35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Multiscale Product Approach for an Automatic Classification of Voice Disorders from Endoscopic High-Speed Videos

Jakob Unger[1], Maria Schuster[2], Dietmar J. Hecker[3], Bernhard Schick[3], Joerg Lohscheller[1]

Abstract— **Direct observation of vocal fold vibration is indispensable for a clinical diagnosis of voice disorders. Among current imaging techniques, high-speed videoendoscopy constitutes a state-of-the-art method capturing several thousand frames per second of the vocal folds during phonation. Recently, a method for extracting descriptive features from phonovibrograms, a two-dimensional image containing the spatio-temporal pattern of vocal fold dynamics, was presented. The derived features are closely related to a clinically established protocol for functional assessment of pathologic voices. The discriminative power of these features for different pathologic findings and configurations has not been assessed yet. In the current study, a collective of 220 subjects is considered for two- and multi-class problems of healthy and pathologic findings. The performance of the proposed feature set is compared to conventional feature reduction routines and was found to clearly outperform these. As such, the proposed procedure shows great potential for diagnostical issues of vocal fold disorders.**

I. INTRODUCTION

The lateral vibration of the vocal folds (VF) modulate the air from the lungs generating the carrier signal of speech. Disordered voice production is commonly due to irregularities or asymmetries of the VF movement leading to disturbances of the voice signal [1]. In this regard, direct observation of the VFs is indispensable to make a reliable diagnosis. As fundamental frequency of VF oscillation is usually within the range of 100 to 400 Hz, modern high-speed cameras provide sampling rates of 2,000 to 10,000 frames per second in order to capture the intra-cycle vibratory characteristics of the VFs [2].

Subjective analysis of endoscopic high-speed videos, however, is time-consuming and it relies on the clinician's knowledge and experience to make judgments about the vibratory behavior and is furthermore restricted by insufficient reproducibility. Hence, objective analysis of endoscopic videos gained increasing attention over the last decades. Physical abnormalities of laryngeal tissues are usually identified by texture analysis [3] of single color images. Current approaches targeting objective analysis of endoscopic high-speed videos quantify glottal perturbation [4], asymmetries [5] or correlation along the anterior-posterior dimension [6]. However, to distinguish between different pathologic findings the complete vibration pattern has to be taken into consideration. In model-based approaches, the VF movement can be adapted to optimally fit the given model by performing mathematical optimization procedures. The most common approaches are lumped element [7] and finite element models [8]. A combination of model parameter estimation and electroglottographic signal analysis is employed by Qin et al. [9]. Biomedical models found a broad acceptance but interpretation and optimization are still extremely challenging tasks.

A comprehensive documentation of laryngeal dynamics was achieved by introducing phonovibrograms (PVG) [10]. The PVG contains the full spatio-temporal pattern of vocal fold dynamics. It allows visualization for clinical interpretation and furthermore, provides the basis for image processing routines to extract valuable features for an objective analysis of VF dynamics. PVG-based classification of voice disorders was made by Voigt et al. by discriminating healthy and paralytic findings [11] as well as healthy and functional voice disorders [12]. The limitations of the approach are twofold: On the one hand, the separation of individual oscillation cycles may be difficult due to aperiodicities often occurring with strong pathologies and on the other hand, the large number of correlated features may reduce the predictive power of the underlying support vector machine (SVM) classifier [13].

Recently, we introduced a wavelet-based analysis of VF vibration and showed that the proposed feature set is closely related to the subjective ELS guideline for functional assessment of pathologic voices. In the current study, we are going one step further by considering healthy subjects and three groups of pathologies: VF paresis, muscle tension dysphonia (MTD) and polyps. Therefore, the corresponding feature vector was extended and optimized and is compared to feature vectors obtained from diverse dimensionality reduction procedures.

II. MATERIAL AND METHOD

A. Phonovibrography

In order to extract VF dynamics from high-speed video sequences, the glottal area is segmented in each video frame by employing a modified region growing algorithm that involves just minimal user intervention [16]. The glottal area is bounded by left and right VF representing the periodical movement of the VFs during phonation. A compact visualization of the glottal area segmentation can be achieved by constructing the PVG. The PVG encodes the spatio-temporal deflection of both VFs by calculating the distances from the

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

[1]Department of Computer Science, University of Applied Science Trier, Trier, Germany

[2]Department of Audiology and Phoniatrics, Campus Grosshadern, University of Munich, Germany

[3]Department of Otorhinolaryngology, Saarland University Hospital, Homburg/Saar, Germany

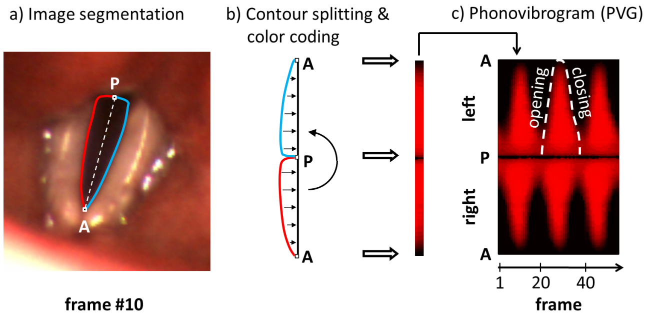

Fig. 1. Construction process of the PVG representation. a) The glottal area is segmented in each video frame, b) the left contour is rotated by 180° and the distances from the VFs to the glottal axis are color coded, c) the color-coded strips are finally concatenated and form the PVG representation.

glottal axis defined by the most anterior (A) and posterior (P) ending of the glottis to the segmented VF contours (Fig. 1). A detailed description of the PVG assembling process can be found in Lohscheller et al. [10].

B. *Wavelet-based analysis*

The PVG exhibits recurring vibration patterns that fundamentally characterize VF dynamics. Recently, we developed a wavelet based approach to quantify the geometrical structure of the vibration pattern involving a minimum number of parameters [14]. Therefore, opening and closing instants are identified forming a characteristic contour within the PVG representation. In Fig. 1 c), these instants (white dashed lines) form a triangular shape that represents a ”zipper-like” opening and closing of the VFs. This characteristic shape is quantified by evaluating the distance between estimated phase values of opening and closing instants along the glottal axis.

C. *Feature vector*

Features of three categories (A,B and C) are used for the assessment of the different classification tasks and are presented in the following.

1) Glottal closure type (category A): The distance from opening to closing instants along the glottal axis quantifies the glottal closure type of VF vibration. However, the distance vector comprises 256 highly correlated entries for each VF. To provide a compact representation of the vibration pattern a ”norm-map” is computed from left and right VF vibration of healthy subjects (reference group). This is achieved by using dimensionality reduction procedures that provide so called out-of-sample extensions allowing to apply a trained model to out-of-sample points. In the current study, several linear and non-linear dimensionality reduction procedures were evaluated: principal component analysis (PCA), kernel PCA (KPCA), Isomaps and locally linear embedding (LLE). For PCA and kernel PCA the out-of-sample extension is actually quite straightforward. For the latter two techniques, out-of-sample extensions have been proposed by Bengio et al. [15]. The raw contour information of left and right VF vibration are projected into the reference space spanned by a group of healthy subjects and the corresponding components constitute the feature set compactly representing glottal closure characteristics.

2) Phase information (category B): The PVG vibratory pattern is also characterized by anterior-posterior and left-right phase relations. A linear regression of the phase displacement in the anterior-posterior dimension is performed for left and right VF individually using the wavelet phase signal defined in [14]. Additionally, the mean phase displacement between left and right VF is estimated and provides a measure of asynchronism. The standard deviation of the phase displacement values over time specifies the coupling of the left and right phase signal.

3) Asymmetry and irregularity (category C): The best classification accuracy was achieved by employing 5 dominant PCA components for left and right VF. It was shown [14] that the first eigenvectors constitute an objective pendant to the current subjective European Laryngological Society (ELS) classification guideline [17]. According to the ELS, the glottal closure can be classified as longitudinal, dorsal, ventral, irregular, oval, and hour-glass shaped. The longitudinal type is characterized by the first eigenvector, dorsal and ventral by the second and oval and hour-glass by the third one (Fig 2). The contribution of higher order eigenvalues characterizes deviations from the reference group. Irregularity of left and right side is therefore quantified as the absolute sum of all components of order 6 and higher. Vibration symmetry is specified by evaluating the Euclidean distance between left and right projection within the PCA space spanned by the first three ELS related components.

D. *Subjects and equipment*

The VF vibrations of 220 subjects were captured with the HS Endocam 5562 high-speed camera (Richard Wolf GmbH, Knittlingen, Germany). The camera provides a sampling rate of $4,000$ frames per second with a spatial resolution of 256×256 pixels and is equipped with a supplemental cold light

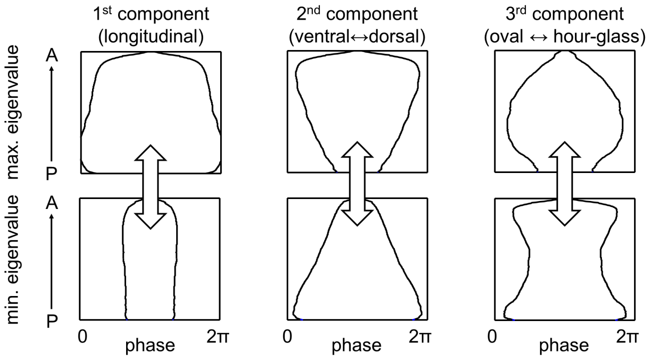

Fig. 2. First, second and third eigenvector component obtained from a PCA that was performed for 100 healthy subjects. The corresponding contours are closely related to the ELS guideline for subjective video assessment of VF vibration.

TABLE I

Classification accuracy (in percent), standard deviations and dimensionality of the feature space for the 2-class, 3-class and 4-class problems and feature categories: A) glottal closure characteristics, B) phase information and C) irregularity and asymmetry.

	A								**A,B**								**A, B, C**	
	PCA		**KPCA**		**Isomaps**		**LLE**		**PCA**		**KPCA**		**Isomaps**		**LLE**		**proposed feature set**	
	acc	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**	**acc**	**dim**
healthy vs. MTD	**77.4** ±4.2	20	71.6 ±5.5	18	71.9 ±4.6	16	70.1 ±4.8	14	**78.3** ±4.7	20+4	73.0 ±4.6	20+4	72.1 ±5.2	40+4	71.5 ±5.2	40+4	**79.5** ±3.3	10+4+3
healthy vs. paresis	81.2 ±2.7	4	**81.3** ±3.4	22	80.4 ±3.7	6	81.4 ±4.5	24	88.7 ±2.5	4+4	**89.3** ±3.7	18+4	88.5 ±2.8	4+4	88.9 ±3.0	16+4	**92.4** ±1.8	10+4+3
healthy vs. polyp	**79.5** ±3.8	14	76.1 ±4.9	36	73.2 ±5.6	40	68.3 ±6.9	40	**81.7** ±3.2	10+4	77.0 ±5.4	20+4	78.0 ±4.4	40+4	74.0 ±4.9	34+4	**89.5** ±2.1	10+4+3
healthy vs. MTD vs. paresis	**53.7** ±2.7	6	52.8 ±3.9	14	52.3 ±2.9	4	52.0 ±4.6	30	**63.6** ±2.9	20+4	61.1 ±2.6	6+4	62.4 ±2.8	4+4	60.6 ±3.3	8+4	**67.5** ±2.6	10+4+3
healthy vs. MTD vs. polyp	**69.9** ±2.7	26	64.7 ±3.9	40	61.8 ±5.3	36	59.3 ±5.1	40	**69.4** ±2.6	18+4	64.4 ±3.8	40+4	65.1 ±4.0	22+4	62.0 ±4.8	36+4	**78.4** ±2.4	10+4+3
healthy vs. paresis vs. polyp	**74.1** ±3.1	14	68.8 ±4.0	22	67.6 ±4.4	30	64.6 ±4.2	36	**77.3** ±2.6	14+4	70.9 ±4.1	30+4	73.2 ±4.0	36+4	68.3 ±4.0	36+4	**86.3** ±1.7	10+4+3
healthy vs. MTD vs. paresis vs. polyp	**56.0** ±2.9	14	51.8 ±4.1	36	49.0 ±3.9	30	46.2 ±3.6	40	**59.6** ±2.2	14+4	53.8 ±3.8	38+4	55.7 ±3.3	20+4	52.3 ±4.1	38+4	**69.0** ±2.0	10+4+3

source.

From the 220 subjects, 40 were with a diagnosed MTD (29 f, 44.03 ± 14.48 yr, 11 m, 57.17 ± 13.27 yr), 40 were found to suffer from unilateral VF paresis (19 f, 51.89 ± 20.70 yr, 21 m, 59.78±8.54 yr) and for 40 subjects (19 f, 54.14±12.97 yr, 21 m, 59.40 ± 12.60 yr), a polyp was found on the VFs. For the remaining collective of 100 subjects (63 f, 41.27 ± 15.84 yr, 37 m, 41.27 ± 15.84 yr) no signs of voice disorders were found. All subjects were instructed to phonate the vowel /ae/ at comfortable pitch and loudness during the examination procedure. For each subject, a sequence of 1,000 frames (= 0.25 s) was considered meaning a total number of 220, 000 segmented video frames.

E. Classification

To evaluate the predictive power of the proposed feature vector, a SVM was trained to assign high-speed videos to one of the four classes: healthy, paresis, MTD or polyp. In the current study, the RBF kernel was used that was found to perform best. Due to a restricted number of subjects within the individual classes the leave-one-out cross validation strategy was pursued. Classification was performed using 1. the raw contour information characterized by 512 highly correlated features and 2. diverse dimensionality reduction procedures. The validation was repeated 100 times and for each iteration a group of 60 subjects was randomly selected from all healthy subjects. Therefore, the selected collective of healthy subjects was used to span the reference space where the remaining subjects (40 per class) were projected into. This strategy ensures balanced class distributions. Multiclass classification was realized through the "one-against-one" strategy [18].

III. RESULTS

As the classifier achieved merely 26.1% accuracy when using the high correlated raw contour information for discriminating between all four classes the high dimensional feature vector is not suited for an automated diagnosis. Consequently, procedures for dimensionality reduction (PCA, KPCA, Isomaps and LLE) were employed. The results are presented in Table I. For each procedure the dimensionality of the feature vector (dim) is shown that achieved the highest classification accuracy (acc). Features were taken from different categories: A) glottal closure characteristics, B) phase information and C) irregularity and asymmetry.

In 6 from 7 cases, PCA was found to perform best. Furthermore, PCA achieved the highest accuracy with lower dimensional feature vectors than KPCA, isomaps and LLE. The additional phase information clearly increased the accuracy values, especially when differentiating between healthy and paretic vibration patterns. Again, PCA showed the best performance.

The proposed feature set (A, B, C) comprising 5 dominant

PCA eigenvectors for each side, phase information, irregularity and asymmetry parameters shows substantial improvements for all classification tasks. This is particularly true for identifying VF polyps. Minor improvements, however, were achieved for identifying functional dysphonia.

IV. DISCUSSION

PCA clearly outperformed other dimensionality reduction procedures providing higher accuracy in combination with lower dimensionality of the feature space. The first three PCA eigenvectors were found to correlate with the subjective ELS guideline. Hence, they have a descriptive meaning helping to further optimize and extend the current feature set. It was also shown that including phase information is essential for comprehensively describing and judging VF dynamics. This is particulary true for organic voice disorders (paresis, polyp). Finally, the proposed feature set additionally comprising irregularity and asymmetry measures was found to have the best overall performance. Eigenvectors of higher order characterize high-frequent disturbances along the anterior-posterior dimension and are therefore summarized in a single irregularity parameter helping to reduce the dimensionality of the feature space.

Generally, higher classification accuracy was achieved when differentiating between healthy vibration patterns and organic pathologies. Organic voice disorders are caused by physical abnormalites in structure of the vocal tract or problems in the nervous system disturbing the vibration pattern in terms of periodicity and symmetry. VF polyps appear localized as a swelling or bump on one or rarely on both VFs. As only global features are used in this study, it is remarkable that polyps localized on arbitrary positions along the VFs are identified quite accurately. Therefore it can be inferred that the entire vibration pattern is significantly altered in the case of an existing polyp.

Classifying functional voice disorders is much more complex. Table I clearly shows the difference between the classifier performances that were achieved for identifying functional (79.5%) and organic (92.4%, 89.5%) voice disorders. However, it has to be kept in mind that the clinical picture of functional dysphonia is quite ambiguous and clinical diagnosis can be a complex process involving history and auditory, acoustic and visual examination [19].

The successive incorporation of clinically relevant features has shown good results so far. In the future, we will incorporate audio analysis of the synchronously recorded acoustic waveform. As acoustic examination is one of the steps in diagnosing functional voice disorders we hope to improve the corresponding classification accuracy. Furthermore, non-stationary phonation was found to provide further diagnostic features [20]. Hence, alterations of the vibration patterns during phonation onset will be quantified and it will be assessed wether improvements of classification can be achieved.

REFERENCES

[1] U. Hoppe, Mechanisms of hoarseness - visualization and interpretation by means of nonlinear dynamics, Ph.D. dissertation, Aachen, 2001.

[2] 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.

[3] A. Verikas, A. Gelzinis, D. Valincius, M. Bacauskiene, V. Uloza, Multiple feature sets based categorization of laryngeal images, Comput. Meth. Prog. Bio., vol. 85, no. 3, pp. 257-266, 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] D. D. Mehta, D. D Deliyski, T. F. Quatieri, R. E. Hillman, Automated Measurement of Vocal Fold Vibratory Asymmetry From High-Speed Videoendoscopy Recordings, J Speech Lang Hear Res, vol. 54, no. 1, pp. 47-54, 2011.

[6] C. R. Krausert, Y. Liang, Y. Zhang, A. L. Rieves, K. R. Geurink, J. J. Jiang, Spatiotemporal analysis of normal and pathological human vocal fold vibrations, Am J Otol, vol. 33, no. 6, 2012.

[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] F. Alipour, D. A. Berry, I. R. Titze, A finite-element model of vocal-fold vibration, J Acoust Soc Am, vol. 108, no. 6, pp. 3003-12, 2000.

[9] X. Qin, S. Wang, M. Wan, Improving Reliability and Accuracy of Vibration Parameters of Vocal Folds Based on High-Speed Video and Electroglottography, IEEE Trans Biomed, vol.56, no.6, pp.1744-1754, 2009.

[10] 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.

[11] 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.

[12] 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.

[13] G. F. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Trans Inf Theory, vol. 14, no. 1, pp. 55-63, 1968.

[14] J. Unger, T. Meyer, M. Doellinger, D. J. Hecker, B. Schick, J. Lohscheller, A wavelet-based approach for a continuous analysis of phonovibrograms, Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE , pp.4410-4413, Aug. 28 2012-Sept. 1 2012.

[15] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. LeRoux, M. Ouimet, Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. In Advances in Neural Information Processing Systems, vol. 16, Cambridge, MA, USA, 2004. The MIT Press.

[16] 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.

[17] 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.

[18] C. Hsu, C. Lin; , A comparison of methods for multiclass support vector machines, IEEE Trans Neural Netw, vol.13, no.2, pp.415-425, 2002.

[19] A. Sama, P. N. Carding, S. Price, P. Kelly, J. A. Wilson, The Clinical Features of Functional Dysphonia, Laryngoscope, vol. 111, no. 3, pp. 458-463, 2001.

[20] T. Braunschweig, J. Flaschka, P. Schelhorn-Neise, M. Doellinger, High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias, Med. Eng. Phys., vol. 30, no. 1, pp. 59-66, 2008.