Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Classification of functional voice disorders based on phonovibrograms

Daniel Voigt [a,*], Michael Döllinger [a], Thomas Braunschweig [b], Anxiong Yang [a], Ulrich Eysholdt [a], Jörg Lohscheller [c]

[a] Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Bohlenplatz 21, D-91054 Erlangen, Germany
[b] Department of Phoniatrics and Pediatric Audiology, University Hospital Jena, Stoystraße 3, D-07743 Jena, Germany
[c] University of Applied Sciences Trier, Department of Computer Science, Medical Informatics, Schneidershof, D-54293 Trier, Germany

ARTICLE INFO

Article history:
Received 12 November 2008
Received in revised form 20 August 2009
Accepted 10 January 2010

Keywords:
Voice pathology classification
Laryngeal high-speed video
Diagnosis support system
Phonovibrogram
PVG
Feature extraction
Pattern recognition
Machine learning

ABSTRACT

Objective: This work presents a computer-aided method for automatically and objectively classifying individuals with healthy and dysfunctional vocal fold vibration patterns as depicted in clinical high-speed (HS) videos of the larynx.

Methods: By employing a specialized image segmentation and vocal fold movement visualization technique – namely phonovibrography – a novel set of numerical features is derived from laryngeal HS videos capturing the dynamic behavior and the symmetry of oscillating vocal folds. In order to assess the discriminatory power of the features, a support vector machine is applied to the preprocessed data with regard to clinically relevant diagnostic tasks. Finally, the classification performance of the learned nonlinear models is evaluated to allow for conclusions to be drawn about suitability of features and data resulting from different examination paradigms. As a reference, a second feature set is determined which corresponds to more traditional voice analysis approaches.

Results: For the first time an automatic classification of healthy and pathological voices could be obtained by analyzing the vibratory patterns of vocal folds using phonovibrograms (PVGs). An average classification accuracy of approximately 81% was achieved for 2-class discrimination with PVG features. This exceeds the results obtained through traditional voice analysis features. Furthermore, a relevant influence of phonation frequency on classification accuracy was substantiated by the clinical HS data.

Conclusion: The PVG feature extraction and classification approach can be assessed as being promising with regard to the diagnosis of functional voice disorders. The obtained results indicate that an objective analysis of dysfunctional vocal fold vibration can be achieved with considerably high accuracy. Moreover, the PVG classification method holds a lot of potential when it comes to the clinical assessment of voice pathologies in general, as the diagnostic support can be provided to the voice clinician in a timely and reliable manner. Due to the observed interdependency between phonation frequency and classification accuracy, in future comparative studies of HS recordings of oscillating vocal folds homogeneous frequencies should be taken into account during examination.

1. Introduction

Discriminating healthy and pathological vocal fold vibration patterns is essential to the clinical diagnosis of voice functioning, which is usually carried out by speech/voice pathologists, phoniatricians, and vocologists. A common quality criterion for a normal voice is the degree of symmetry and regularity of the oscillating vocal folds [1,2]. In order to clinically assess these dynamic aspects, the vocal folds' movement patterns need to be captured during phonation. As the fundamental frequency of oscillating vocal folds ranges from approximately 80 to 300 Hz (given the habitual pitch speaking level of an adult), the temporal resolution of conventional visual recording systems does not suffice to capture the details of the underlying vibratory patterns. Hence, a variety of specialized technologies have been developed to allow for the observation of the rapidly moving vocal folds [3–7]. However, the most common clinically used examination approach, namely stroboscopy [8], shows serious diagnostic deficits, as only periodic laryngeal movements can be adequately investigated due to sampling rate restrictions [9].

Endoscopic high-speed (HS) camera systems are a state-of-the-art examination technique for the visual inspection of a patient's laryngeal dynamics [10]. In doing so, the voice clinician subjectively assesses the occurring mode of vocal fold movement and the symmetry between left and right vocal fold side. The HS technology even allows for capturing irregular movement patterns

* Corresponding author. Tel.: +49 9131 85 32602; fax: +49 9131 85 32687.
E-mail address: daniel.voigt@uk-erlangen.de (D. Voigt).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.01.001

[11], as the vocal fold oscillations are recorded with a frame rate of 4000 frames/s and more. However, the amount of recorded image data rapidly exceeds the limit of what can be evaluated in a usual clinical time-frame. Additionally, plenty of experience regarding the analysis of HS videos is needed on the part of the examiner. This is due to the fact that the human eye is much more adapted to the processing of static visual information than to moving images. Consequently, the clinical assessment of vocal fold movement as captured in HS recordings is inherently imprecise and exhibits a rather low inter- and intra-rater reliability. To overcome the limitations of subjective evaluation, a computerized video analysis is required.

The focus of this work was on HS recordings of patients with functional voice disorders. The clinical picture of this particular kind of dysphonia is quite diffuse, and as a consequence, its rating is highly subjective [12,13]. Unlike organic dysphonias (e.g. Reinke's edema, polyp) where an appropriate diagnosis can be made based almost solely on a single image of a patient's vocal folds [14], in case of functional voice disorders the diagnostic process is much more complex. This is because the corresponding vocal fold movement can only be diagnosed in the context of overall vibratory behavior, which, to date, is only captured in an adequate manner by HS examination. Moreover, during the diagnostic process other factors like muscle tension and mental condition of the patient should be regarded as well [15–17]. Accordingly, there is significant demand for an objective method to differentiate between functional voice disorders and healthy movement patterns.

As yet, a lot of promising approaches have been introduced to facilitate the objective analysis of HS recordings [18–21]. For the most part, these methods focus on the segmentation and analysis of the one-dimensional glottal area signal over time. Another approach consists in extracting the position of individual vocal fold points and observing their displacements in regard to a fixed line. Furthermore, to quantify asymmetries and irregularities of the vibrations, the parameters of biomechanical multi-mass-models are automatically fitted to detected vocal fold deflections and used as an indicator for pathological behavior. Lately, Nyquist plots [22], Hilbert transform-based approaches [23], and methods from nonlinear systems analysis [24] are also applied. Still, all mentioned methods lack the ability to analyze the complete oscillation pattern of the vocal folds at once.

Phonovibrography, a recently developed visualization technique, is a fast and clinically evaluated method for capturing the whole spatio-temporal pattern of activity along the entire length of the vocal folds [25]. The deflections of the vocal folds contained in the recorded HS videos are extracted and can be compactly depicted in a single color-coded image, denoted as phonovibrogram (PVG) [26]. The PVG gives insight into the vibratory information of both vocal folds simultaneously. Thus, occurring vocal fold movement irregularities can be identified quite intuitively by visual inspection. The PVG allows a comprehensive analysis of the underlying two-dimensional laryngeal dynamics [27]. Besides being a valuable diagnostic tool, a PVG can also be taken as a basis for extracting a set of numerical features. These features objectively describe the characteristics of the vocal fold vibration patterns. Hence, they can be used for automatically distinguishing between pathological and healthy behavior. In contrast to the more traditional approaches, where features are derived from the one-dimensional glottal signal, the PVG allows for the extraction of more extensive two-dimensional feature descriptions.

In this work, the HS recordings of a collective of 75 healthy and pathological female subjects has been analyzed with a novel method for describing the spatio-temporal PVG dynamics and classifying them according to normal and dysfunctional vocal fold movement patterns. The obtained PVG features were analyzed using a nonlinear support vector machine (SVM) approach [28–30] in combination with an evolutionary parameter optimization. Subsequent to building a model of the data, new examples were classified according to different binary classification tasks which are relevant to the identification of functional disorders. As a reference, the same classification tasks were also carried out on a set of traditional glottal features. With the resulting cross-validated classification accuracies, the different feature sets were compared to each other regarding their ability to describe vocal fold movement.

2. Data

The vocal fold movements of $n = 75$ patients were captured with state-of-the-art HS recording technique. The diagnoses that subsequently served as a gold standard for classification and evaluation were made by clinically experienced physicians and speech therapists according to the basic protocol of voice pathology assessment of the European Laryngological Society [2]. At this, five different examination steps were accomplished consecutively for each individual: auditory-perceptual assessment, videolaryngoscopic examination, aerodynamic and acoustic analysis, and not least, self-rating of the patient.

In this manner, a population of $n = 50$ women with a diagnosed functional voice disorder was obtained. This clinical picture is also referred to as primary muscle tension dysphonia, and is diagnosed in case of dysphonia given normal vocal fold morphology and motion, and the absence of organic pathological conditions [15]. The considered population included $n = 25$ cases with a hyperfunctional and $n = 25$ cases with a hypofunctional disorder. The distinction between these two dysfunctional types is clinically made based on the patient's overall muscle tension status, the amount of laryngeal muscle tension applied during phonation, the varying degree of hoarseness during crescendo [16] and abnormal laryngeal posture during connected speech [15]. Furthermore, as a reference population for normal voices, the laryngeal dynamics of $n = 25$ female candidate speech therapists were recorded. These healthy individuals exhibited no voice irregularities. Table 1 sums up the age distribution and two acoustic perturbation measures of the considered population.

3. Methods

3.1. High-speed videos

The laryngeal images were recorded with a digital HS camera system, model *Wolf High Speed Endocam 5542*. The camera sensor takes images at a frame rate of 4000 frames/s and a spatial resolution of 256 × 256 image points with 8-bit grayscale (see Fig. 1 for example pictures). The sensor receives the optical images of a patient's vibrating vocal folds through a rigid 90° endoscope (*Wolf Endoscope 8454*) mounted in front of the camera. A typical

Table 1
Age distribution and acoustic perturbation measures of the normal and dysphonic population.

	Diagnoses		
	Healthy	Hyper	Hypo
Number of individuals	25	25	25
Average age (in years)	19.9 ± 1.3	42.7 ± 14.8	40.4 ± 20.8
Jitter [a] (%)	0.30 ± 0.15	0.34 ± 0.23	0.36 ± 0.22
Shimmer [a] (%)	2.60 ± 1.15	2.77 ± 1.29	3.07 ± 1.17

[a] For the determination of jitter and shimmer only reliable values $<5\%$ were considered [31].

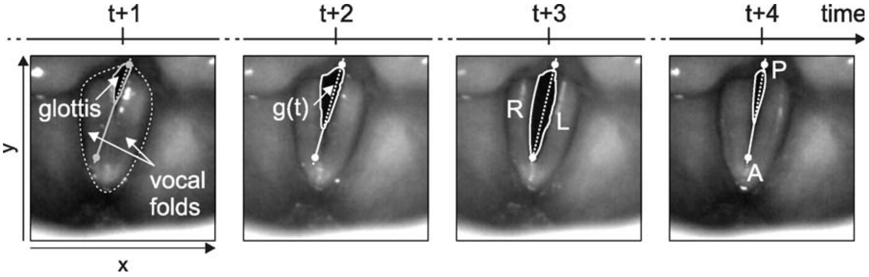

Fig. 1. Excerpt of four single HS video frames illustrating the vocal fold movement and the segmented vocal fold edges. In the first frame the position of the vocal folds and the resulting glottis is shown. The second and the third frame illustrate the glottal axis and the edges of the left and right vocal fold side, respectively. In the last frame the position of the anterior and posterior end is depicted.

record is a sequence of 2000–6000 frames which corresponds to a total duration of 0.5–1.5 s. During examination the patient was asked to phonate the sustained vowel /a/.

3.2. Image segmentation and visualization

In order to identify the position of the vocal folds over time, the recorded HS videos were segmented to detect the glottal area $a(t)$. In Fig. 1 the glottis can be perceived as the expanding and contracting opening in the center of the images. As a result, the position of the left and right vocal fold was obtained for each frame of the video using the glottal axis $g(t)$ as a splitting point for the segmented edge [26].

To obtain a compact representation of the determined vocal fold edges over time, for each HS recording the according PVG was computed. A PVG encodes the deflections of both vocal folds in regard to the glottal axis as graded color intensities—the brighter the color of a certain PVG pixel, the farther away the corresponding vocal fold point from the glottal axis $g(t)$. For a detailed description of the PVG generation process refer to [26].

The vocal folds' spatio-temporal movement patterns are transformed into geometrical shapes using the PVG (see Fig. 2 for an example). Thus, they can be readily employed by a voice clinician to quickly gain information on the overall laryngeal dynamics of a patient. So, for example, the PVG shown in Fig. 2 reveals a triangular vocal fold movement pattern which is quite stable and symmetrical—indicating healthy laryngeal dynamics. Moreover, the PVG provides the opportunity to describe vocal fold movement in a quantitative manner. To this end, the derived PVG data matrix was subsequently analyzed to capture the underlying vocal fold movement patterns by extracting a set of descriptive features.

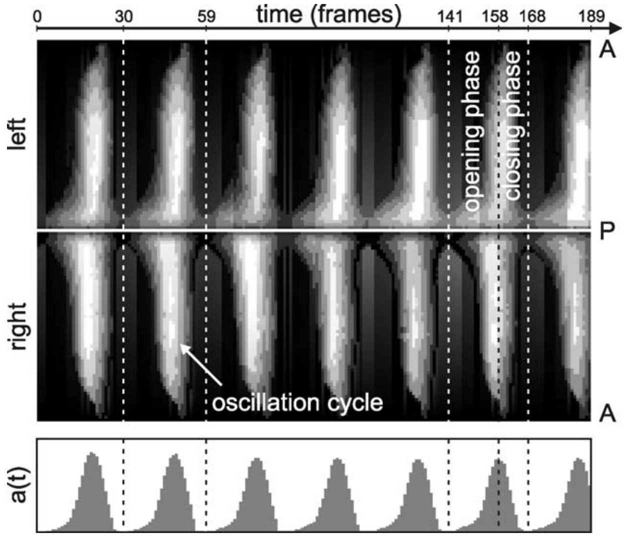

Fig. 2. PVG representation of vocal fold dynamics of a healthy individual ($f_0 = 144.9$ Hz). The movement of both vocal fold sides can be compared to each other, as the deflections of a certain frame are depicted as color-coded pixels in a single PVG column. Usually a PVG consists of three distinct colors (red, black, blue), but the black-and-white representation shown here suffices to give an idea of its basic structure: while bright sections visualize large distances from the glottal axis, dark pixels represent proximity to the midline.

3.3. Feature extraction

3.3.1. Cycle detection

Vocal fold vibration comprises recurring movement patterns which bear a certain degree of similarity to each other (see the shape of the opening and closing phases in Fig. 2). Hence, the first step towards feature extraction consisted in automatically detecting the individual oscillation cycles within the continuous PVG (see dashed white lines in Fig. 2). The boundaries of all captured cycles were determined by applying a peak-picking approach in the image domain [26]. Because the first and the last vocal fold oscillation cycle may have already been truncated in the original HS recording (e.g. rightmost cycle in Fig. 2), the two outermost cycles found by the algorithm were withheld from further analysis. As a result, a robust approximation of the points in time when an individual oscillation cycle starts and ends was obtained. The amount of frames used for cycle detection was set to $K = 1000$ for all included PVGs, respectively.

After boundary detection, all found cycles were normalized to a standard width of $L = 256$ frames to reduce the effects of differing phonation frequencies and varying endoscope positioning in the oral cavity (see Fig. 3). Thus, for the PVG of a single vocal fold side a set of normalized oscillation cycles $C^{\alpha,i}$ (with $\alpha \in [\mathrm{Left}, \mathrm{Right}]$; $i \in \{1, \ldots, I^\alpha\}$ and I^α denoting the total amount of cycles found) was obtained. The individual deflection values of the cycle are accessed via function $c^{\alpha,i}(x, y)$ with $x, y \in \{1, \ldots, 256\}$. Henceforth, a single row $c_y^{\alpha,i}(x)$ with $x \in \{1, \ldots, 256\}$ of a normalized cycle will be referred to as a trajectory.

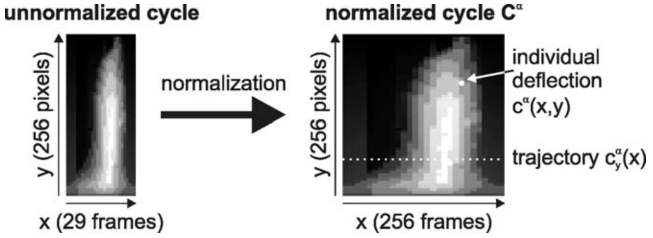
unnormalized cycle
normalized cycle C^α
y (256 pixels)
x (29 frames)
normalization
y (256 pixels)
individual deflection $c^\alpha(x,y)$
trajectory $c_y^\alpha(x)$
x (256 frames)

Fig. 3. Normalization of a detected PVG cycle. The width alignment procedure is performed for all PVG cycles of both vocal fold sides. These normalized cycles provide a basis for extracting quantitative shape information from the data.

3.3.2. PVG contour features

The shape information contained in the oscillation cycles was captured by means of novel PVG contour features. Firstly, the points in time with maximum and minimum deflection were identified for all trajectories of the oscillation cycles:

$$\begin{aligned} d_{y,max} &:= \arg\max_x \; c_y(x) \\ d_{y,min} &:= \arg\min_x \; c_y(x) \end{aligned} \qquad \text{for } C^{\alpha,i}, \quad \forall \alpha, i, y. \tag{1}$$

In this manner, a relative contour threshold

$$t_{y,h} = c_y(d_{y,min}) + h \cdot (c_y(d_{y,max}) - c_y(d_{y,min})), \quad \text{with } h \in \{0, \ldots, 1\} \tag{2}$$

was derived from the deflection values of the two detected points in time.

Secondly, starting from position $d_{y,max}$, the deflection values of each trajectory were traced along both temporal directions towards the adjacent closed states of the vocal folds (see Fig. 4c). This bilateral descent step was performed until the first point in time was reached whose deflection value was equal or below the relative contour threshold, respectively. Parameter h of the contour threshold was set to 0.5 which corresponds to the point in time when a vocal fold point has reached 50% of its displacement between minimum and maximum deflection:

$$\begin{aligned} d_{y,0.5,O} &:= \arg_x \; (c_y(x) \le t_{y,0.5}), \quad \text{with } x < d_{y,max} \\ d_{y,0.5,C} &:= \arg_x \; (c_y(x) \le t_{y,0.5}), \quad \text{with } x > d_{y,max}. \end{aligned} \tag{3}$$

Thus, the points in time when the vocal folds are located at the state of half deflection were determined, yielding two distinct contour lines in the opening and closing phase of the cycle. Some example PVG cycles and the corresponding contours are shown in Fig. 4.

The resulting 256 individual points of a contour line were subsumed by averaging over predefined contour intervals [26]. For this purpose the opening and closing contour line were divided into 16 intervals, yielding two averaged contours $O_{0.5}^{\alpha,i}$ and $C_{0.5}^{\alpha,i}$. So an individual contour point included an x-position along the timeline and a z-position quantifying the interval's mean deflection. Based on the two-dimensional PVG signal, the spatio-temporal behavior along the entire vocal fold length was described in terms of numerical features.

Furthermore, to relate the vibration characteristics of both vocal folds to each other, proportions between contour features of the left and the right side were computed:

$$\begin{aligned} P_{O,0.5}^i &= O_{0.5}^{\text{Left},i} / O_{0.5}^{\text{Right},i} \\ P_{C,0.5}^i &= C_{0.5}^{\text{Left},i} / C_{0.5}^{\text{Right},i} \end{aligned} \qquad \forall i. \tag{4}$$

Additionally, as another characterization of symmetries between left and right vocal fold side, the contours' Euclidian distances were computed as follows:

$$\begin{aligned} D_{O,0.5}^i &= \left\| O_{0.5}^{\text{Left},i} - C_{0.5}^{\text{Right},i} \right\|_2 \\ D_{C,0.5}^i &= \left\| C_{0.5}^{\text{Left},i} - C_{0.5}^{\text{Right},i} \right\|_2 \end{aligned} \qquad \forall i. \tag{5}$$

Thus, a set of supplementary PVG features was obtained describing bilateral properties of the vocal folds.

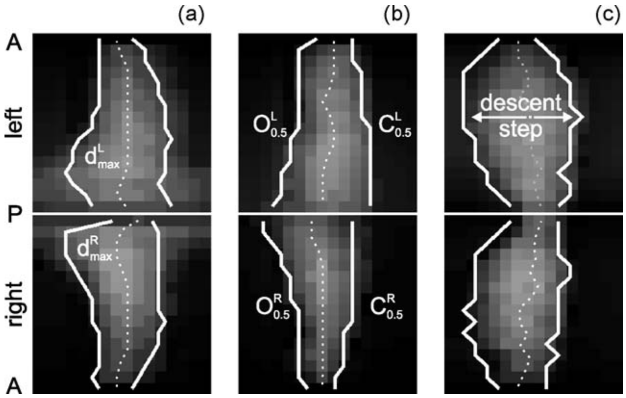

Fig. 4. Contour lines of different normalized vocal fold oscillation cycles of three individuals. While the first cycle (a) is derived from a healthy case, the remaining two represent functional voice disorders (b: hyper-, c: hypofunctional). In addition, for each trajectory the point in time with maximum deflection is shown (dashed line) which serves as a starting point for contour detection.

3.3.3. Reference glottal features

In order to assess the descriptive power of the new PVG-based features, an additional reference feature set was computed for all HS movies. Employing the one-dimensional glottal signal $a(t)$ [22] and the corresponding movement cycles, a set of glottal parameters was derived which until now is part of the standard repertoire for describing vocal fold movement. The features capture the length of the cycles' individual phases, the stability of their glottal deflection modes and their total duration over time. Thus, the vocal folds' movement patterns were described at the level of occurring glottal changes. As these glottal parameters have already been widely used in the voice analysis literature (e.g. [19,32,33]), in the following, they will be referred to as traditional features.

The following parameters were computed from the glottal signal:

- the open quotient Q_o, which quantifies the proportion of time the glottis is open during an oscillation cycle [34],
- the speed quotient Q_s, which represents the temporal proportion between the opening and the closing phase of a cycle [32],
- the glottal insufficiency Q_g, which captures the relation between a cycle's minimum and maximum glottal opening [33],
- the time periodicity index I_{tp}, which describes the temporal stability of the cycle duration [19],
- and the amplitude periodicity index I_{ap}, which measures a vocal fold's deflection stability [19].

To determine the two indices I_{tp} and I_{ap}, pairs of consecutive oscillation cycles were related to each other. Thus, they can be regarded as being equivalent to the voice quality measures jitter and shimmer which are commonly used in the quantitative evaluation of speech (e.g. in [35]).

3.3.4. Feature aggregation

To integrate the individual cycle descriptions and to achieve temporal abstraction, for each PVG and glottal feature the mean and the standard deviation were computed over all cycles. The mean was primarily used for aggregating the individual cycles' feature values and capturing the average vocal fold movement. The standard deviation measured a feature's variability over time, and thus, represented the vocal fold's dynamic changes. An overview of the feature sets used in this study is given in Table 2.

3.4. Data analysis

The computed features subsume the deflections of a subject's vocal folds and their positional changes over time at an abstract level. While feature set F_1 describes both vocal folds' behavior in terms of the two-dimensional PVG signal, feature set F_2 captures the changes of the one-dimensional glottal signal. The different

Table 2
Feature sets derived from the HS recordings of the patients' vocal fold movements. While F_1 contains all the features capturing shape and symmetry of the normalized PVG cycles, F_2 comprises the reference features derived from the glottal signal. Feature set F_3 is obtained by joining F_1 and F_2, yielding another reference description for the identification of beneficial feature combinations.

Feature set	Contained features	Underlying signal
F_1: PVG features	$\bar{O}^{\alpha}_{0.5}, \sigma(O^{\alpha}_{0.5}), \bar{C}^{\alpha}_{0.5}, \sigma(C^{\alpha}_{0.5}), \bar{P}_{O,0.5}, \sigma(P_{O,0.5}), \bar{P}_{C,0.5}, \sigma(P_{C,0.5}), \bar{D}_{O,0.5}, \sigma(D_{O,0.5}), \bar{D}_{C,0.5}, \sigma(D_{C,0.5})$	PVG (2d)
F_2: traditional features	$\bar{Q}_o, \sigma(Q_o), \bar{Q}_s, \sigma(Q_s), \bar{Q}_g, \sigma(Q_g), \bar{I}_{tp}, \sigma(I_{tp}), \bar{I}_{ap}, \sigma(\bar{I}_{ap})$	Glottis (1d)
F_3: combined features	$F_1 \cup F_2$	PVG (2d) + glottis (1d)

Table 3
Considered training sets, classification tasks and their respective class distributions. In classification task C_1 the * indicates merged and undersampled classes.

Classification tasks	Training sets		
	S_1: mixed frequency data	S_2: homogeneous frequency interval	S_3: inhomogeneous frequency interval
C_1: healthy vs. pathological (hyper ∪ hypo)	25–25*	15–15*	15–15*
C_2: hyper vs. hypo	25–25	15–15	15–15
C_3: healthy vs. hyper	25–25	15–15	15–15
C_4: healthy vs. hypo	25–25	15–15	15–15
C_5: healthy vs. hyper vs. hypo	25–25–25	15–15–15	15–15–15

Table 4
Different frequency interval classes derived from the 75 examples of training set S_1 and the resulting class distributions.

Classification tasks	S_1: all frequency data
C_6: high vs. low	37–38
C_7: high vs. medium vs. low	25–25–25
C_8: high vs. upper medium vs. lower medium vs. low	18–19–19–19

quantitative description approaches of the underlying vibratory patterns were used to identify disease-specific particularities in feature space. The discovered structures allow a potential mapping from HS recordings of newly examined patients to certain vocal fold disease classes.

3.4.1. Class structure

In order to analyze the features in terms of class membership and to build models of vocal fold dysfunction, they were integrated into describing feature vectors as shown in Table 2. In addition, a distinct class label indicating the clinical diagnosis of the HS video was attached to all feature vectors. The obtained HS data descriptions were pooled in different combinations to provide adequate training sets S_{1-3} for the considered classification tasks C_{1-5}. The overall set of 25 healthy, 25 hyper-, and 25 hypofunctional examples was partitioned as shown in Table 3.

In standard clinical examination situations phonation frequency is usually determined by the patient's individual voice characteristics. Moreover, it is oftentimes influenced by the vocal dysfunction in question. Hence, an unbalanced frequency distribution of classes will be obtained. The frequency distribution of the three classes included in training set S_1 is shown in Fig. 5. As expected, the underlying frequencies are unbalanced: while the healthy examples show a bias towards the upper frequency spectrum ($\bar{f}_{\text{healthy}} = 275.6 \pm 42.3$ Hz), the functional examples are located mostly in the lower spectrum ($\bar{f}_{\text{hyper}} = 243.0 \pm 54.6$ Hz and $\bar{f}_{\text{hypo}} = 242.8 \pm 39.0$ Hz).

As indicated in the literature [36,37], oscillation frequency of the vocal folds during voice production affects the outcome of the perturbation measurement process. This suggests that the features derived from the HS videos may also be influenced by phonation frequency, and as a consequence, may have an effect on the classification tasks at hand. To verify this assumption, from the available set of 75 learning examples a subset S_2 consisting of 15 healthy, 15 hyper-, and 15 hypofunctional examples was selected. All examples were located in the homogeneous frequency interval $I = [199, 281]$ Hz. Additionally, a complementary training set S_3 was drawn from the data, comprising the remaining examples outside interval I plus a small overlap to retain class balance. In doing so, for S_2 and S_3 the potential influence of oscillation frequency was minimized and maximized, respectively. Moreover, as an additional method to assess the relation between phonation frequency and classification outcome further class structures were examined incorporating different frequency intervals (see Table 4). For this purpose, the overall training set S_1 was subdivided into $n = 2, \ldots, 4$ frequency classes, while entirely disregarding the underlying diagnoses. Hence, the resulting classification tasks consisted in building models of frequency membership. With the two training sets $S_{2,3}$ and the classification tasks C_{6-8} the influence of the frequency range on classification accuracy was evaluated.

Furthermore, to obtain balanced class distributions for classification task C_1 undersampling was performed (marked by *) [38]. By randomly selecting 50% of the data of the two remaining classes (i.e. hyper- and hypofunctional examples), respectively, and merging them into one counter-class (i.e. pathological) equal class sizes were established. To compensate for random side effects caused by undersampling multiple sampling runs were performed, each resulting in a different training set configuration. The individual results were averaged to obtain reliable estimates of classification accuracy.

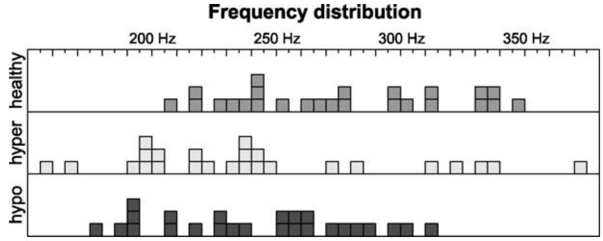

Fig. 5. Distribution of the 75 training examples of dataset S_1 regarding frequency and class membership. The boxes' position indicates the respective dataset's fundamental frequency.

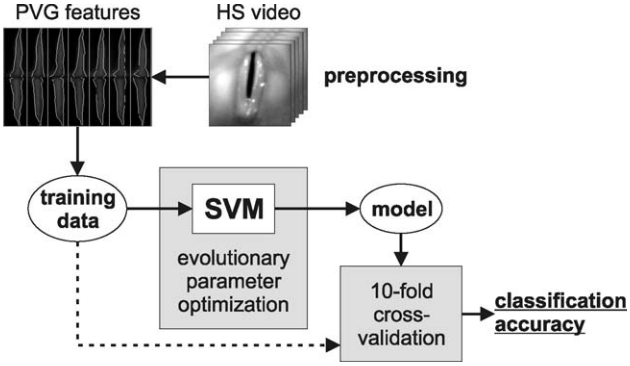

Fig. 6. Workflow diagram of the overall classification process employed to assess the performance of the different feature sets.

3.4.2. Machine learning and evaluation

The derived datasets were subsequently used as an input to an inductive learning scheme which built model descriptions of the data. To this end, a SVM with a Gaussian radial basis function kernel was applied to the training sets [28,39]. On the basis of the resulting nonlinear SVM models the most likely class label was assigned to an unseen feature vector which was withheld from the process of model building.

Appropriate SVM parameters were determined by an evolutionary strategy optimization procedure [29,40]. The parameter space of SVM cost parameter C and the width γ of the radial basis function kernel was automatically searched in order to obtain best classification results [41]. The models' classification accuracy was evaluated via 10-fold cross-validation with stratification [42]. In this manner, the individual results were compared to each other, yielding the best performing classification task and feature set. The applied learning scheme is illustrated in Fig. 6.

4. Results

4.1. Mixed frequency data

In Fig. 7 the results of classification tasks C_{1-5} are shown for feature sets F_{1-3}. The respective models were trained using the overall training set S_1 (without considering the underlying frequency distribution).

The average classification result obtained by PVG features F_1 exceed the one of glottal features F_2 with high significance ($78.5 \pm 15.8\%$ vs. $73.5 \pm 15.7\%$, $p = 0.009$). Hence, vocal fold movement patterns described by PVG features can be distinguished more readily in terms of healthy and pathological behavior than through the use of glottal features. A combination of both feature description approaches does not improve classification performance significantly (F_3: $77.6 \pm 14.6\%$, $p > 0.05$), and hence, is not further considered in the results. Classification tasks C_3 and C_4 yielded better average accuracies than C_1 (77.5% and 76.9% vs. 74.4%). Thus, for classification purposes, it is beneficial to consider the different types of functional voice disorders individually than to merge them into one class. The results' high standard deviations can be ascribed to the relatively small amount of evaluation data available in the individual folds of cross-validation ($\overline{n} = 7.5$). So, depending on the performed split of the training data, classification results with high variability are obtained.

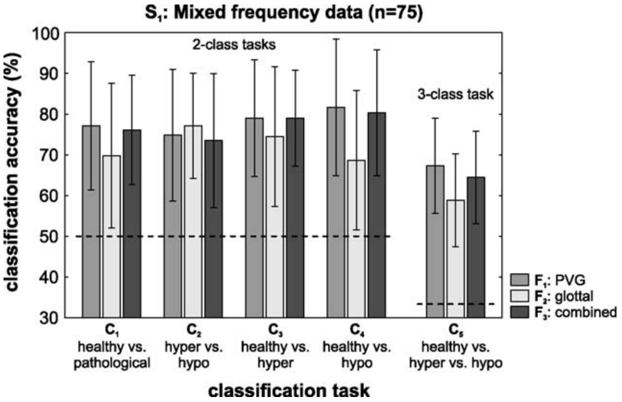

Fig. 7. Classification results obtained through the total amount of 75 datasets for training and evaluation. The two dashed horizontal lines at 50% and 33.3% denote the accuracy levels which can be reached by classifying all examples as the majority class, respectively. The error bars indicate the results' standard deviations arising from cross-validation.

4.2. Frequency classification

Phonation frequency exerts a distinct influence on the measurement of vocal fold movement [36,37]. Therefore, the oscillation patterns are possibly subject to change due to frequency alterations. According to this, a subject's vocal fold movement pattern ought to be automatically assigned to its appropriate frequency interval with high accuracy solely based on the information of its spatio-temporal shape. This means that side effects caused by differing phonation frequencies may have a stronger influence on the classification results obtained via training set S_1 than the disturbed movement pattern of the disease itself. Thus, it must be assumed that the results presented in Fig. 7 are potentially biased by frequency outcomes. In order to assess the actual frequency effect on the data, the examples were arranged into a new class structure by considering only the membership to certain frequency intervals as class information. In Fig. 8 the classification results obtained for the according classification tasks C_{6-8} are shown.

The classification results of the frequency classes outperform the results of the healthy/dysfunctional classification presented in Fig. 7 (2-class average: $81.1 \pm 12.0\%$ vs. $76.0 \pm 23.6\%$). Consequently, the data of the healthy and pathological examples need to be analyzed explicitly taking into consideration the frequency

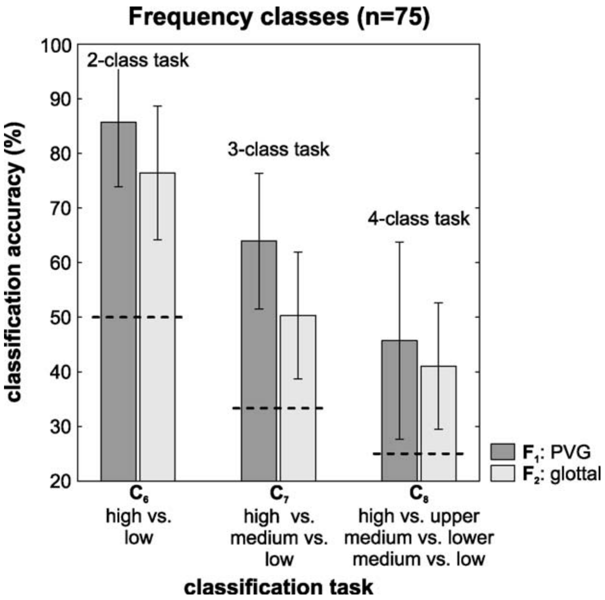

Fig. 8. Classification results obtained by grouping the total amount of 75 training examples into classes educed from different frequency intervals. The dashed horizontal lines mark the baseline accuracy achievable by classifying all cases as the majority class.

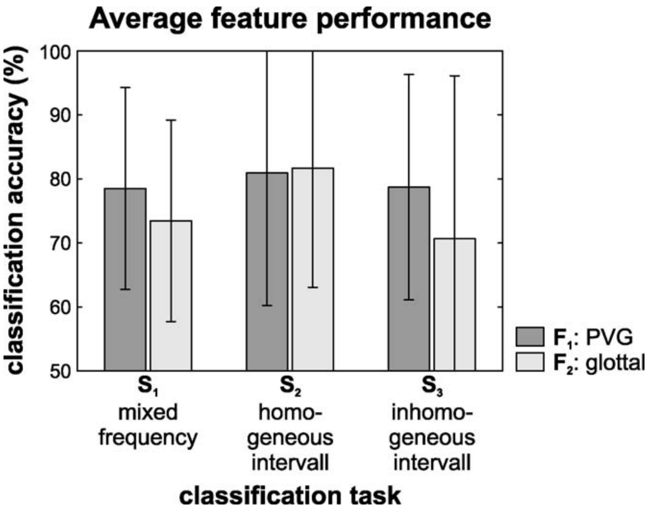

Fig. 9. Average 2-class classification results obtained with PVG feature set F_1 and glottal feature set F_2, respectively. In building the models different training sets were applied varying in terms of considered frequency distribution.

effect outlined above. Furthermore, the PVG feature set F_1 yields better results than the glottal features F_2 throughout all frequency classification tasks.

4.3. Different frequency intervals

To assess the influence of varying phonation frequencies on the discrimination of healthy and pathological examples, the classification performance of F_1 and F_2 were examined individually for the following training sets: all data without regarding frequency at all (S_1), a subset with weakened frequency influence (S_2), and a subset with intensified frequency influence (S_3). In Fig. 9 the averaged 2-class results of the different frequency intervals are contrasted with each other. For averaging only the results of C_{2-4} were considered, as the pathological class underlying C_1 contains examples from both functional classes and its inclusion would yield an overoptimistic performance estimate.

Classification results of training set S_1 in Fig. 9 are significantly exceeded by the ones obtained through subset S_2 (F_1: 78.5% vs. 80.9%; F_2: 73.5% vs. 81.7%; $p = 0.015$). Thus, healthy and pathological vocal fold movements can be differentiated with higher reliability if homogeneous frequency intervals are examined. Classification results obtained through feature set F_1 are relatively stable for all three examined training sets, resulting in a standard deviation of 1.3% across frequency intervals. In homogenized subset S_2 PVG features F_1 and glottal features F_2 perform equally well in average. Despite exhibiting a particular improvement between training sets S_1 and S_2 (+8.2%), the relative decline of F_2 in the S_3 results is much stronger than for F_1 (−11.0% vs. −2.2%). At this, the error bar for F_2 using inhomogeneous training set S_3 even falls below the baseline accuracy of 50%. Thus, in total the classification results of glottal features F_2 are more sensitive to occurring frequency influences than PVG features F_1 (standard deviation across training sets: 5.7%). Due to this fact, PVG features were found to be more suitable to describe laryngeal dynamics under varying frequency distributions than glottal features.

4.4. Discussion

An important finding of this study was that PVG feature set F_1 outperforms glottal feature set F_2. Thus, the benefit of considering the overall vocal fold movement pattern over time as represented in PVGs could be shown with high statistical significance. To focus only on feature descriptions of the changing glottal area yields suboptimal classification results. The reason for this can be seen in the fact that F_2 describes the laryngeal dynamics only at a rather coarse level, and as a consequence, does not capture the necessary details of vocal fold vibration. Moreover, it lacks the ability to distinguish the two vocal fold sides, and thus, to measure left-right asymmetries. Due to this, classification results obtained through glottal features F_2 are clearly surpassed by PVG features F_1 in training sets S_1 and S_3 (see Fig. 9). In addition, F_2 possesses a distinct sensitivity to phonation frequency, as the quite large variability of the obtained results reveals. The combination of both feature set approaches as represented in F_3 achieves no classification improvement.

The subsumption of the two types of functional voice disorders into one class (classification task C_1) has been shown to be obstructive in terms of differentiating between healthy and pathological vocal fold movement. Therefore, it is beneficial to analyze and model hyper- and hypofunctional diagnoses individually (classification tasks C_3 and C_4), as the corresponding classification results are better than in the class pooling approach. In the hyperfunctional case the patient's muscular tension tends to be increased, whereas in the hypofunctional case it is considerably reduced [16]. Since healthy laryngeal dynamics are essentially situated between these two diseased states, the merging of both diagnoses into one class results in an increased overlap of pathological and healthy examples. Accordingly, the identification of adequate class boundaries is aggravated by merging functional diagnoses.

From the fact that classification accuracy could be improved by focusing on homogeneous frequency data it may be concluded that a constant phonation frequency should be seeked for during examination. However, a mixture of phonation frequencies is a more realistic assumption in a standard clinical setting, as habitual pitch phonation plays an important part in making proper diagnoses of voice disorders. Hence, under these clinical considerations the proposed PVG features have been shown to be more reliable for the identification of dysfunctional behavior than the set of glottal parameters. The frequency-dependent classification performance can be accounted for in practice by including supplementary features capturing relevant phonation frequency information which facilitates the discrimination of vocal fold movement patterns. So the presented approach shows a lot of promise in regard to the characterization and classification of functional voice disorders. For the analysis of organic disorders, the PVG features are actually expected to perform at least equally well in terms of distinguishing healthy and pathological cases. So, for instance, for the identification of vocal fold paresis it is of particular importance to capture the left-right asymmetries—a property which cannot be described with the glottal signal. Notwithstanding, in future studies the HS recordings of the patients should be made under more controlled phonation conditions in order to allow for a systematic analysis of the frequency influence.

As the PVG method presented in this work is a very novel approach to the classification of moving vocal folds, only relatively few comparable results can be found in the literature. Some works focus on the extraction of features from the glottal signal and the subsequent identification of normal and dysphonic feature ranges (e.g. [19]). But most commonly, the voice signal of a patient is recorded and analyzed using acoustic features. So, for example, in [43] a classification accuracy of 80% is reported by applying artificial neural networks to training data derived from the audio signal of 120 individuals. By analyzing voice signal features Awan et al. achieved an accuracy of 74.6% for the classification of healthy and functional examples into voice quality classes using stepwise discriminant analysis [44]. A sophisticated method from nonlinear dynamical systems theory combined with quadratic discriminant analysis yields results as good as 91.8% [45]. As a matter of fact, other studies employing logistic regression analysis [46] and kernel principal component analysis [47] report on errorless classification.

In the context of these works, the following facts should be kept in mind. Firstly, in contrast to organic voice disorders, the clinical picture of functional voice disorders is pretty vague and cannot be easily covered by a distinct set of diagnostic rules which apply under any clinical circumstances (see [16,12,13]). As a result, the subjective assessment of the symptoms can lead to quite low inter-rater reliability. So the PVG description approach presented here is a promising step towards the objectification of clinical criteria underlying the diagnostic process of functional voice disorders. By means of numerical PVG features extracted from laryngeal HS videos a level of inter-individual comparability is achieved which effectively enables the realization of evidence-based medicine in the field of clinical voice diagnosis.

Secondly, only vibratory vocal fold information was utilized. The results obtained from the classification of PVG cycle shape features may possibly be improved by incorporating additional information characterizing a patient's vocal fold behavior, phonation frequency or acoustic outcome. So, for example, in further studies the PVG features can be combined with parameters derived from the recorded voice signal [48] or features capturing the transient oscillations during phonation onset [49].

Thirdly, another essential aspect that needs to be regarded is the existence of the superimposed phonation frequency effect. Its influence on classification accuracy of vocal fold movement patterns was substantiated through the improving results shown in Fig. 9. Even though the frequency influence on the data could be reduced to a certain extent by narrowing down the bandwidth of the analyzed HS videos to 82 Hz in the homogeneous frequency interval, its effect is still existent in the data. As a consequence, the results obtained from training set S_2 may still be affected by the frequency influence.

5. Conclusion

A novel method for capturing the movement of vocal folds was presented which can be used to discriminate healthy and pathological vibration modes. To this end, the laryngeal dynamics of a collective of individuals with normal voices and diagnosed functional voice disorders were recorded with state-of-the-art HS examination technique. The resulting videos were automatically analyzed and transformed into PVGs. Subsequently, the movement patterns contained in these PVG representation were described by a set of PVG features capturing spatio-temporal shape and symmetry of vocal fold oscillation. As a way to evaluate the descriptiveness of the derived PVG features, another set of traditional glottal parameters was determined for the HS data. These two feature sets were used as a basis for building nonlinear models of the healthy and pathological examples by employing a SVM. Thus, different classification tasks that are relevant to the diagnosis of functional voice disorders were analyzed. This allowed to draw conclusions with respect to the adequacy of the feature sets and the general classification accuracy.

The features computed from PVGs are more suitable for the differentiation of voice disorders than the traditional glottal parameters. The average classification results based on PVG features yield considerably better results throughout all considered learning tasks and exhibit higher stability under different clinical conditions. In average, a classification accuracy of 81% was obtained for 2-class tasks concerning the identification of functional dysphonia. This is very promising given the vague clinical picture of the disease and its difficult subjective diagnosis. A further finding of this study is that the choice of the phonation frequency plays an important role in the process of discriminating healthy and pathological behavior, and thus, needs to be especially considered during analysis. By and large, the presented approach to combine knowledge-based feature extraction techniques with methods from machine learning in order to develop objective medical decision support systems can be regarded as being successful and should be refined in the future.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (DFG) grants no. LO1413/2 1-3 and EY15/11 3-4. The authors would like to thank the High Performance Computing Group at Regionales Rechenzentrum Erlangen (RRZE) for providing computational resources.

References

[1] Hoppe U. Mechanisms of hoarseness—visualization and interpretation by means of nonlinear dynamics. Aachen, Germany: Shaker; 2001.

[2] Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. guideline elaborated by the committee on phoniatrics of the European laryngological society (els). Eur Arch Otorhinolaryngol 2001;258(February 2):77–82.

[3] van Michel C, Pfister KA, Luchsinger R. Electroglottography and slow-motion films of the larynx, comparison of results. Folia Phoniatr 1970;22(2):81–91.

[4] Childers DG, Larar JN. Electroglottography for laryngeal function assessment and speech analysis. IEEE Trans Biomed Eng 1984;31(December 12):807–17.

[5] Raes J, Lebrun Y, Clement P. Videostroboscopy of the larynx. Acta Otorhinolaryngol Belg 1986;40(2):421–5.

[6] Švec JG, Schutte HK. Videokymography: high-speed line scanning of vocal fold vibration. J Voice 1996;10(June 2):201–5.

[7] Wittenberg T, Tigges M, Mergell P, Eysholdt U. Functional imaging of vocal fold vibration: digital multislice high-speed kymography. J Voice 2000;14(September 3):422–42.

[8] Olthoff A, Woywod C, Kruse E. Stroboscopy versus high-speed glottography: a comparative study. Laryngoscope 2007;117(June 6):1123–6.

[9] Yumoto E. Aerodynamics, voice quality, and laryngeal image analysis of normal and pathologic voices. Curr Opin Otolaryngol Head Neck Surg 2004;12(June 3):166–73.

[10] Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatr Logop 2008;60(1):33–44.

[11] Eysholdt U, Rosanowski F, Hoppe U. Vocal fold vibration irregularities caused by different types of laryngeal asymmetry. Eur Arch Otorhinolaryngol 2003;260(September 8):412–7.

[12] Morrison MD, Nichol H, Rammage LA. Diagnostic criteria in functional dysphonia. Laryngoscope 1986;96(January 1):1–8.

[13] Altman KW, Atkinson C, Lazarus C. Current and emerging concepts in muscle tension dysphonia: a 30-month review. J Voice 2005;19(June 2):261–7.

[14] Verikas A, Gelzinis A, Bacauskiene M, Uloza V. Towards a computer-aided diagnosis system for vocal cord diseases. Artif Intell Med 2006;36(January 1): 71–84.

[15] Rosen CA, Murry T. Nomenclature of voice disorders and vocal pathology. Otolaryngol Clin North Am 2000;33(October 5):1035–46.

[16] Wendler J, Seidner W, Eysholdt U. Lehrbuch der Phoniatrie und Pädaudiologie, 4th ed., Stuttgart, Germany: Thieme; 2005.

[17] Seifert E, Kollbrunner J. Stress and distress in non-organic voice disorder. Swiss Med Wkly 2005;135(July 27/28):387–97.

[18] Švec JG, Sram F, Schutte HK. Videokymography in voice disorders: what to look for? Ann Otol Rhinol Laryngol 2007;116(March 3):172–80.

[19] Qiu Q, Schutte HK, Gu L, Yu Q. An automatic method to quantify the vibration properties of human vocal folds via videokymography. Folia Phoniatr Logop 2003;55(3):128–36.

[20] Wurzbacher T, Döllinger M, Schwarz R, Hoppe U, Eysholdt U, Lohscheller J. Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters. J Acoust Soc Am 2008;123(April 4):2324–34.

[21] Mergell P, Herzel H, Titze IR. Irregular vocal-fold vibration—high-speed observation and modeling. J Acoust Soc Am 2000;108(December 6):2996–3002.

[22] Yan Y, Ahmad K, Kunduk M, Bless D. Analysis of vocal-fold vibrations from high-speed laryngeal images using a hilbert transform-based methodology. J Voice 2005;19(June 2):161–75.

[23] Yan Y, Damrose E, Bless D. Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings. J Voice 2007;21(September 5):604–16.

[24] Zhang Y, Tao C, Jiang JJ. Parameter estimation of an asymmetric vocal-fold system from glottal area time series using chaos synchronization. Chaos 2006;16(June 2):023118.

[25] Lohscheller J, Toy H, Rosanowski F, Eysholdt U, Döllinger M. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from

endoscopic digital high-speed videos. Med Image Anal 2007;11(August 4): 400–13.

[26] Lohscheller J, Eysholdt U, Toy H, Döllinger M. Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Trans Med Imaging 2008;27(March 3):300–9.

[27] Lohscheller J, Eysholdt U. Phonovibrogram visualization of entire vocal fold dynamics. Laryngoscope 2008;118(April 4):753–8.

[28] Vapnik VN. The nature of statistical learning theory. New York, NY, USA: Springer-Verlag New York, Inc.; 1995.

[29] Yang SY, Huang Q, Li LL, Ma CY, Zhang H, Bai R. An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs. Artif Intell Med 2009;46(June 2):155–63.

[30] Asl BM, Setarehdan SK, Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intell Med 2008;44(September 1):51–64.

[31] Titze IR. Workshop on acoustic voice analysis: summary statement, 1995.

[32] Sapienza CM, Stathopoulos ET, Dromey C. Approximations of open quotient and speed quotient from glottal airflow and egg waveforms: effects of measurement criteria and sound pressure level. J Voice 1998;12(March 1):31–43.

[33] Bielamowicz S, Kapoor R, Schwartz J, Stager SV. Relationship among glottal area, static supraglottic compression, and laryngeal function studies in unilateral vocal fold paresis and paralysis. J Voice 2004;18(March 1):138–45.

[34] Henrich N, D'Alessandro C, Doval B, Castellengo M. Glottal open quotient in singing: measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. J Acoust Soc Am 2005;117(March 3 Pt 1):1417–30.

[35] Deliyski DD, Shaw HS, Evans MK, Vesselinov R. Regression tree approach to studying factors influencing acoustic voice analysis. Folia Phoniatr Logop 2006;58(4):274–88.

[36] Orlikoff RF, Baken RJ. Consideration of the relationship between the fundamental frequency of phonation and vocal jitter. Folia Phoniatr (Basel) 1990;42(1):31–40.

[37] Rasp O, Lohscheller J, Döllinger M, Eysholdt U, Hoppe U. The pitch rise paradigm: a new task for real-time endoscopy of non-stationary phonation. Folia Phoniatr Logop 2006;58(3):175–85.

[38] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6(5):429–49.

[39] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2(2):121–67.

[40] Beyer HG, Schwefel HP. Evolution strategies—a comprehensive introduction. Nat Comput 2002;1(May):3–52.

[41] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[42] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. 1995. p. 1137–45.

[43] Linder R, Albers AE, Hess M, Pöppl SJ, Schönweiler R. Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. J Voice 2008;22(March 2):155–63.

[44] Awan SN, Roy N. Acoustic prediction of voice type in women with functional dysphonia. J Voice 2005;19(June 2):268–82.

[45] Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 2007;6:23.

[46] Eadie TL, Doyle PC. Classification of dysphonic voice: acoustic and auditory-perceptual measures. J Voice 2005;19(March 1):1–14.

[47] Alvarez M, Henao R, Castellanos G, Godino JI, Orozco A. Kernel principal component analysis through time for voice disorder classification. Conf Proc IEEE Eng Med Biol Soc 2006;1:5511–4.

[48] Döllinger M, Lohscheller J, McWhorter A, Kunduk M. Variability of normal vocal fold dynamics for different vocal loading in one healthy subject investigated by phonovibrograms. J Voice 2009;23(March 2):175–81.

[49] Braunschweig T, Flaschka J, Schelhorn-Neise P, Döllinger M. High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias. Med Eng Phys 2008;30(January 1): 59–66.