34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

A Wavelet-Based Approach for a Continuous Analysis of Phonovibrograms

Jakob Unger[1], Tobias Meyer[1], Michael Doellinger[2], Dietmar J. Hecker[3], Bernhard Schick[3], Joerg Lohscheller[1]

***Abstract*— Recently, endoscopic high-speed laryngoscopy has been established for commercial use and constitutes a state-of-the-art technique to examine vocal fold dynamics. Despite overcoming many limitations of commonly applied stroboscopy it has not gained widespread clinical application, yet. A major drawback is a missing methodology of extracting valuable features to support visual assessment or computer-aided diagnosis. In this paper a compact and descriptive feature set is presented. The feature extraction routines are based on two-dimensional color graphs called phonovibrograms (PVG). These graphs contain the full spatio-temporal pattern of vocal fold dynamics and are therefore suited to derive features that comprehensively describe the vibration pattern of vocal folds. Within our approach, clinically relevant features such as glottal closure type, symmetry and periodicity are quantified in a set of 10 descriptive features. The suitability for classification tasks is shown using a clinical data set comprising 50 healthy and 50 paralytic subjects. A classification accuracy of 93.2% has been achieved.**

I. INTRODUCTION

Approximately 3 to 9 percent of the US population suffers from voice disorders [1]. Furthermore, about 25 percent of the working population depends on their voice functionality due to their jobs [2]. Consequently, computer-aided diagnosis of voice disorders has gained increased attention within the last years.

Among modern imaging techniques, high-speed videoendoscopy (HSV) constitutes the only technique that captures the true intra-cycle vibratory behavior through a full image of the vocal folds (VF) [3]. The endoscope is inserted into the oral cavity and captures up to 10,000 frames per second of the rapidly moving VFs during phonation (Fig. 1). Assessment of laryngeal function on the basis of HSV videos is time consuming and costly; one second of phonation corresponds to thousands of video frames and analysis must be performed by trained professionals.

To overcome these limitations, automated assessment of HSV video sequences has been adressed by several authors. In [4] assessment is based on the extracted glottal area waveform, specifying the time-dependent area enclosed by both VFs. Others employ two- [5], [6] or multi-mass models [7], which are adapted to the recorded VF movement. However, glottal area evaluation does not discriminate between left and right VF and due to the complexity of VF vibration it is sophisticated to define adequate model assumptions.

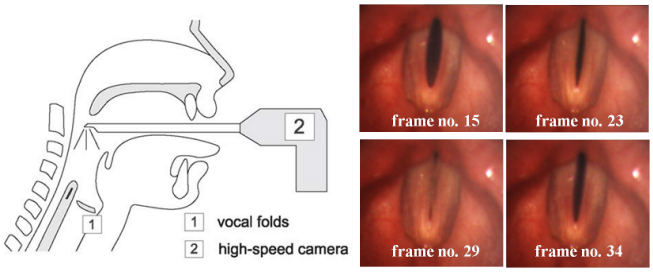

Fig. 1. Schematic representation of rigid high-speed endoscopy. The endoscope is inserted into the oral cavity and captures several thousand frames per second of the larynx during phonation.

A novel classification strategy maps the whole spatio-temporal pattern of VF vibration to single color graphs called phonovibrograms (PVG) [8]. The periodic VF movement remains in individual vibration patterns that can be analyzed by image processing methods. The PVG data matrix can be used to derive characteristic features that describe the fundamental properties of both the PVG and VF dynamics. Voigt et al. [9], [10] extracts iso-contour lines from an averaged PVG-cycle representing constant deflection states of VFs. This approach takes a lot of individual features into consideration, but nevertheless, has two main drawbacks. First, it requires quasi-periodic VF vibration since subsequent cycles have to be identified and secondly, the corresponding feature vecor comprises 448 highly correlated quantitative features. However, strong pathologies may be accompanied by a general loss of periodicity and the identification of individual cycles becomes impracticable.

In this study, we designed a wavelet-based classification system that fuses glottal closure details and phase information forming a low-dimensional feature vector without the need of identifying individual cycles. The features' discriminative power is evaluated by classifying normophonic and paralytic voices.

II. MATERIAL AND METHOD

A. Subjects and equipment

One hundred patients, fifty of them without any signs of voice disorders (20m, mean age 53.6 ± 15.00, 30w, mean age 45.5 ± 18.65) and 50 with unilateral VF paresis (22m, mean age 56.4 ± 12.81, 28w, mean age 48.8 ± 18.43) were investigated with a HS Endocam 5562 high-speed camera (Richard Wolf GmbH, Knittlingen, Germany). This camera provides a spatial resolution of 256×256 pixels and a

This work is supported by the German Research Foundation (DFG), LO-1413/2.

[1]Department of Computer Science, University of Applied Science Trier, Trier, Germany
[2]Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Erlangen, Germany
[3]Deptartment of Otorhinolaryngology, Saarland University Hospital, Homburg/Saar, Germany

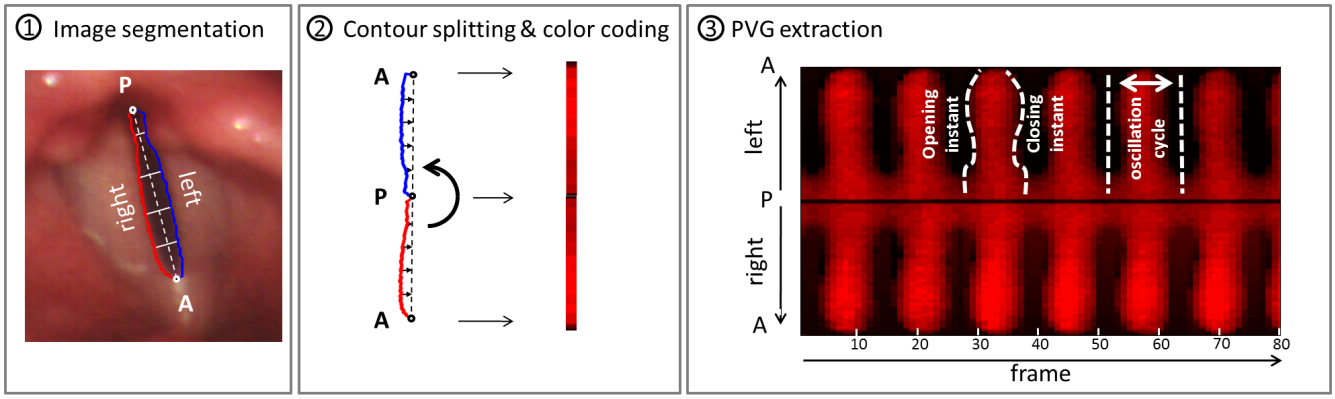

Fig. 2. PVG construction process. (1) Image segmentation and vocal fold identification. (2) Contour splitting and color coding. (3) Concatenation of the color-coded strips forming the phonovibrogram (PVG)

sampling rate of 4000 frames per second. All patients were requested to phonate the vowel /ae/ at habitual pitch and loudness for at least one second during rigid endoscopy.

B. Phonovibrogram extraction

Phonovibrography is a method for mapping high speed videos into two dimensional color graphs. The construction process is summarized in Figure 2. First, the glottal area is segmented in each video frame using a modified region growing algorithm [11] (Figure 2, step 1). The left contour is rotated by 180° around the P commissure. Along the glottal axis, from most anterior (A) to most posterior (P) ending, the distances to the left and right VF contour are color-coded (Figure 2, step 2) and finally, the color-coded strips are concatenated forming a two-dimensional image called PVG (Figure 2, step 3). The PVG color intensities reflect the deflection as a time-dependent function. While bright colors correspond to large distances from the glottal axis, dark colors indicate closed states of the glottis. A detailed description of the assembling process can be found in [8]. The PVG-graph provides the basis for further feature extraction routines that will be described in the following section.

C. Feature extraction

Characteristic patterns of regular vocal vibration are embodied in geometrical structures within the PVG image determined by the opening- and closing instant (Fig. 2, step 3). According to the European Laryngological Society (ELS) [12], these can be classified as triangle, V-shape, rectangle, convex and concave glottal closure type. As this classification is performed in a subjective manner, it is not suitable for an adequate description of VF dynamics.

The continuous wavelet transform (CWT) is a convolution of an arbitrary signal with a set of functions $\Psi_{a,b}$ generated by the mother wavelet Ψ and is given by the inner L^2-product

$$\mathcal{W}_\Psi\{g\}(a,b) = \langle g, \Psi_{a,b}\rangle_{L^2}. \tag{1}$$

Two wavelet bases are employed for the treatment of PVG-data: Ψ_1 denotes a complex Gaussian wavelet of the 8-th order and Ψ_2 the real valued Mexican hat. For each position k along the glottal axis the wavelet transform is given by

$$W_i(k,a,b) = \mathcal{W}_{\Psi_i}\{PVG_k(t)\}(a,b), \qquad i = 1,2. \tag{2}$$

Accordingly, the corresponding wavelet-phase reads as

$$P(k,b) = \arg \frac{\Re(W_1(k,a_0,b))}{\Im(W_1(k,a_0,b))} \tag{3}$$

where a_0 denotes the scale with maximum entropy [14]. In this context, entropy is seen as a measure of disorder of a system or process [13], [14]. Hence, the scale a_0 relates to the fundamental frequency, if the signal is periodic.

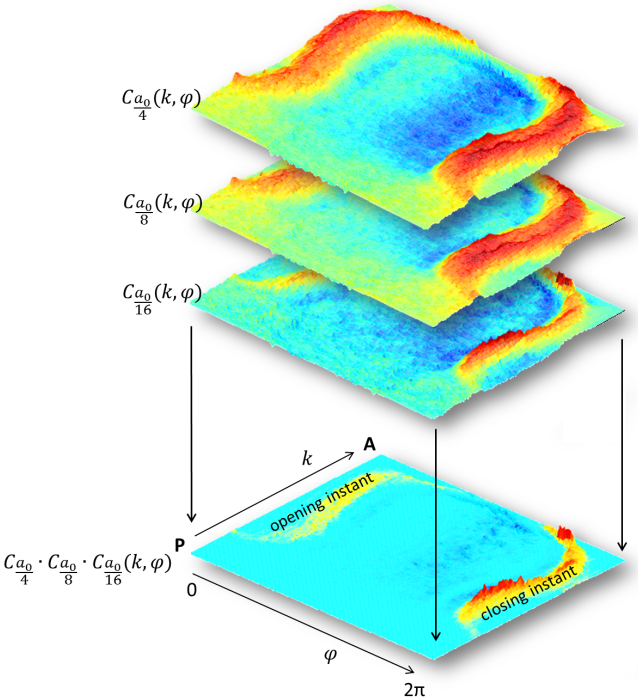

Fig. 3. Representative cycles at different frequency bands. The multiscale product provides a robust and precise localization of opening- and closing instants.

The glottal closure type essentially depends on glottal

opening- and closing instants. These instants are indicated by a rapid behavior change of the glottis [15], resulting in sharp transitions within W_2. In this context, singularities can be localized adequately by multiplying the wavelet coefficients of different scales [16] of W_2. The combination of phase and amplitude information yields a representative cycle

$$C_a(k,\varphi) = \sum_{M=\{b|P(k,b)=\varphi\}} \frac{1}{|M|} W_2(k,a,b) \qquad (4)$$

shown in Figure 3. Opening and closing instants are extracted from the multiscale product $C_{\frac{a_0}{4}} \cdot C_{\frac{a_0}{8}} \cdot C_{\frac{a_0}{16}}$ for left and right VF separately.

To enable a compact representation of the PVG-geometry, a principal component analysis (PCA) was performed projecting the feature space to three dimensions. The first three eigenvectors of the PCA represent different closure types determined by the European Laryngological Society [12] which confirms the reasonableness of this classification. The distance between the projection of left and right VF vibration in PCA space then provides a measure of vibration-symmetry and the energy-distribution of the first eigenvectors directly reflects the vibration periodicity.

Besides geometrical patterns, symmetry and periodicity, an averaged phase shift between left and right vibration was considered. In this context, g_l denotes the area enclosed by the left vocal fold and the glottal axis and g_r the remaining area for the right side. The wavelet-phases of left and right VF vibration read

$$\phi_l(b) = \arg\left(\mathcal{W}_{\Psi_1}\{g_l(t)\}(a_0,b)\right) \qquad (5)$$

$$\phi_r(b) = \arg\left(\mathcal{W}_{\Psi_1}\{g_r(t)\}(a_0,b)\right). \qquad (6)$$

Hence, an averaged phase delay ϕ of n frames is obtained by

$$\phi = \frac{1}{n}\sum_{b=1}^{n} |\arg\left(e^{i(\phi_l(b)-\phi_r(b))}\right)| \qquad (7)$$

providing a measure of synchronicity.

Taken together, the feature vector comprises three parameters for quantifying the geometrical structure of the vibration pattern as well as the sum of the energy of the first three eigenvectors for each VF, respectively. Furthermore, the distance between left and right VF projection in PCA space and a mean phase shift are employed for classification defining a 10-dimensional feature vector.

D. Classification

For the evaluation of the features' discriminative power, a Support Vector Machine (SVM) was employed to build a predictive model in order to decide whether a subject belongs to the healthy or paralytic class. Due to a restricted amount of data, cross-validation with the leave-one-out method was used for evaluation and reproducibility validity. The SVM was further configured with different kernel functions: linear, polynomial and radial basis function. Kernel parameters were found by using a combination of coupled simulated annealing and a standard simplex method.

III. Results

The classification accuracies are shown in Figure 4. The best performance of 93.2% classification accuracy was archived by SVM with RBF kernel functions. The corresponding standard deviation of 0.63% clearly emphasizes the consistency and stability of the classifier results.

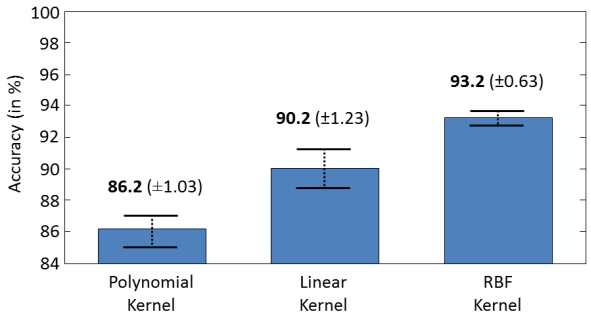

Fig. 4. Classification accuracy of SVM classification employing different kernel functions.

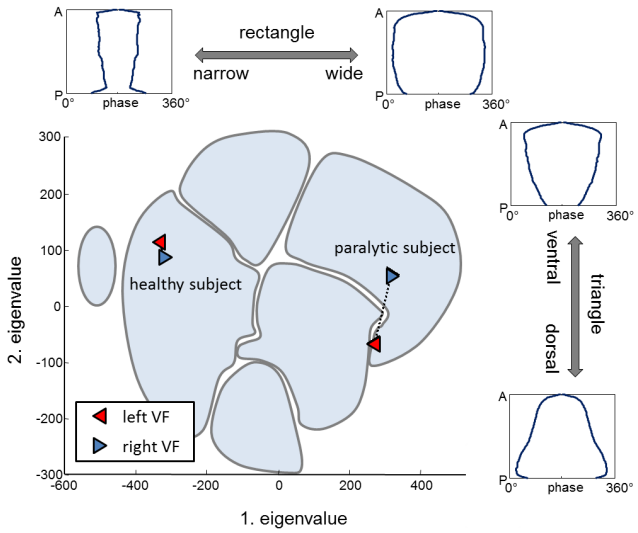

Fig. 5. PCA-subspace of vibrations patterns classified in six clusters. The first eigenvectors correspond to the ELS classification [12]: rectangle narrow-wide and triangle ventral-dorsal.

The PCA subspace spanned by first two eigenvectors is depicted in Figure 5. The first eigenvector corresponds to a rectangle shaped vibration where high eigenvalues correlate with a broad contour and vice versa, low values come along with a narrow geometry. Accordingly, the second eigenvalue determines triangle and V-shape contours. The six clusters are resulting from an agglomerative cluster analysis of 100 healthy subjects. Exemplarily, the projections of left and right VF vibration patterns of a healthy (male, 60yrs) and a paralytic subject (female, 54yrs) are shown in Figure 6. As already mentioned, the distance between left and right

vibration pattern provides a measure of vibration symmetry, which is 25.65 for the healthy and 129.38 for the paralytic subject, indicating a strong decrease of symmetry of the paralytic subject.

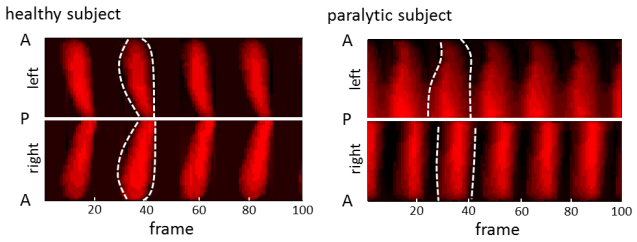

Fig. 6. PVGs computed from a healthy subject (left side, male, 60yrs) and a paralytic subject (right side, female, 54yrs).

IV. Discussion

Classification results demonstrate the suitability of the proposed feature set for the classification of voice disorders. The compact feature space is spanned merely by 10 quantitative features providing stable classification results with low variance of accuracy. Our studies did not show any improvent of the accuracy when taking more than three eigenvectors into consideration.

The reason for the misclassification-rate of 6.8% may be seen in compensatory vocal behaviors at patients' presentation [17] implicating a highly regular and symmetric vibratory behavior. However, extension of the feature set as well as optimization of the classifier will most likely enhance the performance in the near future.

Currently, the procedure is validated within a comprehensive study, where normative data is acquired for the classification of vibratory patterns. Further research will extend the classification task to multiple diagnostic findings including functional voice disorders. Therefore, additional features concerning phase correlations are assessed in terms of their discriminant power. Also, the synchronously recorded acoustic waveform will be incorporated into the analysis.

References

[1] K. Verdolini, L. O. Ramig, Review: Occupational risks for voice problems. Logopedics Phoniatrics Vocology, vol. 26, no. 1, pp. 37-46, 2001.

[2] National Center for Voice and Speech. Occupation and voice data. National Center (Iowa City, IA). 1993.

[3] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, B. Martin-Harris, R. E. Hillmann, Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatr Logop, vol. 60, no. 1, pp. 33-44, 2007.

[4] Y. Yan, K. Ahmad, M. Kunduk, and D. Bless, Analysis of vocal-fold vibrations from high-speed laryngeal images using a hilbert transform-based methodology, J Voice, vol. 19, no. 2, pp. 161-175, 2005.

[5] K. Ishizaka, J. L. Flanagan, Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Syst Tech, vol. 51, pp. 12331268, 1972.

[6] M. Doellinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, U. Eysholdt, Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE Transactions on Biomedical Engineering, vol. 49, no. 8, pp. 773-781, 2002.

[7] R. Schwarz, M. Doellinger, T. Wurzbacher, U. Eysholdt, J. Lohscheller, Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model. J Acoust Soc Am, vol. 123, no. 5, pp. 2717-2732, 2008.

[8] J. Lohscheller, U. Eysholdt, H. Toy, M. Doellinger, Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibrations Into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics. IEEE Transactions on Medical Imaging, vol. 27, no. 3, pp. 300-309, 2008.

[9] D. Voigt, M. Doellinger, A. Yang, U. Eysholdt, J. Lohscheller, Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. Computer Methods and Programs in Biomedicine, vol. 99, no. 3, pp. 275-288, 2010.

[10] D. Voigt, M. Doellinger, T. Braunschweig, A. Yang, U. Eysholdt, J. Lohscheller, Classification of functional voice disorders based on phonovibrograms. Artificial Intelligence in Medicine, vol. 49, no. 1, pp. 51-59, 2010.

[11] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Doellinger, Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Medical Image Analysis, vol. 11, no. 4, pp. 400-413, 2007.

[12] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. V. D. Heyning, M. Remacle, V. Woisard, 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, vol. 258, no. 2, pp. 77-82, 2001.

[13] D. J. Hecker, W. Delb, F. I. Corona, D. J. Strauss, Possible Macroscopic Indicators of Neural Maturation in Subcortical Auditory Pathways in School-Age Children. Engineering in Medicine and Biology Society, 2006. EMBS '06, pp. 1173-1176, 2006.

[14] R. Quian Quiroga, 0. Rosso, and E. Basar, Wavelet-entropy in event related potentials: a new method shows ordering of EEG oscillations, Biol Cybern, vol. 84, no. 4, pp. 291-299, 2001.

[15] D. G. Childers, A. M. Smith, G. P. Moore, Relationships Between Electroglottograph, Speech, and Vocal Cord Contact. Folia Phoniatr, vol. 36, no. 3, pp. 105-118, 1984.

[16] L. Zhang, P. Bao, Edge Detection by Scale Multiplication in Wavelet Domain. Pattern Recognition Letters, vol. 23, no. 14, pp. 1771-1784, 2002.

[17] J. A. Koufman, G. N. Postma, M. M. Cummins, P. D. Blalock, Vocal fold paresis. Otolaryngol Head Neck Surg, vol. 122, no. 4, pp. 537-541, 2000.