Here, the output of the discrete vocal fold model u_n is fed directly into the input of a standard, linear model of the vocal tract and lip radiation impedance, to obtain a discrete pressure signal p_n. The input to the vocal fold model is a set of parameters, usually changing in time, that represent the change in configuration of the vocal organs (such as the muscles of the larynx and the lungs) as particular speech sounds are articulated.

The second main application is *analysis* by parameter identification. Here quasi-linear prediction is used to identify the five parameters of the model directly from discrete measurements of the vocal fold flow rate signal u_n. Typically, this signal will be obtained by inverse linear digital filtering of the speech pressure signal from recordings obtained using a microphone.

Figure 2 shows flow diagrams of both synthesis and analysis applications.

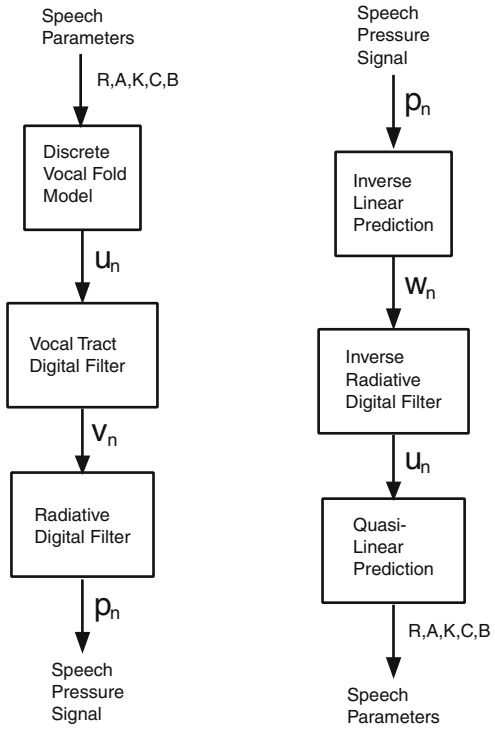

Fig. 2. (Left) Typical arrangement of forward signal processing when the model is used for speech synthesis applications, (Right) Similarly for parameter identification (analysis) applications

2 Deriving the Continuous Model

Figure 3 shows the geometric setup and configuration of the model of the vocal folds. We assume inviscid, laminar air flow in the vocal tract. The areas of the two points A and B in the vocal folds are:

$$a_A = 2lx_A, \quad a_B = 2lx_B \tag{1}$$

where x_A, x_B are the positions of points A and B. Point A is assumed to be stationary, therefore x_A is a parameter of the model.

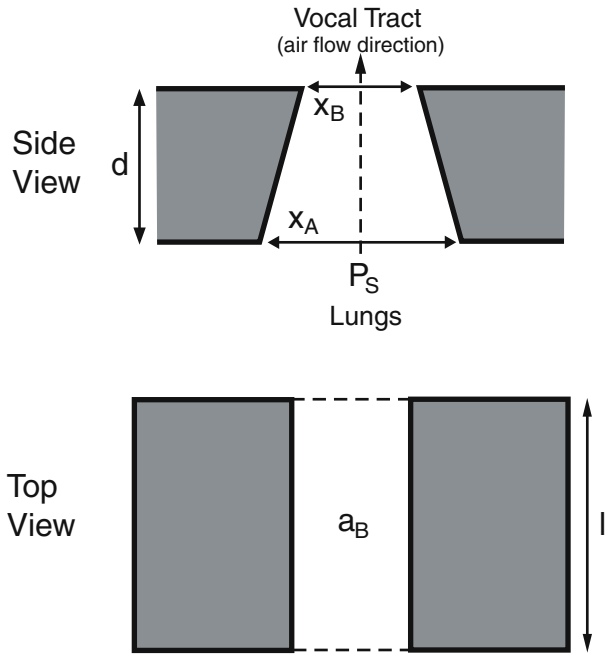

Fig. 3. Geometry and configuration of the nonlinear vocal fold model

The Bernoulli pressure at both points is:

$$\frac{1}{2}\rho_0\frac{U^2}{a_A^2} + P_A = P_\mathrm{S}, \quad \frac{1}{2}\rho_0\frac{U^2}{a_B^2} + P_B = P_\mathrm{S} \tag{2}$$

where U is the air flow rate through the vocal folds, ρ_0 is the equilibrium density of air, and P_S is the (static) lung pressure.

At the top of the vocal folds, a jet is assumed to form such that there is no air pressure. Therefore $P_B = 0$ and so:

$$U = 2l\sqrt{\frac{2P_\mathrm{S}}{\rho_0}}x_B\Phi(x_B) \tag{3}$$

where the Heaviside step function $\Phi(x)$ is used to indicate that there is no air flow when point B is negative (the vocal folds are completely closed at the top). Therefore the air flow rate is proportional to the position of x_B when positive:

$$U \propto x_B, \quad x_B > 0 \tag{4}$$

The pressure at point A is:

$$P_A = P_{\mathrm{S}} - \frac{1}{2}\rho_0 \frac{U^2}{a_A^2} = P_{\mathrm{S}} \left[1 - \Phi\left(x_B\right) \frac{x_B^2}{x_A^2} \right] \tag{5}$$

and the force acting on the vocal fold tissue is assumed to be the average of that at points A and B:

$$F = \frac{1}{2}\left(P_A + P_B\right) ld = \frac{1}{2} ld P_{\mathrm{S}} \left[1 - \Phi\left(x_B\right) \frac{x_B^2}{x_A^2} \right] \tag{6}$$

where l and d are the length and height of the folds respectively.

From now on we write $x = x_B$ for convenience. For the vocal folds, the tissue is assumed to have the following, nonlinear stress-strain relationship [8]:

$$s\left(x, \dot{x}\right) = kx + ax\dot{x} \tag{7}$$

where k is the stiffness of the vocal fold tissue that depends highly upon the tightness of the vocal muscles in the larynx. The parameter a controls the extent of velocity-dependent stiffness of the vocal folds. It is this velocity-dependence [8] of the relationship that causes the important *time asymmetry* of the vocal fold flow rate signal U which is observed in real speech signals [9].

With damping effects of the vocal fold tissue proportional to the velocity the equation of motion for the system is:

$$m\ddot{x} + r\dot{x} + s\left(x, \dot{x}\right) = F = b - c\Phi\left(x\right) x^2 \tag{8}$$

where $b = P_S ld/2$, $c = P_S ld/(2x_A^2)$ and r is the frictional damping constant that depends upon the biomechanical properties of vocal fold tissue.

3 Deriving the Discrete Model

Making use of the *discrete variational calculus* [10] we can derive the discrete equations of motion as:

$$m\left(\frac{x_{n+1} - 2x_n + x_{n-1}}{\Delta t^2}\right) + r\left(\frac{x_n - x_{n-1}}{\Delta t}\right) + ax_n\left(\frac{x_n - x_{n-1}}{\Delta t}\right) - b + kx_n + c\Phi\left(x_n\right) x_n^2 = 0 \tag{9}$$

where n is the time index and Δt is the time difference between samples of a speech signal. Such a discretisation has a *discrete energy expression* that represents the mechanical energy in the vocal folds:

$$E_n = \frac{1}{2}\left(x_{n+1} - x_n\right)^2 + \frac{1}{2}Kx_n^2 \tag{10}$$

and the corresponding rate of change of discrete energy is:

$$\mathrm{d}E_n = -\left(x_{n+1} - x_n\right)\left[R\left(x_{n+1} - x_n\right) + Ax_n\left(x_{n+1} - x_n\right) - B + Cx_n^2\right] \tag{11}$$

where:

$$R = \frac{r\Delta t}{m}, \quad A = \frac{a\Delta t}{m}, \quad B = \frac{b\Delta t^2}{m}, \quad K = \frac{k\Delta t^2}{m}, \quad C = \frac{c\Delta t^2}{m} \tag{12}$$

The discrete equations of motion (9) can be used as an explicit integrator for the model:

$$x_{n+1} = 2x_n - x_{n-1} - R\left(x_n - x_{n-1}\right) - Ax_n\left(x_n - x_{n-1}\right) + B - Kx_n - C\Phi\left(x_n\right)x_n^2 \tag{13}$$

This is a quasi-linear discrete system for which the method of *quasi-linear prediction*, described in the next section, can be used to obtain the parameters from

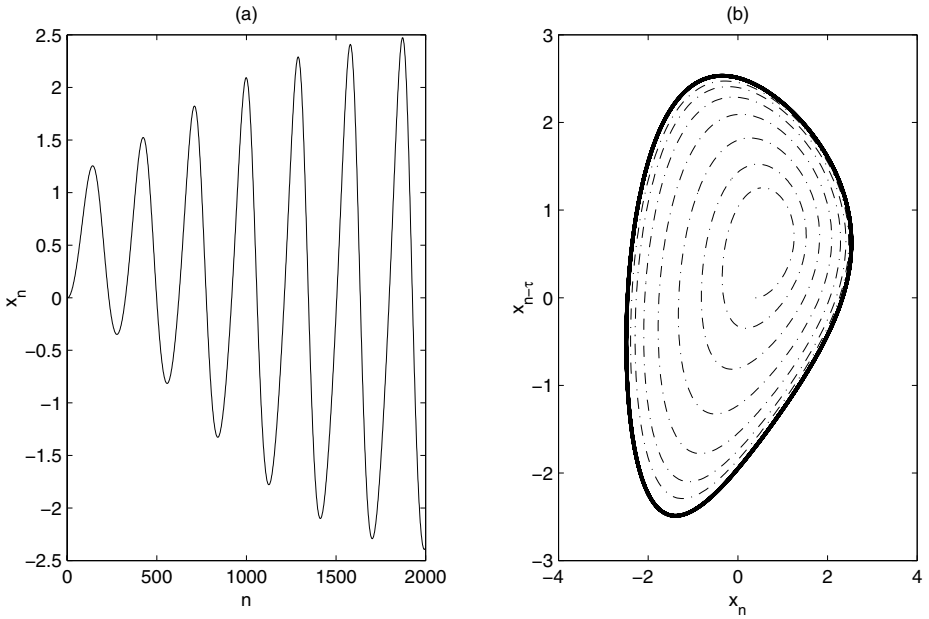

Fig. 4. Typical vocal fold model behaviour with $R = 0.001$, $A = -0.007$, $B = 0.00025$, $K = 0.00026$ and $C = 0.00024$. (a) Time series x_n, (b) Two-dimensional embedding of x_n with embedding delay $\tau = 60$. Thick black line is the limit cycle.

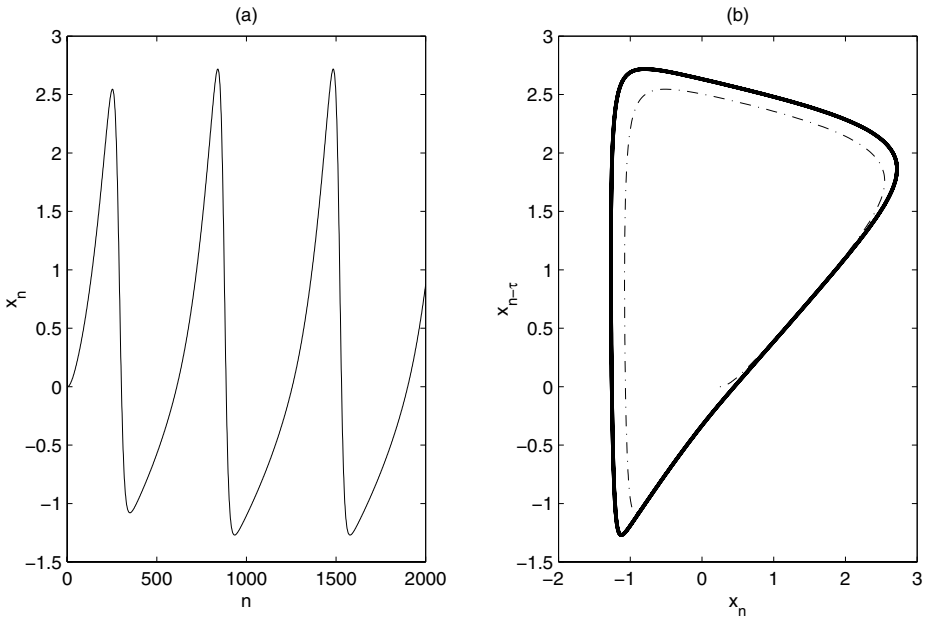

Fig. 5. Typical vocal fold model behaviour x_n with $R = 0.03125$, $A = -0.0375$, $B = 0.000234$, $K = 3.906e-6$ and $C = 0.0002343$. (a) Time series x_n, (b) Two-dimensional embedding of x_n with embedding delay $\tau = 60$. Thick black line is the limit cycle.

a recording of the vocal fold behaviour using a straightforward matrix inversion. The discrete output flow rate is:

$$u_n = 2l\sqrt{\frac{2P_{\mathrm{S}}}{\rho_0}}x_n\Phi\left(x_n\right) \tag{14}$$

Figures 4 and 5 show typical behaviour of the vocal fold model for certain ranges of parameters. Figure 4 shows oscillation in a limit cycle, and Fig. 5 shows asymmetric oscillation, with the rising slope being slower than the falling slope. This is typical of vocal fold behaviour as identified from real, voiced speech signals [9].

In general, when we have obtained a recorded pressure signal p_n we cannot know the real scale. In other words, the recording equipment leads to unknown amplification. This means that, subsequently, the scaling factor $2l\sqrt{\frac{2P_{\mathrm{S}}}{\rho_0}}$ in equation (14) cannot be known. Similarly, there is an additional ambiguity for the mass parameter m which cannot be resolved. This implies that the parameters are scaled by some unknown factor. Therefore we can only compare values obtained from different recordings, and the parameters have no absolute physical interpretation.

4 Quasi-linear Prediction for Parametric Identification

Since the position of point A in the vocal folds, by equation (4), is proportional to the discrete flow rate signal u_n, we assume, initially that $x_n = u_n$. Then we exclude all values of u_n that are negative. Then, by defining the *residual error* as:

$$e_n = x_{n+1} - 2x_n + x_{n-1} + R\left(x_n - x_{n-1}\right) + Ax_n\left(x_n - x_{n-1}\right) - B + Kx_n + C\Phi\left(x_n\right)x_n^2 \quad (15)$$

we can extend the linear prediction process [2], assuming that e_n has a zero-mean, independent Gaussian distribution. This leads to the least-squares solution to find the best fit parameters of the model.

For a non-negative speech signal of length N the system that is the solution to the least-squares problem is:

$$\sum_{n=2}^{N-1} \mathbf{M}\mathbf{a} = -\sum_{n=2}^{N-1} \mathbf{d} \quad (16)$$

where:

$$\mathbf{a} = \left[\, R\ A\ B\ K\ C \,\right]^T \quad (17)$$

The 5×5 system matrix is:

$$\mathbf{M} = \begin{bmatrix} x_n^2 - 2x_nx_{n-1} + x_{n-1}^2 & x_n^3 - 2x_n^2x_{n-1} + x_nx_{n-1}^2 & x_{n-1} - x_n & \cdots \\ x_n^3 - 2x_n^2x_{n-1} + x_nx_{n-1}^2 & x_n^4 - 2x_n^3x_{n-1} + x_n^2x_{n-1}^2 & x_nx_{n-1} - x_n^2 & \\ x_{n-1} - x_n & x_nx_{n-1} - x_n^2 & 1 & \\ x_n^2 - x_nx_{n-1} & x_n^3 - x_n^2x_{n-1} & -x_n & \\ \Phi\left(x_n\right)x_n^2\left(x_n - x_{n-1}\right) & \Phi\left(x_n\right)x_n^3\left(x_n - x_{n-1}\right) & -\Phi\left(x_n\right)x_n^2 & \end{bmatrix}$$
$$\begin{bmatrix} x_n^2 - x_nx_{n-1} & \Phi\left(x_n\right)x_n^2\left(x_n - x_{n-1}\right) \\ x_n^3 - x_n^2x_{n-1} & \Phi\left(x_n\right)x_n^3\left(x_n - x_{n-1}\right) \\ \cdots \quad -x_n & -\Phi\left(x_n\right)x_n^2 \\ x_n^2 & -\Phi\left(x_n\right)x_n^3 \\ \Phi\left(x_n\right)x_n^3 & -\Phi\left(x_n\right)x_n^4 \end{bmatrix} \quad (18)$$

and:

$$\mathbf{d} = \begin{bmatrix} x_{n+1}x_{n-1} - x_{n+1}x_n + 2x_n^2 - 3x_nx_{n-1} + x_{n-1}^2 \\ x_{n+1}x_nx_{n-1} - x_{n+1}x_n^2 + 2x_n^3 - 3x_n^2x_{n-1} + x_nx_{n-1}^2 \\ x_{n+1} - 2x_n + x_{n-1} \\ 2x_n^2 - x_{n+1}x_n - x_nx_{n-1} \\ -\Phi\left(x_n\right)x_n^2\left(x_{n+1} - 2x_n + x_{n-1}\right) \end{bmatrix} \quad (19)$$

The coefficients, $\mathbf{a}$ of the quasi-linear model, taken together with the residual e_n and the initial conditions x_1, x_2 as a set, form a one-one representation of the modelled data x_n, and we can exactly reconstruct x_n using only this information.

5 Discussion and Conclusions

This paper has introduced a new, simplified model of vocal fold dynamics. Making use of variational integration methods, a corresponding discrete counterpart is derived. This discrete model can then be used to synthesise vocal fold flow rate signals or used to identify model parameters from estimated vocal fold flow rate signals obtained from speech pressure recordings.

The main advantage of this model is that it captures the overall features of vocal fold oscillation (limit cycles, open/close quotient and pulse asymmetry) whilst the computational complexity is low, meaning that it can be implemented on standard digital signal processing hardware. The corresponding disadvantage is that certain interesting vocal pathologies cannot be replicated, for example *creaky voice* which is suggested to originate from period doubling bifurcations [11].

In a related study, this model has been used to identify and resynthesise estimated vocal fold flow rate signals. Discovering that the model was capable of replicating actual flow rate signals with some successes and some failures, the suggested method for parametric identification is to ensure positivity of all parameters (non-negative least squares [12]), and subsequently reversing the sign of the parameter A. This guarantees that the parameters conform to the modelling assumptions.

References

1. Markel, J.D., Gray, A.H.: Linear prediction of speech. Springer-Verlag (1976)
2. Proakis, J.G., Manolakis, D.G.: Digital signal processing: principles, algorithms, and applications. Prentice-Hall (1996)
3. Kleijn, K., Paliwal, K.: Speech coding and synthesis. Elsevier Science, Amsterdam (1995)
4. Story, B.H.: An overview of the physiology, physics and modeling of the sound source for vowels. Acoust. Sci. & Tech. **23** (2002)
5. Ishizaka, K., Flanagan, J.: Synthesis of voiced sounds from a two-mass model of the vocal cords. ATT Bell System Tech Journal **51** (1972) 1233–1268
6. Steinecke, I., Herzel, H.: Bifurcations in an asymmetric vocal-fold model. J. Acoust. Soc. Am. **97** (1995) 1874–1884
7. Little, M.A., Moroz, I.M., McSharry, P.E., Roberts, S.J.: System for generating a signal representative of vocal fold dynamics. (2004) UK patent applied for.
8. Chan, R.W.: Constitutive characterization of vocal fold viscoelasticity based on a modified arruda-boyce eight-chain model. J. Acoust. Soc. Am **114** (2003) 2458
9. Holmes, J.: Speech synthesis and recognition. Van Nostrand Reinhold (UK) (1988)
10. Marsden, J., West, M.: Discrete mechanics and variational integrators. Acta Numerica (2001) 357–514
11. Mergell, P., Herzel, H.: Bifurcations in 2-mass models of the vocal folds – the role of the vocal tract. In: Speech Production Modeling 1996. (1996) 189–192
12. Lawson, C., Hanson, R.: Solving least square problems. Prentice Hall, Englewood Cliffs NJ (1974)