A Wavelet-based Approach for Analyzing Non-Stationary Phonation Sequences from Endoscopic High-Speed Videos

Jakob Unger[1]
unger@hochschule-trier.de
Maria Schuster[2]
Dietmar Hecker[3]
Bernhard Schick[3]
Joerg Lohscheller[1]

[1] Dept. of Computer Science
Trier University of Applied Sciences
Trier, Germany

[2] Dept. of Audiology and Phoniatrics
University Hospital Munich
Munich, Germany

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

Abstract

Direct observation of vocal fold dynamics is an essential part of clinical diagnosis of voice disorders. Presently, high-speed videoendoscopy is a state-of-the-art procedure to capture the vibratory behaviour of the vocal folds in full images. In order to enable an objective assessment of high-speed video sequences, a wavelet-based method for extracting descriptive features from phonovibrograms, a two-dimensional image containing the spatio-temporal pattern of vocal fold dynamics, was presented. The method quantifies basic characteristics of vocal fold dynamics that are closely related to a basic set of video-based measurements categorized by the European Laryngological Society for a subjective assessment of pathologic voices. Here, features are quantified from sustained phonation at habitual pitch and loudness. Stationary voice production mechanisms, however, do not fully reflect the complex mechanisms of voice production as a whole. Therefore, pathologies might not be predicted reliably. This study addresses an extension of the wavelet-based procedure to non-stationary phonation sequences. The enhanced discriminative power is demonstrated by healthy and pathologic subjects that were examined during sustained and non-stationary phonation.

1 Introduction

The voice production mechanism is primarily given by the vocal folds (VF) modulating the passing flow of air. Pathologic voice is commonly caused by irregular or asymmetric vibratory behaviour [2]. Clinical diagnostics therefore demands visual inspection of the VFs during phonation. The rapid movement of the VFs is recorded with endoscopic high-speed video cameras capturing thousands of frames for each second. In order to facilitate a comprehensive documentation of laryngeal dynamics, phonovibrograms (PVG) [5] were introduced

that allow -besides visualizing even long high-speed videos in compact graphs- also a computerized and hence, objective analysis of VF vibration. From the PVG, a limited number of clinical relevant parameters can be derived by employing a wavelet-based analysis [8]. The wavelet-based analysis characterizes vibration patterns examined during sustained phonation with constant pitch and loudness. The feature's discriminative power has been evaluated for different pathologic findings [7] and showed a good performance so far. However, the VFs' vibration patterns strongly depend on pitch and intensity settings [9]. Therefore, the vibration during sustained phonation characterizes merely a part of the complex voice production mechanism. First studies using non-stationary phonation and a model-based approach showed promising results [9]. Defining adequate models, however, remains extremely challenging. In the current study, we extended the wavelet-based approach to non-stationary phonation sequences. It is shown that healthy and pathological vibration patterns can be differentiated more reliably when analysing non-stationary phonation sequences.

2 Materials and Methods

2.1 Subjects and Equipment

Laryngeal high-speed recordings were performed for a total number of 10 female subjects. For five subjects no signs of voice disorders were found whereas the other five subjects were with a diagnosed unilateral VF paresis. Each subject was examined twice, during sustained phonation at habitual pitch and loudness and during a pitch raise from low to high pitch. The recordings were made with the HS Endocam 5562 highspeed camera system (Richard Wolf GmbH, Knittlingen, Germany) providing a sampling rate of $4,000$ frames per second and a spatial resolution of 256×256 pixels.

2.2 Phonovibrography

In order to quantify the VFs' lateral movement along the visible anterior-posterior dimension the glottal area, enclosed by left and right VFs, is segmented in each video frame using a modified region growing algorithm (Fig. 1 (a)) that proved to provide reliable segmentation results even for low image quality [4]. The segmentation is then transformed to the phonovibrogram (PVG) visualization [5] that encodes the time varying deflection of both VFs from the glottal midline in a single colour image (Fig. 1 (b) and (c)). Besides being a valuable visualization for a subjective analysis of laryngeal function the PVG provides the basis for image processing routines to classify different types of pathologic vibration patterns.

2.3 Wavelet-based Analysis of Phonovibrograms

The PVG exhibits periodically occurring geometric structures for each VF, respectively. An example is illustrated in Figure 1 (c). Here, the triangular shape (white dashed line) represents a "zipper-like" opening and closing. In order to quantify the periodical structures, the wavelet phase is estimated using a complex Morlet wavelet ψ_1. Let $g_{l,r}$ denote the area enclosed by the left or right VF and the glottal axis (Fig. 1 a). The phase of left and right VF vibration reads as

$$\phi_{l,r}(b) = \arctan \frac{\Im\{\mathscr{W}_{\psi_1}\{g_{l,r}\}(a^{l,r}(b),b)\}}{\Re\{\mathscr{W}_{\psi_1}\{g_{l,r}\}(a^{l,r}(b),b)\}} \tag{1}$$

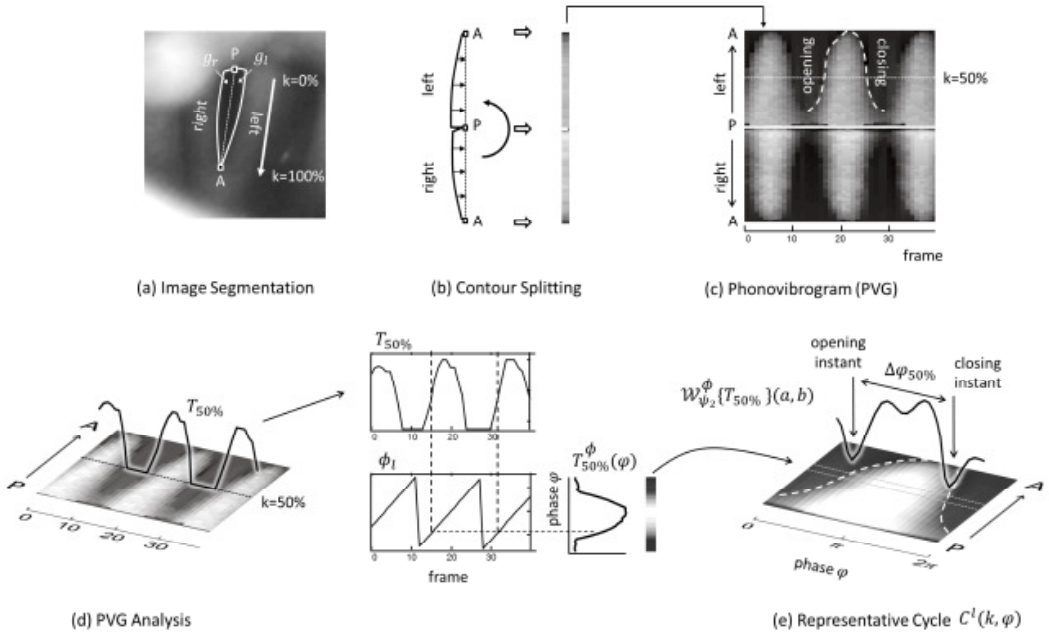

Figure 1: (a), (b), (c) Phonovibrogram construction process, (d), (e) Wavelet-based PVG analysis

where $\mathscr{W}$ denotes the wavelet transform and $a^{l,r}(b)$ denotes the wavelet scale chosen for a specific frame b. In [8], a constant scale $a(b) = a_0$ was taken assuming a constant pitch during laryngoscopy. In order to evaluate non-stationary phonation, in this study, the scale $a(b)$ is chosen from the wavelet ridges as proposed by Carmona et al. [1]. A mean PVG cycle can then be obtained by assigning the PVG values $T_k^{l,r}$ at position k along the anterior-posterior axis to the corresponding phase (Fig. 1 (e)).

$$C^{l,r}(k,\varphi) = \sum_{M=\{b|\phi_k^{l,r}(b)=\varphi\}} \frac{1}{|M|} T_k^{l,r}(b) \tag{2}$$

The graph $C^{l,r}$ shows the dominant geometric structure periodically appearing within the PVG representation. This pattern is essentially determined by the opening and closing instants forming the triangular shape. To localize these instants a second order Gaussian wavelet ψ_2 is employed. The wavelet has two vanishing moments and hence, acts as differential operator of the smoothed source signal [6] providing minima at opening and closing instants (Fig. 1 f). A more precise localization of these instants can be achieved by employing a multiscale product (MSP) as done by Unger et al. [8] involving the product of several successive wavelet scales. Thereby, opening and closing instants are indicated by sharp peaks whereas noise components are suppressed.
Finally, in order to separate opening and closing instants from each other a first order Gaussian wavelet ψ_3 is employed. Consequently, the sign of the wavelet transform

$$\operatorname{sgn}\left(\mathscr{W}_{\psi_3}\{C^{l,r}\}(a,b)\right) \tag{3}$$

indicates whether the MSP minimum refers to glottal opening or closing.

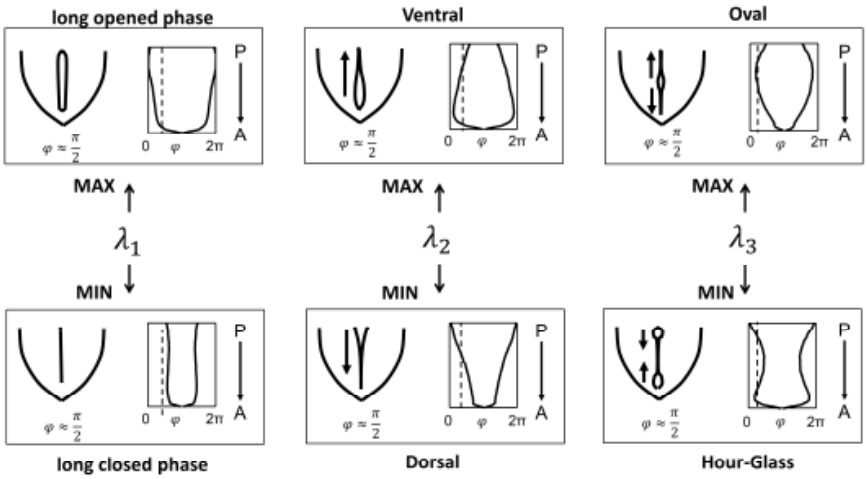

Figure 2: Correlation of the eigenvalue decomposition and VF vibration patterns.

2.4 Principal Component Analysis

A compact description of the contours is achieved by computing the distance from opening to closing instants $\Delta\varphi_k$ along the entire length of the glottal axis (Fig. 1 (f)). From the distance vector $\Delta\varphi_k$, a PCA is performed for 100 healthy subjects during sustained phonation. The PCA space spanned by these subjects serves as norm-map where the remaining subjects to be analysed are projected into. The eigenvectors spanning the PCA space and examples of corresponding glottal configurations are depicted in Fig. 2. The first eigenvector exhibits a rectangular shape affecting the contour's width in phase direction. A long and complete closure is encoded by low λ_1 values whereas short or incomplete closure corresponds to high λ_1 values. Triangular geometries comply with a zipper-like opening and closing. Opening from anterior to posterior is given by high λ_2 (ventral) and vice versa, opening from posterior to anterior corresponds to low values of λ_2 (dorsal). Values of λ_2 around zero can be interpreted as superposition of both types meaning that the VFs open and close simultaneously alongside the glottal axis. Finally, the third eigenvector encodes oval and hour-glass contour shapes meaning opening from medial to terminal and vice versa, from terminal to medial. The first three eigenvectors make up more than 90% of the overall variance and hence, comprehensively describe vibration patterns for each VF individually. Furthermore, the first three eigenvalues are closely related to a categorical scheme for a subjective assessment of pathologic voices defined by the European Laryngological Society [3] demonstrating the clinical interpretability of the eigenvalue decomposition.

2.5 Assessment of non-stationary Phonation Sequences

By projecting a PVG into the given PCA space the vibration characteristics can be assessed with merely three parameters for each VF. Furthermore, when taking windowed PVG segments, non-stationary phonation sequences will form a trajectory through the PCA space comprehensively describing alterations of vibration characteristics over time. In the current study, windowed segments with a length of 400 frames ($= 0.1sec$) were shifted by 200 frames in each time step for the non-stationary recordings whereas the entire length of the signal was taken for the stationary sequences.

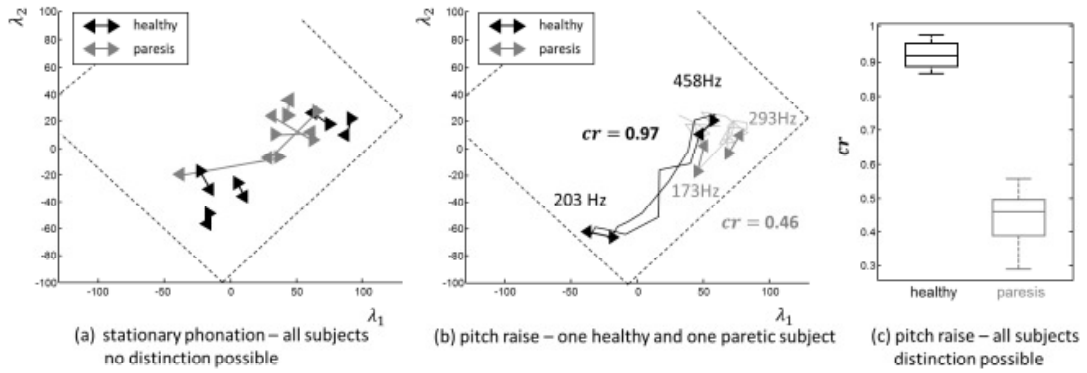

Figure 3: (a), (b) PCA space characterizing VF vibration patterns during pitch raise from low to high frequency, (c) Linear correlation of left and right vibration patterns during pitch raise.

In order to evaluate whether the vibration patterns of left and right VF are changing synchronously, the correlation coefficients c_{λ_i} between the eigenvalues λ_i of left and right VF vibration patterns are computed. Let s_{λ_i} denote the range of the i^{th} eigenvalue paramter (during pitch-raise) and $s_0 = \sum_{i=1}^{3} s_{\lambda_i}$. In this respect, the measure

$$cr = \frac{s_{\lambda_1}}{s_0}\left|c_{\lambda_1}\right| + \frac{s_{\lambda_2}}{s_0}\left|c_{\lambda_2}\right| + \frac{s_{\lambda_3}}{s_0}\left|c_{\lambda_3}\right| \tag{4}$$

quantifies the weighted linear correlation between left and right vibration patterns during pitch raise.

3 Results and Discussion

Figure 3 (a) shows the projections of left and right VF vibration patterns of all subjects examined during sustained phonation. A higher asymmetry can be seen for the paretic subject illustrated by the higher distance between the projections of left and right VF within the PCA space. Although asymmetry is obviously higher for the paretic group, no clear distinction between healthy and paretic subjects can be made from the eigenvalue measures alone.
Figure 3 (b) exemplarily demonstrates the change of the vibration pattern during a pitch raise for a healthy and a paretic subject. For the healthy subject, a systematic change from dorsal zipper-like vibration to a simultaneous opening and closing (longitudinal closure) accompanied by a gradually increasing opened-closed ratio can be seen. Contrarily, no systematic change can be observed for the paretic subject. For the entire sequence, high λ_1 values are observed for the paretic subject corroborating the fact that paresis is often accompanied by a complete absence of VF contact. For the healthy subject high λ_1 values are merely seen for higher pitches. Furthermore, the correlation measure $cr = 0.46$ indicates a reduced coupling between both VFs compared to the healthy subject ($cr = 0.97$). A boxplot of the correlation measure cr evaluated for all subjects is given in Figure 3 (c). From this measure, healthy and paretic subjects can be separated adequately.

4 Conclusion

Analysis of non-stationary phonation provides additional information of the underlying laryngeal dynamics that cannot be derived from stationary processes. It has been shown that additional clinical relevant measures can be obtained providing valuable information for a diagnosis of voice disorders. In order to verify these advantages more quantitative measures have to be defined and their relevance for different pathologic findings have to be clarified in more extensive studies.

5 Acknowledgement

This work is supported by the German Research Foundation (DFG), Grant No. Lo-1413/2-2.

References

[1] R.A. Carmona, W.L. Hwang, and B. Torresani. Multiridge detection and time-frequency reconstruction. *IEEE Trans Signal Processing*, 47(2):480–492, 1999.

[2] U. Eyshold, F. Rosanowski, and U. Hoppe. Irregular vocal fold vibrations caused by different types of laryngeal symmetry. *Eur Arch Otorhinolaryngol*, 260(8):412–417, 2003.

[3] G. Friedrich and P. H. DeJonckere. The voice evaluation protocol of the european laryngological society (els) – first results of a multicenter study. *Laryngo Rhino Otol*, 84(10): 744–752, 2005.

[4] J. Lohscheller, H. Toy, U. Eysholdt, and M. Doellinger. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Med Image Anal*, 11(4):400–413, 2007.

[5] J. Lohscheller, U. Eysholdt, H. Toy, and M. Doellinger. Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics. *IEEE Trans Med Imag*, 27(3):300–309, 2008.

[6] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, ISBN: 0123743702, 2009.

[7] J. Unger, M. Schuster, D.J. Hecker, B. Schick, and J. Lohscheller. A multiscale product approach for an automatic classification of voice disorders from endoscopic high-speed videos. In *Engineering in Medicine and Biology Society (EMBC)*, pages 7360–7363, 2013.

[8] J. Unger, D.J. Hecker, M. Kunduk, M. Schuster, B. Schick, and J. Lohscheller. Quantifying spatiotemporal properties of vocal fold dynamics based on a multiscale analysis of phonovibrograms. *IEEE Trans Biomed Eng, in press*, 2014.

[9] T. Wurzbacher, R. Schwarz, M. Doellinger, U. Hoppe, U. Eysholdt, and J. Lohscheller. Model-based classification of nonstationary vocal fold vibrations. *J Acoust Soc Am*, 120 (2):1012–1027, 2006.