Biomedical Signal Processing and Control 13 (2014) 71–78

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Testing software tools with synthesized deviant voices for medicolegal assessment of occupational dysphonia

N. Ruqja[a], P.H. Dejonckere[b,c], G. Cantarella[d], J. Schoentgen[e], S. Orlandi[a,f], S.D. Barbagallo[a], C. Manfredi[a,*]

[a] *Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy*
[b] *Catholic University of Leuven, Neurosciences, Exp. ORL, Belgium*
[c] *Federal Institute of Occupational Diseases, Brussels, Belgium*
[d] *IRCCS Ca' Granda Foundation, Ospedale Maggiore General Hospital, Milano, Italy*
[e] *Laboratories of Images, Signals and Telecommunication Devices, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium*
[f] *Department of Electrical Energy and Information Engineering, Università degli Studi di Bologna, Bologna, Italy*

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form 25 March 2014
Accepted 27 March 2014
Available online 25 April 2014

Keywords:
Occupational dysphonia
Synthetic signals
Jitter
Noise
Software tools

ABSTRACT

The measure of acoustic parameters related to dysphonia has long been one of the main research challenges in voice analysis. When the origin of the voice disorder is occupational, the medico-legal and insurance aspects must be considered, involving, when relevant, the definition of a percentage of physical impairment (usually 1–10% according to the guidelines). An objective aid capable of quantifying these levels is desirable for assisting the medico-legal expert. Vocal fold nodules in voice professionals are mentioned in the European List of Occupational Diseases. Generally, such voices are slightly to moderately deviant, and mainly characterized by audible air escape (breathiness), due to insufficient vocal fold closure. Mild amounts of jitter may also be present.

No studies have been published so far testing the ability of software tools to discriminate levels of added noise in signals when the amount of noise is exactly known. In the present study, four program tools (BioVoice, PRAAT, MDVP and AMPEX) are tested on realistic synthesized voice signals corrupted by 4 slightly increasing levels of jitter and 10 levels of additive noise close to the degree of dysphonia occurring in real patients with vocal fold nodules. The results show that the four tools are able to correctly estimate both jitter and noise levels. Specifically for noise, in some cases the agreement is close to 100%. Thus they could be of help to clinicians in determining the level of impairment to quantify compensation in patients affected by an occupational voice disorder.

1. Introduction

Speech is the most important means of communication in social interactions between humans. Moreover, in modern society the number of people using their voice professionally is increasing. Teachers, salespeople, singers, gym-instructors, counselors, clergymen, drill sergeants, politicians, etc. depend on their voice quality as a primary tool in their professional activity, a voice disorder thus severely interferes with the activity and may cause considerable economical damage both individual and social due to absenteeism.

Quite often, occupational voice disorders, or occupational dysphonias, are due to vocal overuse, dryness of air and dust, psychological factors, especially stress, etc. [1]. The diagnosis of occupational voice disorders also requires a good clinical experience in distinguishing between voice diseases occurring in professional voice users, but not due to the occupation, and those related to the high load of voice and other factors due to occupation [2]. Research shows that women are more affected by voice disorders, and especially occupational voice disorders, due to anatomical and physiological factors [3] as well as to the type of profession. In fact, female teachers are the most affected by occupational voice disorders. They also represent a relevant part of the workforce in comparison to other voice professionals. Vocal fold nodules are the most common laryngeal pathology in teachers [3,4].

When dealing with occupational voice, the medico-legal and insurance aspects must be considered. Voice professionals may attribute their voice disorder to profession and accordingly claim

* Corresponding author at: Department of Information Engineering, Università degli Studi di Firenze, 50139 Firenze, Italy. Tel.: +39 55 479 6410; fax: +39 55 4796497.
E-mail addresses: claudia.manfredi@unifi.it, manfredi@det.unifi.it (C. Manfredi).

http://dx.doi.org/10.1016/j.bspc.2014.03.011
1746-8094/

compensation from the national insurance system for occupational diseases.

Most European countries have set up Institutes of Occupational Diseases in charge of prevention and compensation of working impairments. Unfortunately, although vocal fold nodules are mentioned in the European List of Occupational Diseases, voice disorders are not officially recognized as an occupational disease in all countries [2]. Moreover, there is no well-defined international agreement about how to evaluate as objectively as possible the severity of a permanent occupational dysphonia and the related working impairment in these patients. Today, referring to existing medico-legal guidelines, clinicians and insurance experts usually quantify the level of impairment due to dysphonia with a scale from 1% to 10% (in extreme cases up to 30%) [5], based on perceptual rating. Thus, it would be desirable to have an objective aid for quantifying the level of impairment related to the level of dysphonia in these patients with vocal fold nodules, whose voices are in general slightly to moderately deviant, and mainly characterized by audible air escape due to insufficient vocal fold closure [6].

The use of software tools in clinical practice requires, as with any device, the verification of their reliability. This study tests some commonly used software tools for voice analysis (MDVP [7], PRAAT [8], Ampex [9] and BioVoice [10]). We compared their performance based on well-known benchmarks obtained from synthetic data. In doing so it was possible, for the first time, to exactly define the limits of the tools as regards the estimation of dysphonia, measured as the signal to noise ratio. This has considerable importance, especially in applications related to the occupational field, where an accurate assessment of the level of impairment is required that may help medico-legal experts with the task of quantifying the severity of the occupational dysphonia, and thus the level of compensation.

In general, a hoarse voice is acoustically characterized by an excess of noise in the frequency range corresponding to the vocal tract resonances, but also in the high frequency range with loss of high frequency harmonic components. Moreover, some degree of irregularity of vocal folds vibration, named jitter, frequently contributes to perceived dysphonia.

One of the innovative aspects of this work concerns the data set used to test the tools, that is made up by synthesized voice signals that resemble real signals obtained with a recently developed formant-based synthesizer [11] allowing controlling exactly as well the level of perturbation as the level of additive noise. Synthetic signals consist here of a sequence of five repetitions of the fragment/ aiu/ that simulate an ideal short sentence, voiced throughout, with intonation and formants changes. Different levels of f0-perturbation (jitter) and additive noise, both causing voice to be perceived as hoarse, are added to the signal to simulate different levels of perceived pathology. In previous work (dealing with sustained /a/'s), it was shown that perceptual raters could not differentiate such synthesized voices from true pathological voice emissions [11]. In the current study, the levels of jitter and noise are defined on the basis of what is commonly observed in patients with vocal fold nodules: in such patients, noise due to turbulent transglottal air escape prevails over irregular vibrations, because nodules are usually small and symmetrical mass lesions. BioVoice, PRAAT, MDVP are tested with regard to both jitter and noise, while Ampex is applied to jitter estimation only because it does not report additive noise.

A large number of studies aim at quantifying perceived hoarseness and relating perceptual scales to quantitative indexes. A complete review of all studies in this field is out of the scope of the present paper, thus only few are quoted here, in particular with regard to methodology.

The relationship between levels of noise measured in the signal and perceived hoarseness (e.g. GIRBAS scale or visual inspection of spectrograms) has been studied for decades. It started with the pioneering work reported in Ref. [12] and [13] and later in Ref. [14]. Imaizumi et al. [15] estimated the ratio of the energies in the non-harmonic and the harmonic parts of the signal and used these ratios to evaluate the level of hoarseness. Kojima et al. [16] used the Fourier expansion to separate the noise from the periodic components and estimated the signal to noise ratio (SNR). Later, Yumoto [17,18] proposed the Harmonics to Noise Ratio, which was improved further by Awan [19]. Hiraoka [20] proposed the relative harmonic intensity, de Krom a cepstrum-based approach [21], the relationship of which with perceptual scale was exploited in Ref. [22]. In 1986, Kasuya [23] introduced the normalized noise energy (NNE) based on adaptive comb filtering in the frequency domain. Methods based on nonlinear analyses in pseudo-state spaces were also proposed [24,25]. Research continued in more recent times and is still active. An exhaustive list of papers on this topic is outside the scope of this work, so we just quoted only those on which the voice analysis tools (commercial or free) currently in use in clinical and in research centers are based.

In this study, as we focus on noise, we compare some of the latter, i.e. MDVP [7], PRAAT [8] and BioVoice [10]. In particular, MDVP computes the NHR (noise to harmonic ratio) in pre-fixed frequency ranges, PRAAT makes use of the autocorrelation of the signal in the time domain as an estimate of the harmonics-to-noise ratio (HNR) on short signal frames and BioVoice uses an adaptive version of the NNE for which the width of the window is inversely related to fundamental frequency f0. To take into account differences between methods with regard to the frequency range over which noise is estimated, the comparison is performed applying statistical indexes.

As mentioned earlier, another feature related to the level of hoarseness – but of less importance in cases of vocal fold nodules – is jitter, that reports the irregularity of vocal fold vibration: the higher the irregularity the higher the jitter (in %). For jitter estimation, the most widely accepted approach is based on quantifying cycle-to-cycle variations of the fundamental period T_0. All the tools considered here, including Ampex, implement the same formula thus allowing for direct performance comparison.

2. Materials and methods

2.1. The synthesizer

The synthesizer developed and used in this work is a formant based synthesizer for disordered voices; it is an upgrade of the previous synthesizer developed by Fraj et al. [11]. The synthesizer uses a model of the glottal area based on a polynomial distortion function [26,27] that transforms two excitatory harmonic functions in the desired waveform. The polynomial coefficients are obtained by constant, linear and invertible transforms of the Fourier series coefficients of Klaat's template cycle that is asymmetric and skewed to the right [27]. This waveform is in fact typical for the glottal area cycle, allowing a maximal glottal area of 0.2 cm^2.

The discrete phase of the harmonic excitation functions changes from one iteration to the other with a step defined by the inverse of the sampling frequency. The sampling frequency is set at 200 kHz to simulate voices the frequency modulation of which is of the order of 1% of the fundamental frequency, thus requiring a high temporal resolution. The harmonic excitation functions are low-pass filtered and down-sampled to 50 kHz before their transformation by the distortion function.

To simulate voice perturbations as jitter (fast modulation) and tremor (slow modulation of the fundamental frequency in the 2–15 Hz range), phase and/or amplitude disturbances of the harmonic excitation functions are introduced. These models are based on a modulation of the instantaneous frequency of the harmonic