Artificial Intelligence in Medicine 66 (2016) 15–28

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A generalized procedure for analyzing sustained and dynamic vocal fold vibrations from laryngeal high-speed videos using phonovibrograms

Jakob Unger[a,*], Maria Schuster[b], Dietmar J. Hecker[c], Bernhard Schick[c], Jörg Lohscheller[a]

[a] Department of Computer Science, Trier University of Applied Sciences, Schneidershof, 54293 Trier, Germany
[b] Department of Otorhinolaryngology and Head and Neck Surgery, University of Munich, Campus Grosshadern, Marchioninistr. 13, 81366 München, Germany
[c] Department of Otorhinolaryngology, Saarland University Hospital, Kirrbergerstr., 66424 Homburg/Saar, Germany

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form
28 September 2015
Accepted 20 October 2015

Keywords:
Wavelet-based analysis
Multiscale product
Wavelet ridge
Voice disorder
High-speed laryngoscopy
Dynamic phonation

ABSTRACT

Objective: This work presents a computer-based approach to analyze the two-dimensional vocal fold dynamics of endoscopic high-speed videos, and constitutes an extension and generalization of a previously proposed wavelet-based procedure. While most approaches aim for analyzing sustained phonation conditions, the proposed method allows for a clinically adequate analysis of both dynamic as well as sustained phonation paradigms.
Materials and methods: The analysis procedure is based on a spatio-temporal visualization technique, the phonovibrogram, that facilitates the documentation of the visible laryngeal dynamics. From the phonovibrogram, a low-dimensional set of features is computed using a principle component analysis strategy that quantifies the type of vibration patterns, irregularity, lateral symmetry and synchronicity, as a function of time. Two different test bench data sets are used to validate the approach: (*I*) 150 healthy and pathologic subjects examined during sustained phonation. (*II*) 20 healthy and pathologic subjects that were examined twice: during sustained phonation and a glissando from a low to a higher fundamental frequency. In order to assess the discriminative power of the extracted features, a Support Vector Machine is trained to distinguish between physiologic and pathologic vibrations. The results for sustained phonation sequences are compared to the previous approach. Finally, the classification performance of the stationary analyzing procedure is compared to the transient analysis of the glissando maneuver.
Results: For the first test bench the proposed procedure outperformed the previous approach (proposed feature set: accuracy: 91.3%, sensitivity: 80%, specificity: 97%, previous approach: accuracy: 89.3%, sensitivity: 76%, specificity: 96%). Comparing the classification performance of the second test bench further corroborates that analyzing transient paradigms provides clear additional diagnostic value (glissando maneuver: accuracy: 90%, sensitivity: 100%, specificity: 80%, sustained phonation: accuracy: 75%, sensitivity: 80%, specificity: 70%).
Conclusions: The incorporation of parameters describing the temporal evolvement of vocal fold vibration clearly improves the automatic identification of pathologic vibration patterns. Furthermore, incorporating a dynamic phonation paradigm provides additional valuable information about the underlying laryngeal dynamics that cannot be derived from sustained conditions. The proposed generalized approach provides a better overall classification performance than the previous approach, and hence constitutes a new advantageous tool for an improved clinical diagnosis of voice disorders.

1. Introduction

Voice quality is known to be closely interconnected with vocal fold (VF) vibration [1]. A rough and hoarse voice is often caused by asymmetric and irregular vibration essentially affecting the modulation of the air-stream coming from the lungs [2]. In this regard, endoscopy is an inherent part of clinical examination routines. Currently, videostroboscopy is still the most prevalent imaging technique in clinical practice [3], capturing VF dynamics during sustained phonation at constant pitch and loudness. During stroboscopy, light flashes are synchronized to the fundamental frequency of VF vibration with a slightly increasing phase

* Corresponding author. Tel.: +49 2418027866.
E-mail address: jakob.unger@lfb.rwth-aachen.de (J. Unger).

http://dx.doi.org/10.1016/j.artmed.2015.10.002
0933-3657/

[4]. Thus, stroboscopy simulates real-time playback by merging frames from different oscillation cycles. Consequently, stroboscopy is restricted to highly periodic VF vibration, and thus clinically to a sustained phonation examination paradigm. Rapidly changing and non-periodic signals cannot be clearly resolved by stroboscopy.

Contrarily, high-speed videoendoscopy (HSV) captures full laryngeal images in real-time with sampling rates of 2000–20,000 fps, allowing new insights into physiologic and pathologic mechanisms of the VFs [5,6]. Despite allowing for analyzing transient and thus dynamic phonation paradigms, HSV examination is still mostly performed during sustained phonation at constant pitch and loudness. However, the vibration observed during sustained phonation represents only a very limited subset of the entire complex voice production mechanisms. It is known, that the degree of symmetry and regularity of VF vibrations highly depends on the examined frequency interval [7]. Thus, pathologic indications may not be observable in a sustained phonation condition at only one specific pitch or intensity level.

From a clinical point of view, there is a particular demand for extracting valuable and interpretable parameters from both: (A) sustained phonation sequences focusing on the type and stability of VF vibrations at specific pitch and intensity, and (B) dynamic phonation sequences focusing on the systematic change of VF vibration types during pitch or intensity change. While most approaches focus on analysis of sustained conditions [1,8–10], just a limited number of studies deal with the analysis of transient phonation paradigms. Braunschweig et al. reported promising results for diagnosing functional voice disorders from phonation onset analysis [11,12]. They demonstrated their method in a total of 71 subjects, and found statistically significant differences between both groups. However, the approach involves eight to ten examinations at different sound pressure levels for each subject, which limits the clinical applicability. Transient laryngeal dynamics were examined by Rasp et al. [13] during a monotonous increase of the fundamental frequency. Generally, pitch is a result of changing the VFs' eigenfrequency controlled by length, tension and mass of the vibrating VFs [14]. While the authors found decreasing VF amplitudes during the glissando, the change of the glottal closure characteristics could only be classified into six rough categories. A continuous characterization was not achieved in this context.

Quantitative analysis of dynamic phonation was performed by Wurzbacher et al. [7,15]. They extended a stationary biomechanical two-mass model to dynamic phonation paradigms and examined its benefit for identifying pathologies from a glissando using a parameter optimization strategy [16,17]. However, since the lumped element models imply a very strong discretization of the vibrational structure along the anterior-posterior dimension, the analysis of the longitudinal propagation of VF vibrations is quite limited. Furthermore, it had not been proven yet that the computed model parameters, which are obtained using an optimization procedure, are unique. A proper clinical interpretation of the parameters remains challenging.

A continuous analysis of VF dynamics along the anterior-posterior (AP) dimension was achieved by introducing phonovibrograms (PVG) [18]. The principle of PVG construction is demonstrated in Fig. 1(1–3). The PVG encodes the lateral displacement of VF vibrations in a single image by color-coding the distance between the glottal axis and the VF contours for left and right side, respectively. Concatenating the color-coded stripes of the subsequent images of video results in a two-dimensional image, the PVG. Depending on the specific type of VF vibrations, geometric patterns occur within the PVG, which represent the glottal opening and closing process. In Fig. 1(3), the white dotted contour line exemplarily represents the vibration pattern of the left VF for the second oscillation cycle. The PVG contains the visible spatio-temporal pattern of VF vibration along the entire glottal axis and allows for a visual assessment and further clinical interpretation without inspecting the entire high-speed video sequence in slow motion.

The underlying data within a PVG provide further information for a quantification of the VF vibration characteristics. Recent works showed that quantitative features can be derived from PVGs that characterize VF vibrations very precisely [20–24]. However, there are two main limitations with these initial approaches: Firstly, the number of extracted features is extremely high. Voigt et al. employ about 200 numerical features for classification. High-dimensional feature vectors, however, may reduce the predictive power of the classifier – a phenomenon known as the "curse of dimensionality" [25]. Secondly, the features are not understandable within a clinical context causing problems with the clinical interpretability and acceptance.

To solve both problems, we recently proposed a wavelet-based procedure in combination with a principal component analysis (PCA) to analyze the geometric structure of a PVG. It extracts the most significant properties of VF vibration patterns by quantifying the main characteristics of the PVG geometry (dotted contour line Fig. 1(3)) on the basis of three dominant PCA eigenvalues [19]. The three eigenvalues $\lambda_1 - \lambda_3$ provide a compact representation of the complex VF vibration patterns along the entire glottal axis. Furthermore, the eigenvalues are clinically interpretable as they correlate with a rating scheme elaborated by the European Laryngological Society for a perceptive assessment of glottal closure characteristics [26]. As shown in Fig. 1(4), the first eigenvalue λ_1 encodes the open-closed ratio, whereas λ_2 distinguishes between a zipper-like opening from posterior to anterior (dorsal) or anterior to posterior (ventral). The third eigenvalue λ_3 (not shown here) distinguishes between oval (opening from medial to terminal) and hour-glass (opening from terminal to medial) vibration types. From the eigenvalue distribution, clinically relevant measures can be derived that allow distinguishing between physiologic and pathologic vibration patterns [27]. Moreover, it could even be shown that the procedure differentiates between malignant and precancerous lesions solely on the basis of laryngeal dynamics analysis [28].

The PVG eigenvalues allow one to uncover precisely even the intra-individual vibratory changes induced by different phonation conditions in a quantitative manner. Fig. 1(4) exemplarily illustrates the two-dimensional PCA subspace for λ_1 and λ_2 comprising the results of five distinct examinations of a single healthy male subject obtained at different fundamental frequencies (F_0: 143 Hz – 333 Hz). The different fundamental frequencies provoke a systematic change of the VF vibration patterns, which is clearly reflected by the different magnitudes of both eigenvalues.

The hitherto existing PVG wavelet analysis approach is limited to the characterization of VF vibrations during sustained phonation (stationary analysis), however. It thus does not take advantage of a dynamic examination paradigm that incorporates additional information about VF dynamics compared to the sustained one. The goal of the work presented here is to propose a generalized procedure allowing a clinically adequate analysis of both dynamic as well as sustained phonation sequences in order to enhance the computer assisted diagnosis of voice disorders (non-stationary analysis). The discriminatory power of the new features introduced here is validated using a supervised machine learning approach.

2. Methods

2.1. Clinical data

The proposed method was developed and verified on the basis of clinical HSV recordings, including two different clinical test benches and phonatory tasks. Endoscopic imaging was performed with the HS Endocam 5562 high-speed camera system (Richard