Biomedical Signal Processing and Control xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Technical note

Comparison of an audio-based and a video-based approach for detecting diplophonia

Philipp Aichinger[a,b,*], Imme Roesner[a], Matthias Leonhard[a], Berit Schneider-Stickler[a], Doris-Maria Denk-Linnert[a], Wolfgang Bigenzahn[a], Anna Katharina Fuchs[b], Martin Hagmüller[b], Gernot Kubin[b]

[a] Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Waehringer Guertel 18-20, 1090 Vienna, Austria
[b] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c/EG, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 22 March 2014
Received in revised form 30 August 2014
Accepted 1 October 2014
Available online xxx

Keywords:
Laryngeal high-speed videos
Diplophonia
Audio signal processing
Video signal processing
Degree of subharmonics
Pathologic voice

ABSTRACT

Background and objectives: Diplophonia is a common symptom in voice disorders. Depending on the underlying aetiology, diplophonic patients typically need treatment such as phonosurgery or speech therapy. In current clinical practice, the presence of diplophonia is assessed by auditive rating. To avoid subjectivity in voice assessment and to follow principles of evidence based medicine, objective instrumental assessment methods are needed. In order to gain insight into instrumental assessment of diplophonic voice, comparisons between different assessment approaches are necessary. The aim of the study is to compare the performance of two independent objective approaches on their ability to detect diplophonia. The compared approaches are the formerly published degree of subharmonics (DSH), and a newly proposed measure for spatial bimodality of the vocal fold vibration.
Material and methods: From a clinical database of 352 laryngeal high-speed videos with synchronous audio recordings, 60 phonation segments (20 euphonic, twenty diplophonic and twenty non-diplophonic dysphonic) were auditively selected. For all phonation segments, the DSH and the newly proposed measure for spatial bimodality were determined. The DSH is the occurrence rate of audio analysis blocks with ambiguous fundamental frequency in percent. The bimodality measure quantifies the spatial occurrence of secondary oscillation frequencies along the vocal folds' edges. Both the DSH and the bimodality measure are evaluated on their ability to detect diplophonia by means of cut off threshold classification.
Results and conclusions: The DSH showed excellent classification rates for separating diplophonic from euphonic phonation (sensitivity: 98.4%, specificity: 100%). In separating diplophonic from non-diplophonic dysphonic phonation, the bimodality measure slightly outperforms the DSH approach (sensitivity: 54.6%, specificity: 92.7%). The separation of diplophonia from other kinds of dysphonia is challenging, and more sophisticated methods are needed. It is concluded that auditive and glottal diplophonia must be distinguished. As the clinical assessment of diplophonia primarily aims at determining glottal conditions, the video-based approach might deliver clinically more relevant data than the auditive approach.

Abbreviations: DSH, degree of subharmonics; LHSV, laryngeal high-speed video; STP, spatio-temporal plot; ROC, receiver operating characteristic; AUC, area under the curve; NUM, non-unimodality measure.

* Corresponding author at: Corresponding author. Tel.: +43 1 40400 11670; fax: +43 1 40400 42840.
E-mail addresses: philipp.aichinger@meduniwien.ac.at (P. Aichinger), imme.roesner@meduniwien.ac.at (I. Roesner), matthias.leonhard@meduniwien.ac.at (M. Leonhard), berit.schneider-stickler@meduniwien.ac.at (B. Schneider-Stickler), doris-maria.denk-linnert@meduniwien.ac.at (D.-M. Denk-Linnert), wolfgang.bigenzahn@meduniwien.ac.at (W. Bigenzahn), anna.fuchs@tugraz.at (A.K. Fuchs), hagmueller@tugraz.at (M. Hagmüller), gernot.kubin@tugraz.at (G. Kubin).

http://dx.doi.org/10.1016/j.bspc.2014.10.001
1746-8094/

1. Introduction

Diplophonia is a common symptom in disordered voice, and characterized by the simultaneous presence of two separate pitches in the voice [1]. Depending on the underlying aetiology, diplophonic patients typically need treatment such as phonosurgery or speech therapy. In order to provide solid indications for treatment as well as accurate treatment outcome measures, the presence of diplophonia needs to be detected accurately. In current clinical practice it is assessed by auditive rating, which mostly suffers from its intrinsic subjectivity. Moreover, the clinical assessment is subject-global and neglects the time-variant nature of diplophonia.

In voice assessment, several kinds of physical phenomena are often summarized into the terms "irregularity" [2] or "roughness" [1]. In fact, there are several subcategories of irregular phonation (i.e., diplophonia, creak, fry, double pulsing, glottalization, laryngealization, pulse register phonation, squeak), which intersect each other [3]. By generalizing these phenomena, the accuracy of voice assessment is reduced. The presented research is based on the fact that the subcategories of irregular phonation are not solidly defined. In [4], the authors stated that "variations in scope, purpose, terminology, measurement technique, and level of description make it difficult to compare vocal phenomena across disciplines, or even across studies within a single discipline", and describe the need for an update in terminology. Beside the perceptive definition of diplophonia there is also a definition based on the wave shape: Diplophonia can be characterized by a period-two up-down pattern of an arbitrary cyclic parameter (i.e., double pulsing) [5]. As the perceptive definition and the wave shape definition are not equivalent, the current terminology is questionable.

To overcome these restrictions and to meet principles of evidence based medicine, objective instrumental assessment methods are needed. Despite great efforts in the development of such methods [6], their performance is not yet satisfying. In order to get an insight into instrumental assessment of diplophonic voice, the aim of the study is to compare the performance of two independent objective approaches on their ability to detect diplophonia. The compared measures are the formerly published degree of subharmonics (DSH) [7], and a newly proposed analysis method for measuring oscillation frequencies from laryngeal high-speed videos (LHSVs). In the absence of a unified definition and a solid ground truth of diplophonia, auditive annotations are used as a baseline categorization in this study, and the instrumental methods will be analysed on their ability to detect auditive diplophonia.

A useful concept with respect to the definition of diplophonia is the concept of metacycles. Metacycles arise when two oscillators are combined, either additively, producing a beat frequency phenomenon, or multiplicatively, where both oscillators modulate the amplitude and frequencies of each other. Two frequencies in a rational ratio are commensurable and produce a metacycle. Its length is equal to the least common multiple of the oscillators' individual period lengths.

However, in addition to this very comprehensive concept of metacycles, diplophonia must also be analysed from the perspective of laryngeal dynamics. With respect to asymmetric vocal fold vibration, several types of diplophonia have been described [5,8–14], while the vocal folds create two distinct oscillators at different frequencies. Diplophonia can either arise from left-right asymmetry [8,9] (i.e., the left and the right vocal fold are vibrating at different frequencies), anterior–posterior asymmetry [9,10] (i.e., the anterior and the posterior part of the vocal folds are vibrating at different frequencies), or combinations of both [8]. Double pulsing (i.e., alternating amplitudes or period lengths) [7,9–12] can arise from slight anterior–posterior asymmetry [10], or from inferior–superior asymmetry (i.e., vertical modes [15]), which is not well understood.

The existence of various types of asymmetry in diplophonic voice motivates the use of LHSVs. Unfortunately the vast amount of data makes the video analysis complex, time consuming and expensive. Although the clinical application of LHSVs has been investigated intensively, it is still limited [16–18]. There is a lack of clinically relevant interpretation guidelines for LHSVs, thus this technique needs more investigation, which is addressed in this study.

There is very little knowledge about the relation of audio recordings and LHSVs. From literature, it must have been expected that diplophonia determined from the audio signal is similar to diplophonia determined from the video. In common sense, one could expect that the relation between oscillation phenomena and perceptual cues is not straight forward, as psychoacoustic phenomena are manifold [19,20]. Nevertheless, this distinction has not yet been described systematically and incorporated in the terminology of voice research. It will be shown, by comparing the audio-based and the video-based detection method, that the audio domain and the video domain must be treated separately. Hence, the term "glottal diplophonia" is introduced. In contrast to auditive diplophonia, glottal diplophonia is characterized by two distinct oscillation frequencies of the vocal fold edges, as determined from LHSVs. One possible reason for the divergence between the audio domain and the video domain, is the possibility of extraglottal sound sources that provoke the sensation of a secondary pitch.

The idea of quantitatively analysing spatio-temporal plots of LHSVs (STPs, i.e., phonovibrograms) is not new. For example, in empirical eigenfunction analysis [9], the plots are decomposed into spatially orthogonal oscillation modes (i.e., eigenmodes). The spatial irregularity is measured by summarizing the eigenmodes' weights in an entropy measure. Although this decomposition is related to our work, it does not primarily aim at decomposing separate fundamentals, which limits the applicability for detecting double pitch phenomena.

Mechanical modelling of vocal fold vibration is accomplished by optimizing the parameters of a mechanical model with a predefined structure onto the empirical STPs [21]. This approach shows the potential to extract the mechanical properties of the vocal fold tissue, which would be a highly desirable goal in voice research. However, due to the limited possibilities to observe the complex vocal fold movement (i.e., LHSVs are a two-dimensional projection of a three-dimensional movement) we accept the fact that the extraction of all mechanical parameters of the vocal folds is not possible in the near future. As a realistic research assignment, we stick to the goal of detecting the presence of diplophonia, without estimating its underlying tissue dynamics.

Yet another approach for analysing STPs is artificial intelligence [22]. Although such data driven approaches can achieve good classification rates, its acceptance in clinical practice is generally low, mainly because of the black box nature of the prediction rules. It is important that voice analyses for clinical use are intuitively understood by clinicians. Hence, artificial intelligence is beyond the scope of this paper.

An approach of decomposing the glottal area waveform or the audio signal into distinct oscillators, is empirical mode decomposition [23]. In empirical mode decomposition, time series are iteratively analysed in order to extract amplitude/frequency modulated signals, which is related to the analysis of coupled oscillators. Diplophonia is produced by glottal oscillators that are coupled, either anatomically (i.e., the oscillators modulate each other via connecting tissue), or aerodynamically (i.e., the oscillators modulate each other via a common airflow). Although empirical mode decomposition seems well suited for detecting coupled oscillators, the extracted modes are difficult to interpret. In addition, similarly to empirical eigenfunction analysis, the empirical mode decomposition does not aim at extracting distinct fundamental frequencies.

Please cite this article in press as: P. Aichinger, et al., Comparison of an audio-based and a video-based approach for detecting diplophonia, Biomed. Signal Process. Control (2014), http://dx.doi.org/10.1016/j.bspc.2014.10.001