ELSEVIER

Medical Engineering & Physics 24 (2002) 453–460

Medical Engineering & Physics

www.elsevier.com/locate/medengphy

Hybrid parametric-physiological glottal modelling with application to voice quality assessment

Carlo Drioli [*], Federico Avanzini

Dipartimento di Elettronica e Informatica, Università di Padova, Via Gradenigo 6/a, Padova, Italy

Received 25 April 2002; accepted 24 May 2002

Abstract

A glottal model based on physical constraints is proposed. The model describes the vocal fold as a simple oscillator, i.e. a damped mass-spring system. The oscillator is coupled with a nonlinear block, accounting for fold interaction with the airflow. The nonlinear block is modelled as a regressor-based functional with weights to be identified, and a pitch-synchronous identification procedure is outlined. The model is used to analyse voiced sounds from normal and from pathological voices, and the application of the proposed analysis procedure to voice quality assessment is discussed. © 2002 IPEM. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Nonlinear models of vocal emission; Physical models; Estimation and modelling of the glottal flow; Nonlinear identification; Voice quality assessment

PACS: 43.70.Bk; 43.70.Dn

1. Introduction

Many of the acoustic and perceptual features of an individual's voice are believed to depend on specific characteristics of the quasi-periodic excitation signal (glottal flow waveform) provided by the vocal folds. Different approaches have been adopted for the modelling of the glottal flow, the most important ones being the parametrization by analytical functions [1] and the physiological modelling of the glottis. The use of parametric and physiological models of vocal emission has been proposed for a wide range of applications, namely speech synthesis, speech coding and compression, and voice quality analysis and assessment [2–4].

The first physiological model that attempted to represent the vocal folds vibration by means of a mass-spring analogy was proposed by Flanagan and Landgraf in 1968 [5]. It is made of a single spring-mass oscillator driven by airflow from the lungs. Although such a simple system with one degree of freedom cannot accurately simulate the fine details of the functioning of the glottis, it is able to capture the basic mechanisms that initiate self-sustained oscillations of a pressure-controlled valve [6]. Due to low computational loads, one-mass models are used by many authors in articulatory speech synthesizers (see [7]).

However, the model by Flanagan and Landgraf [5] is able to produce self-sustained oscillations only when interfaced with an inertive vocal tract load. The reason is that the system is not able to reproduce the vertical movement of the vocal fold tissue that accompanies the main lateral movement [8,9]. This feature, often referred to as vertical phase difference, has the important role of producing different glottal shapes within a glottal flow cycle and provides a mean to generate a driving force which is asymmetrical within the cycle. If the driving force in the closing phase is lower than the driving force in the opening phase (i.e., the force is dependent on the displacement velocity), a positive flow of energy from the airstream to the tissue can be realized, and a flow-induced sustained oscillation can be achieved [8].

Among the many methods that have been proposed to take into account the driving force asymmetry, the two-mass model proposed by Ishizaka and Flanagan (*IF*) [10]

[*] Corresponding author.
E-mail address: drioli@dei.unipd.it (C. Drioli).

PII: S1350-4533(02)00057-7

is the most widely known: the *IF* model represents each of the vocal folds as a pair of damped mass-spring systems coupled to each other. However, also the *IF* model (and in general multi-mass lumped models) is derived using crude approximations in the mechanical, acoustical and fluid dynamic modelling. As noted by Villain et al. [11], elementary mechanical constraints on the physiological problem are completely neglected: it is assumed that the elastic structure is fixed to a rigid wall, while in reality a significant radiation of surface waves from the throat can be noticed when voiced sounds are produced. Another limitation is concerned with the glottal closure: glottal areas are assumed to be rectangular in the *IF* model. As a consequence, closure of the glottis occurs in an abrupt manner and this introduces additional energy at high frequencies. In natural 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).

In this paper we propose a lumped-mass modelling approach which is suited for parametric identification from real data. The model is made of two main parts, a linear mass-spring system which describes the vocal folds as a harmonic oscillator, and a nonlinear block which accounts for the interaction between glottal flow, glottal pressure, and position of the vocal folds [12]. The one-mass paradigm is adopted due to its simplicity and its low-dimensional parameter set, and the model is provided with the necessary modifications in order to make it suitable for voice analysis/identification purposes.

The proposed model is then used for the analysis and assessment of voice quality, including the detection and classification of voice pathologies. Voice quality assessment is traditionally based on subjective perceptual ratings which is considered, still today, the only reasonable way to classify certain types of voice disorders. The objective assessment of voice, however, remains an open problem that calls for reliable analysis tools and algorithms to aid the clinicians in the diagnosis.

Acoustical analysis based on modern signal processing techniques is becoming popular due to its quantitative and noninvasive nature [13,14]. Many researchers propose to assess perceptual voice quality using a set of parameters derived from analysis of the radiated pressure signal. Examples are time-domain measures such as jitter (variation between successive fundamental periods) and shimmer (variation between magnitudes in successive fundamental periods), and frequency-domain measures such as the spectral slope or spectral flatness of the inverse filtered residue [14,15]. However, all of these parameters depend exclusively on features of the signals, while no assumptions are made about the physiology of the source. We suggest here that the information gathered from simple physically-informed models of the glottis, such as the one proposed in this paper, can add valuable information when combined with the set of traditional cues, and can be helpful in the detection and classification of voice pathologies.

The rest of the paper is organized as follows. In section 2.1, a review of the methods for glottal flow estimation from a measured speech signal is presented. A physically-informed model is proposed in section 2.2, and section 2.3 outlines the approach to parametric identification. In section 3, the use of the model as an analysis tool is proposed and the results of the analysis on normal and pathological voices are discussed. The conclusions are presented in section 4.

2. Physically-informed model

2.1. 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 (the voice source). Most typically, the voice source signal cannot be measured directly, whereas the only measurable signal is the output from a speaker's mouth (i.e., the radiated pressure). Therefore, one fundamental task in many speech analysis and modelling approaches is the accurate estimation of the voice source, and its separation from the effects due to the vocal tract. This estimation allows precise determination of voice source features such as the glottal flow waveform, the glottis opening and closing instants, etc.

Our approach to voice source modelling assumes that, given a flow waveform from a steady portion of a voiced sound, the parameters of the model under study can be adapted so to reproduce the same waveform. A second simplifying assumption is that there is approximately no interaction between the source and the vocal tract. Hence, the modelling procedure is made of two sequential steps, the first being the estimate of the glottal flow from the radiated pressure signal, and the second being the fitting of the model to the resulting flow signal.

The most common techniques rely on linear prediction coding schemes (LPC). These methods estimate the vocal tract filter, and provide the source signal (or *residual*) by inverse filtering the radiated pressure signal with the estimated tract filter. However, the vocal tract characteristics change within a pitch period because of the opening and closing of the glottis: therefore the determination of the poles of the tract filter is often carried out by a covariance LP analysis restricted to the closed glottis period [16]. The remainder of this section describes this procedure, which is used in the following to estimate glottal flow waveforms from radiated pressure waveforms of voiced utterances.

The method requires an initial estimate of the closing glottis instants (CGI). A rough estimate is provided by the peaks in the residual error of a pitch-asynchronous autocorrelation LP analysis. Once a CGI is chosen, a pitch-synchronous covariance LP analysis is used, which starts at time instant CGI+1 and is limited to the closed-phase, and estimates the all-pole vocal tract filter. This filter should model the formant structure of the speech signal, but the resulting polynomial often exhibits poles in excess, that do not contribute to any formant. For this reason, an improved all-pole filter is constructed by finding the roots of the original LP polynomial and discarding the roots that correspond 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 glottal flow waveform (note that this approach amounts to modelling the lip radiation effect with a differentiator filter). Integration of this latter waveform then provides the estimated glottal flow. Since the first CGI estimate is not always accurate, a small number of covariance LP analysis centred around the CGI estimate are usually performed, and the best result is then selected on the basis of the residual characteristics.

2.2. The glottis model

The glottis model proposed here is based on the lumped-mass models described in [5,10]. These are made of two main functional blocks. The first one represents the mechanical behaviour of the vocal folds: each fold is described by means of one or two masses, which are connected to the fold body by springs and dissipative elements, and oscillate driven by the pressure distribution at the glottis. As such, this block is modelled as a quasi-linear differential equation. The resonance frequencies of the oscillators determine some significant features of the glottal signal, such as the pitch and the open quotient.

The second block is highly nonlinear and represents the coupling between the vocal fold motion, the glottal flow, and the glottal pressure distribution. Using very crude approximations (e.g., quasi-steadiness of the flow), Ishizaka and Flanagan [10] derive the nonlinear equations that describe the pressure drops and recoveries along the glottis. These depend upon the displacement of the vocal fold, which in turn is determined by the glottal pressure: as a consequence, the two blocks are coupled in a feedback loop.

Many refinements have been proposed to these models, in which the physical behaviour of the glottis is described in finer details. This approach typically results in more realistic but at the same time more complicated models. The opposite approach is taken here: the overall model structure is retained, but both the linear and nonlinear blocks are drastically simplified by dropping physical information. In the remainder of this section the model is described in the digital domain.

The linear block is modelled as the simplest oscillating system, i.e. a second-order filter $H_{\mathrm{res}}(z)$, which relates the vocal fold displacement x_1 to the glottal flow u_g. The transfer function can be written as

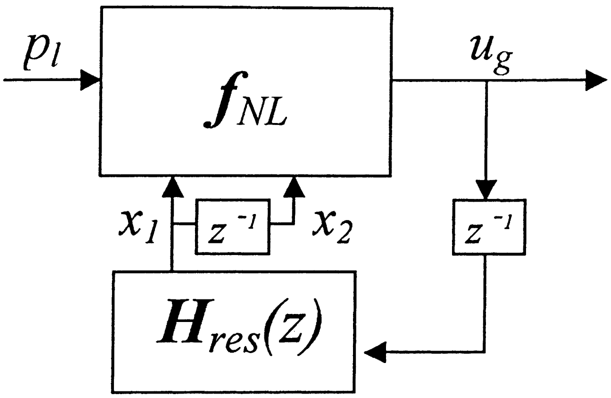

Fig. 1. Parametric model of the glottis with physical constraints. H_{res} is a simple resonant filter, tuned on the pitch of the voiced signal, $f_{\mathrm{NL}}(x_1,x_2,p_l)$ is a nonlinear function, u_g is the glottal flow signal, x_1 and x_2 are respectively the output and the delayed version of the output, and p_l is the lung pressure. A delay z^{-1} is inserted in the loop in order to make it computable in a signal processing environment.

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

where β_0 is a gain factor and the coefficients α_i determine the resonance frequency f_0 and the 3-dB bandwidth Δf of the filter. Note that the filter H_{res} is the digital equivalent of a mechanical system in which a mass is connected to a linear spring and a damping element. In this sense Eq. (1) describes a 1-mass model analogous to the one described in [5]. The output $x_1(k)$ of H_{res} represents the fold displacement. A second displacement variable, $x_2(k) = x_1(k-1)$, is derived as a delayed version of the input $x_1(k)$. This term provides an additional degree of freedom to the system and roughly simulates a vertical phase difference along the vocal fold tissue.

The second block is a nonlinear map f_{NL} that models the interaction between vocal fold motion, flow, and pressure. Given the lung pressure p_l, and the state (x_1,x_2) of the linear filter H_{res}, the mapping $f_{\mathrm{NL}}(x_1,x_2,p_l)$ returns the flow u_g. This is taken as a driving signal for the oscillator H_{res}. The final structure of the model is the one given in Fig. 1. Note that very similar block schemes can be used to describe the models proposed in [5,10], although the internal structures of the blocks are different.

The nonlinear map f_{NL} is modelled with a regressor-based functional:

$$f_{\mathrm{NL}}(x_1,x_2,p_l) = w_0 + \sum_{i=1}^{M} w_i \psi_i(x_1,x_2,p_l) \tag{2}$$

where $\{\psi_i(x_1,x_2,p_l)\}_{i=1}^{M}$ is a set of M regressors of the input data and $\{w_i\}_{i=1}^{M}$ are the corresponding weights to be identified. The choice of the regressor set can be made in several ways. Local models, such as gaussian functions or any other radial basis function, are often used. This approach leads to a model called *Radial Basis Function Network* (RBFN) [17], used in the field of time series analysis and modelling. The use of a polynomial expansion of the input leads to a class of NARMAX models [18], known in the fields of system identification and control. Alternatively, the regressors can be selected on the basis of physical considerations.

We choose here to select a small number of terms from a polynomial expansion of the input. The choice of these terms was made empirically, by fitting the model to a training data set for each new added term, and retaining only those terms that significantly improved the identification results. It was found that a small number of polynomial terms, not exceeding ten, is usually sufficient to provide the required accuracy of the identification. The set of regressors selected in this case was:

$$\begin{array}{ll} \psi_1(x_1,x_2,p_l) = x_1 \cdot p_l & \psi_4(x_1,x_2,p_l) = x_2/(x_1 + \in) \\ \psi_2(x_1,x_2,p_l) = x_1^3 & \psi_5(x_1,x_2,p_l) = x_1^2/(x_2 + \in) \\ \psi_3(x_1,x_2,p_l) = 1/(x_1 + \in) & \psi_6(x_1,x_2,p_l) = x_1 \cdot x_2 \end{array} \tag{3}$$

where $\in$ is a constant offset that prevents the denominator from assuming a zero value.

2.3. Pitch-synchronous parametric identification

The following problem is now addressed: given a *target* glottal flow waveform $\bar{u}_g$ (which is estimated, together with the vocal tract filter, using the inverse filtering technique described in section 2.1) the parameters of the model have to be identified so that the output u_g from the nonlinear block fits the target as closely as possible.

System identification is achieved in three steps.

1. For each period $p = 1,\ldots,P$ of the glottal flow $\bar{u}_g$, the starting time n_p and the period length N_p (in samples) are defined as the CGIs and the difference between the two successive CGIs, respectively. The lung pressure p_l is given a value which is constant for each period and proportional to the energy of $\bar{u}_g$ within the period.
2. The linear block H_{res} is driven using $\bar{u}_g$, and its output is computed. The resonance frequency f_0 is chosen interactively, in such a way that the open and closed phase for the output match those of the target flow $\bar{u}_g$. The bandwidth Δf is chosen so that the quality factor Q of H_{res} matches a reference value Q_0 deduced from physical parameters: the values used in [10] for masses, spring constants, and dissipation elements provide a value $Q_0 \approx 10$. Once the parameters f_0 and Δf are chosen, the output $x_1(k)$ and the state $x_2(k)$ from the oscillator are computed. From Fig. 1, it can be seen that, if the flow segment is $u_g(k)$, $k = n_{p+1},\ldots,n + N_p$, then $x_1(k) = h_{res} * u_g(k-1)$, $k = n_p + 1,\ldots,n + N_p$ and $x_2(k) = x_1(k-1)$, where the operator $*$ denotes convolution. Our analysis focuses on the reconstruction of the complementary relation, i.e. the mapping from x_1 and x_2 back to u_g, via the function f_{NL}.
3. At this point, both the input (x_1,x_2,p_l) and the target output $\bar{u}_g$ of the nonlinear block are available. Then the weights in Eq. (2) are estimated using the following pitch-synchronous nonlinear identification step.

For each period $p = 1,\ldots,P$, two *training data sets* are defined as

$$\mathbf{T}_{\bar{u}_g}(p) = [\bar{u}_g(n_p + 1),\bar{u}_g(n_p + 2),\ldots,\bar{u}_g(n_p + N_p)], \tag{4}$$

$$\mathbf{T}_x(p) = \begin{bmatrix} \psi_1(n_p) & \cdots & \psi_1(n_p + N_p - 1) \\ \vdots & \ddots & \vdots \\ \psi_6(n_p) & \cdots & \psi_6(n_p + N_p - 1) \end{bmatrix}, \tag{5}$$

where $\psi_i(k) = \psi_i(x_1(k),x_2(k),p_l(k))$, i.e.$\psi_i(k)$ is the ith regressor at the discrete time k. The data sets in Eqs (4) and (5) are used to train the regressors in f_{NL}. Specifically, the identification of the weights $\mathbf{w}(p)$ requires the solution of the matrix system

$$\mathbf{w}(p)\begin{bmatrix} \mathbf{1} \\ \mathbf{T}_x(p) \end{bmatrix} = \mathbf{T}_{\bar{u}_g}(p), \tag{6}$$

where $\mathbf{1} = [1,\ldots,1]$ is a row vector of length N_p. The LS solution of Eq. (6) is known to be

$$\mathbf{w}(p) = \mathbf{T}_{\bar{u}_g}(p)\begin{bmatrix} \mathbf{1} \\ \mathbf{T}_x(p) \end{bmatrix}^+, \tag{7}$$

where the symbol + denotes the pseudo-inverse of a matrix [17]. Pseudo-inversion provides a method to invert non-square matrices, and it is required in this case since the number of regressors is typically less than the number of samples in one period of the flow waveform.

The model and the identification procedure described above have been successfully used to obtain modifications of basic sound parameters, such as pitch and voice quality, mainly for voice synthesis and processing purposes [12]. These applications exploit the ability of the identified model to self oscillate and generate autonomously a given glottal flow waveform. As this is a nonlinear iterated system, stability is critical and accurate identification is a key point for the performance.

This work explores the suitability of this model for a different purpose, namely the analysis of voiced sounds for voice quality assessment and glottis pathology detection. It should be highlighted that stability is not a critical point when the model is used for analysis purposes: in this case the model is not required to autonomously generate the glottal pulse, but it is instead used to obtain a new representation of the flow, suitable for the analysis and assessment.

3. Analysis of voiced sounds and voice quality assessment: testing and results

3.1. Testing the identification procedure

An example of the analysis procedure described in the previous sections is shown in Fig. 2. Two voiced frames from two different speakers were used to estimate the glottal flow (first and second plot from top) and to fit the model (third and fourth plots from top). The representation of the f_{NL} input–output time series, called phase portrait, is a common way to represent the qualitative nonlinear dynamics for low-dimensional dynamical systems, and is the basis for a class of nonlinear system identification models. A similar approach to the representation of the dynamics of vocal emission with a focus on speech and vocal disorders analysis can be found in [4,19].

Usually, the phase portraits are directly derived from the time series representing the radiated pressure, thus retaining a mixed information in which the contribution of the vocal tract, the glottis, and their interaction, are indistinguishable. With the approach here described, the use of a physically informed model permits to localize the observation of the time series with respect to the nonlinear excitation mechanism, thus limiting the influence of other elements. One direct consequence is that the phase portrait is more representative of the dynamics of the glottis, since it is independent from the vocal tract and from the pitch of the signal taken into account by the filter H_{res}. The first column of plots shows the analysis of a fragment of a voiced sound from a healthy speaker uttering the vowel /a/. The values of the parameters $\mathbf{w}(p)$, strictly tied to the shape of the phase portrait, are associated with the waveform of the glottal flow, while their relative changes in time are associated with the stability of the waveform. The second column of plots shows a different speaker uttering the vowel /a/ at constant pitch. In this case, a pathological voice was considered from a patient with unilateral cordectomy. The time variation of the model parameters well reflects the period shape instability of the glottal flow waveform.

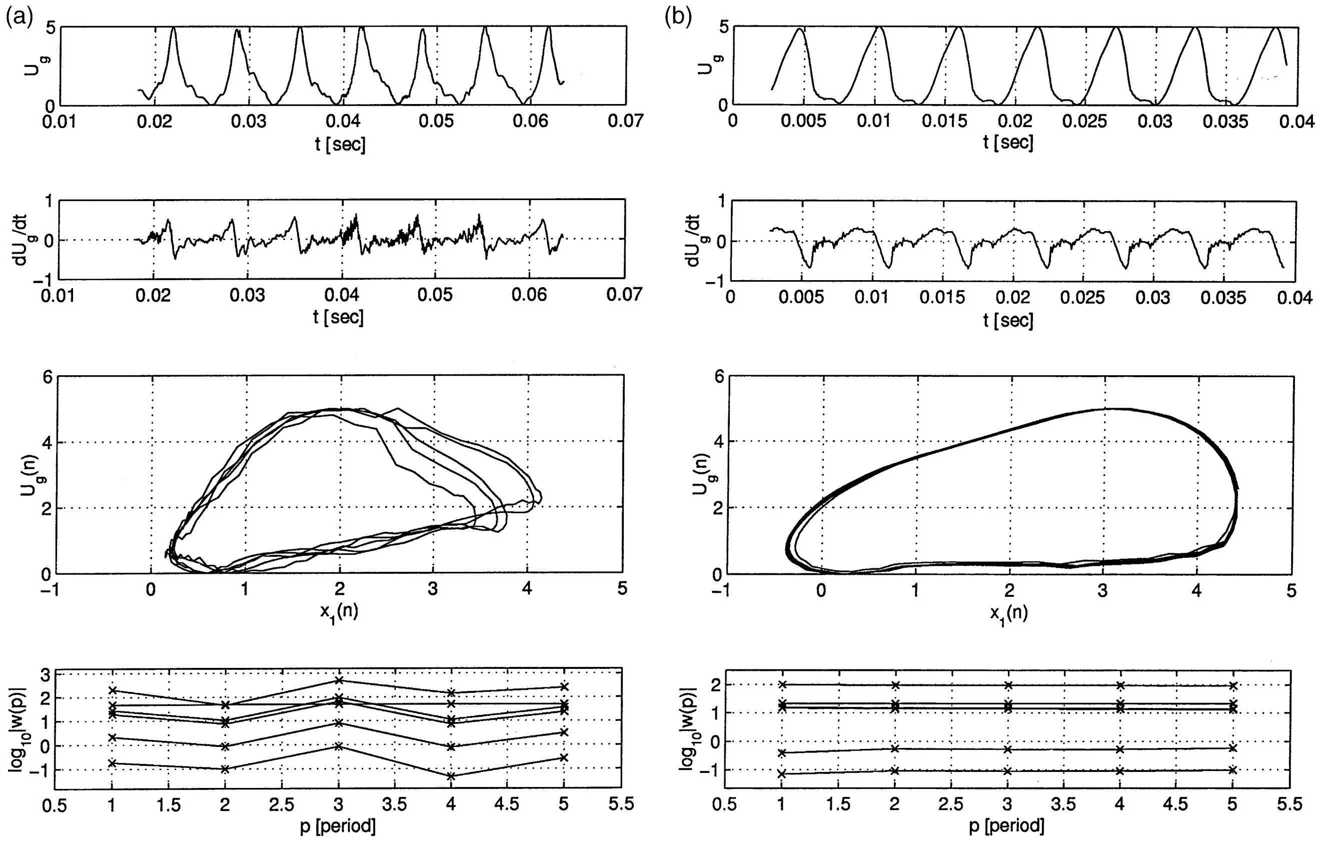

Fig. 2. Result of the model-based analysis for two voice samples, namely (left to right) a male uttering the vowel /a/ at a pitch of 177 Hz, and the same vowel uttered at a pitch of 154 Hz by a patient subjected to partial cordectomy. The plots show (from top to bottom) the estimated glottal flow (normalized to an arbitrary scale), the differentiated glottal flow, the projection of the phase portrait on the (x_1, u_g) plane, and the time-varying parameters $w_i(p)$, of the nonlinear function f_{NL} (log-scale).

3.2. Acoustic analysis for pathological voice assessment

To test the potentialities of the analysis based on physiological models, a number of high quality voice samples were collected from healthy subjects and from dysphonic patients. The pathologies taken into account were mainly partial cordectomy and unilateral paralysis. Utterances of sustained vowel /a/ were recorded with a Sliure SM 48 dynamic microphone and a Kay CSL system, at a sampling rate of 11,025 Hz.

A set of nine different voices is analysed in the remainder of this section. From the generic *i*th voice ($i = 1,\ldots,9$), a suitable stationary portion of the signal was selected and analysed. For each subject, a frame of five periods of the signal was considered ($P_i = 5$, $i = 1,\ldots,9$). Each period of the estimated glottal flow, obtained by inverse filtering, was normalized in magnitude to focus the observation on the shape variations (consequently, shimmer information was extracted at this stage). The time-varying pitch was used to set, at each period, the resonance frequency of the filter $H_{\mathrm{res}}(z)$. Then, the training data sets $\mathbf{T}_{u_g,i}(p)$ and $\mathbf{T}_{x,i}(p)$, $p = 1\ldots P_i$, were computed by feeding $H_{\mathrm{res}}(z)$ with the glottal flow estimate. The training sets were finally used to compute the time-varying parameters $\mathbf{w}(p)$ for each case.

The result of the analysis of nine selected cases is shown in Fig. 3. The plots on the left represent the phase portraits as in Fig. 2. A qualitative inspection suggests that a classification could be based on two main features, namely the shape and the stability of the orbits in the phase space.

To further investigate the behaviour of the **w** parameters, the matrices $\mathbf{W}_i$ (each one having dimension $(M + 1) \times P_i$, i.e.7×5) were collected into a global matrix $\mathbf{W} = [\mathbf{W}_1|\mathbf{W}_2|\ldots|\mathbf{W}_9]$, and performed a principal component analysis based on the singular value decomposition of the matrix **W** [20]. Here, the principal component analysis has the purpose of reducing the dimensionality of the matrix **W**, to better visualize and evaluate the data.

The eigenvalues matrix computed in the decomposition confirmed that two principal axes are sufficient to explain over 90% of the overall variance in the data, thus allowing to represent the data on a two-dimensional space (as in the right plot of Fig. 2). A point in this space represents a single period of the glottal flow waveform, and different positions represent different period shapes. An utterance is then represented in this space by a set of points, as in Fig. 4 (the points that are periods of the same utterance are represented in the figure by the same marker and clustered together).

By observing the overall distribution of the points and comparing the clusters with the corresponding phase portraits on the left, one can deduce that utterances from the healthy voices considered are characterized by clusters made of points tight together (denoting high stability of the period shape). On the other hand, pathological

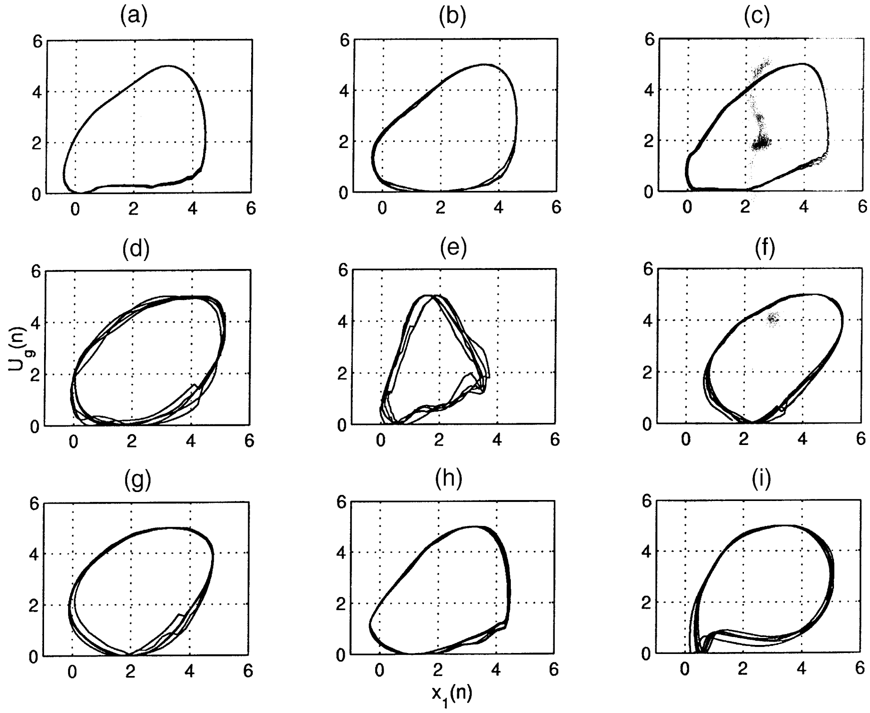

Fig. 3. Analysis of a set of sustained vowels from normal and pathological voices. The plots show the projection of the phase portraits on the (x_1, u_g) plane. All plots refer to five periods of the flow waveform. Cases (a), (b) and (c) are utterances from three healthy male subjects. Cases (d)–(i) refer to subjects with vocal fold pathologies (specifically, (d)–(e) to partial cordectomy, and (f)–(i) to unilateral paralysis).

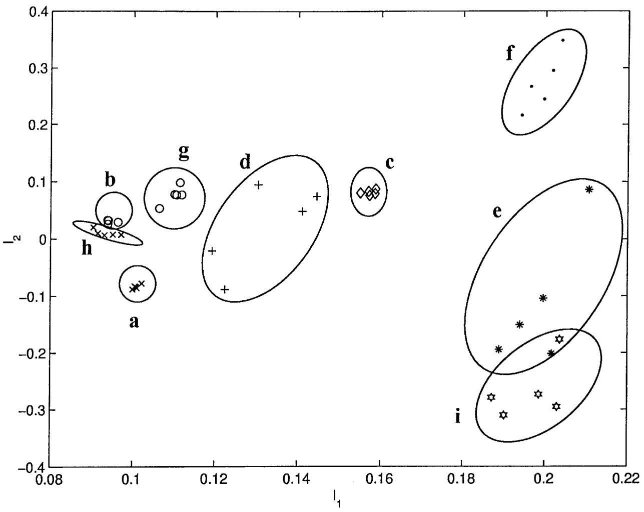

Fig. 4. Clustering of the data in a two-dimensional space after reduction of the dimensionality by principal component analysis. Each point in the plot corresponds to a glottal waveform period. Data from the same subjects are represented with the same marker.

voices, characterized by period shape instability, are represented by spread out clusters. The proximity of different clusters in the two-dimensional space denotes similarity of waveforms, as for cases (b) and (h) in Fig. 4.

It is useful to compare the approach proposed here to other existing voice quality assessment methods based on acoustic features. Measures of perturbation of the waveform shape, usually obtained by correlation of pairs of adjacent waveform periods, are often combined to other measures of periodicity perturbations and of the noise component. An example of such representation is the ‘hoarseness diagram’, which is based on three acoustic measures that assess different aspects of signal periodicity (jitter, shimmer, and mean period correlation, MWC) and on one measuring the noise component (glottal to noise excitation ratio, GNE) [15,21]. The data can be plotted in a 2-dimensional space in which the *x*-axis is the irregularity component and the *y*-axis is the noise component. Typical hoarseness diagrams for pathological voices are characterized by the coverage of large areas, whereas normal voices are clustered in given regions, namely lower left, corresponding to low values of both the noise and the irregularity component.

Whether the method proposed here could present some advantages over this and other representations, is a matter of assessing if it can supply any additional information other than a measure of periodicity perturbation. As noticed above, the position of a voice sample in the two-dimensional space is determined in some way by the shape of the waveform period: this suggests that the position of the clusters can be used to distinguish healthy voices from pathological ones, or to detect and classify certain pathologies.

However, as shown in Figs. 3 and 4, there was no evidence in this experiment that healthy voices and pathological voices occupy mutually exclusive regions, nor that the same pathologies occupy the same regions in the two-dimensional space. We believe that further work on a larger data set could be helpful in understanding these aspects. Moreover, a better understanding of the meaning of the principal components resulting from the dimensionality reduction is also important. These should be related to some gross shape features of the waveform period, such as the closed phase duration (if any), the slope of the opening and closing phases, or others.

4. Conclusions and future work

In this paper we presented a physically-informed model loosely based on the traditional scheme of a mechanical oscillator driven by the glottal pressure, whose structure is adequate for the fitting of real data. We then proposed an analysis procedure that can be used in combination with the model, and we discussed the possibility of using this procedure to assist in the assessment of voice quality and in the detection of vocal fold pathologies. Future work is foreseen to improve the model, in order to take into account important features such as the interaction between the glottis and the vocal tract, and to refine the investigation of voice pathology detection and classification. The analysis of enlarged sets of pathological cases should better clarify, for example, if voices from subjects affected by certain pathologies can be said to occupy the same region of the low-dimensional space, offering in this case a reliable tool for pathology detection and classification.

References

[1] Fant G, Liljencrants J, Lin Q. A four-parameter model of glottal flow. STL-QPSR 1985; 4:1-13.

[2] Schönweiler R, Hess M, Wübbelt P, Ptok M. Novel approach to acoustical voice analysis using artificial neural networks. J Assoc Res Otolaryngol 2000; 1: 270–282.

[3] Cranen B, Schroeter J. Physiologically motivated modelling of the voice source in articulatory analysis/synthesis. Speech Commun 1996;19:1–19.

[4] Herzel H, Berry D, Titze I, Saleh M. Analysis of vocal disorders with methods from nonlinear dynamics. J Speech Hearing Res 1994;37:1008–19.

[5] Flanagan JL, Landgraf LL. Self-oscillating source for vocal-tract synthesizers. IEEE Trans Audio Electroacoust 1968;16:57–64.

[6] Fletcher NH. Autonomous vibration of simple pressure-controlled valves in gas flows. J Acoust Soc Am 1993;93(4):2172–80.

[7] Meyer P, Wilhelms R, Strube HW. A quasiarticulatory speech synthesizer for German language running in real time. J Acoust Soc Am 1989;86(2):523–39.

[8] Titze IR. The physics of small-amplitude oscillations of the vocal folds. J Acoust Soc Am 1988;83(4):1536–52.

[9] Rodet X. One and two mass model oscillations for voice and instruments, Proc. Int. Computer Music Conf. (ICMC’95), Banff, 1995.

[10] Ishizaka K, Flanagan JL. Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Syst Tech J 1972;51:1233–68.

[11] Villain C, Le Marrec L, Op't Root W, Willems J, Pelorson X, Hirschberg A. Towards a new brass player's lip model. Proc. Int. Symp. on Musical Acoustics (ISMA'0l), Perugia, 2001:107–10.

[12] Drioli C, Avanzini F. Model-based synthesis and transformation of voiced sounds. Proc. COST-G6 Conf. on Digital Audio Effects (DAFX-00), 2000:45–9.

[13] de Oliveira Rosa M, Pereira JC, Grellet M. Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Trans Biomed Engng 2000;47(1):96–104.

[14] Gavidia-Ceballos L, Hansen JHL. Direct speech feature estimation using an iterative em algorithm for vocal fold pathology detection. IEEE Trans Biomed Engng 1996;43(4):373–83.

[15] Michaelis D, Fröhlich M, Strube HW. Selection and combination of acoustic features for the description of pathologic voices. J Acoust Soc Am 1998;103(3):1628–39.

[16] Wong D, Markel J, Gray AH. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans Acoust Speech Sig Process 1979;27(4):350–5.

[17] Chen S, Cowan CFN, Grant PM. Orthogonal least squares learning algorithm for radial basis functions networks. IEEE Trans Neural Net 1991;2(2):302–9.

[18] Chen S, Billings SA. Representation of non-linear systems: NARMAX model. Int J Control 1989;49(3):1013–32.

[19] Kumar A, Mullick SK. Nonlinear dynamical analysis of speech. J Acoust Soc Am 1996;100(1):615–29.

[20] Jolliffe I. Principal components analysis. New York: Springer-Verlag, 1986.

[21] Michaelis D, Gramss T, Strube HW. Glottal to noise excitation ratio—a new measure for describing pathological voices. Acustica-Acta Acustica 1997;83:700–6.