Speech Processing in Diagnosis of Vocal Chords Diseases

Milan Jičínský, Jan Mareš
Faculty of Electrical Engineering
University of Pardubice
Pardubice, Czech Republic
milan.jicinsky@student.upce.cz

Ludmila Verešpejová, Martin Chovanec
Department of Otorhinolaryngology
Charles University Prague, 3rd Medical Faculty,
University Hospital Královské Vinohrady
Prague, Czech Republic

***Abstract* – The paper presents new possibilities in the ways of analyzing voice of patients suffering from voice disorders. Specialized software has been developed for this purpose. Application called Voice disorder diagnostician allows creating own patient database, storing patient's data, capturing voice of patients for further analysis and displaying results of feature extraction algorithm. Common diagnostic methods and procedures takes into count only limited number of parameters. The main idea is to combine medical experience and audio processing techniques in order to achieve desired results.**

Keywords – voice disorder; feature extraction; voice processing; diagnostics; vocal cords; voice processing, Voice disorder diagnostician

I. INTRODUCTION

Modern medicine willingly benefits from the development of computer science. It is mainly because these offer all sorts of easy and effective noninvasive diagnostic methods, such as observing and analyzing different patterns.

Computational intelligence techniques are employed for detection of nasopharyngeal cancer for ages [1]. Different studies deal with oral and nasal cancer diagnosis using Raman spectra online analysis [2-5]. Different possible approaches of information engineering in diagnostics employs artificial intelligence tools. Application can be found in diagnosis of neck and head cancer [6]. One of untutored way of research in this field is voice disorder analysis in early diagnosis of othorinolaringeal diseases.

Voice captured researchers' attention because of usefulness in order to assess early vocal pathologies, and neurodegenerative and mental disorders [7]. Many research groups deal with a voice as a source of large amount of information about the speaker as sex, age or regional origin [8]. Mehta et al. introduced usage of a miniature accelerometer on the neck surface to analyze a large set of ambulatory data from patients with hyper-functional voice disorders [9-11].

II. VOICE DISORDERS

Term phonation corresponds to the generation of vocal sounds while speech represents generation of word sounds. The key structures involved in phonation and speech are oral and nasal cavities, pharynx, larynx, trachea, bronchial tree, lungs, thorax, and diaphragm. During the phonation, in response to brain stimuli from the cerebral cortex the respiratory muscles contract and expiratory flow from the lungs is pushed upwards towards the trachea and larynx as a power source, while at the same time both vocal cords are adducted through the laryngeal muscles, closing the glottis. The expiratory flow raises the subglottal pressure, causing the vocal cords to vibrate and generate sound, which passes through the vocal tract. Vocal tract acts as a resonance chamber to produce vocal sound. Consonants and vowels are articulated, becoming speech, and are generated continuously to produce spoken words. Voice disorders are consequences of impairment of either phonation control or motion mechanisms. These can be subclassified as: glottal closure disorder, affected vocal cord stiffness, vocal cord asymmetry, respiration/resonance chamber disorders and neuropsychological dysfunctions.
When examining patients with voice disorders history taking, listening to voice and speech and endoscopic laryngeal examination are crucial tools for proper diagnostic work-up. The main drawback of the mentioned methods is lack of objectivity in voice evaluation.

III. MATHEMATICAL BACKGROUND

In order to learn any information from recording, it has to be processed by voice processing algorithm. Our version consists of pre-processing and feature extraction phase in the form of Matlab function. Pre-processing means conversion of imported audio track to the single channel (mono), pre-emphasis filtering and segmentation of sound into the fixed length frames. This is followed by feature extraction (evaluation of parameter values for each frame). The features are stored in feature vectors which are used for representing patterns found in recording. Classification is made based on feature vector values instead of using original recording. One of fundamental features is log energy defined by the given formula

$$E = \log\left(\sum_{n=0}^{\mathrm{L}-1} x^2(n)\right), \tag{1}$$

where x(n) are the sample values of analyzed signal and L is the number of samples within frame.

ISBN 978-80-261-0812-2,

Zero crossing rate also ranks among basic static features and it is given by

$$ZCR = \frac{1}{2}\left(\sum_{n=1}^{L-1} |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]|\right). \quad (2)$$

Evaluation of spectral energy B_k is very similar to log energy. Samples contained within frames are convoluted with hamming window followed by N-point FFT. The first half $\left(\frac{N}{2}\text{values}\right)$ is divided into k frequency bins. In this case, 8 bins with 500Hz scale were calculated. This corresponds to

$$B_k = \log\left(\sum_{n=0}^{L-1} f^2(n)\right), \quad (3)$$

where k is the current bin number and f is original signal frame edited using window function and FFT. As for the feature selection, cepstral features are crucial for any kind of speech analysis. Since 1990s till nowadays they are considered to be most beneficial for automatic speech recognition, speaker identification and many other speech analysis sectors. Despite many modifications the most common features carrying information about cepstrum are Mel-frequency cepstral coefficients (MFCC). Each frame must pass several steps. A simplified diagram of obtaining MFCC is shown at figure 1.

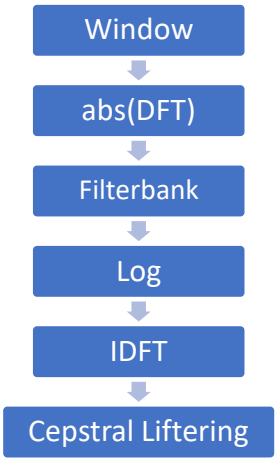

Figure 1 - Obtaining MFCC.

The last part of static feature vector is reserved for fundamental frequency also called pitch. Own pitch detection algorithm is implemented. Algorithm is based on cross-correlation frame analysis. Whenever a maximum value is found, it is compared with threshold. If its greater it is considered to be a rightful lag value. The lags are firstly filtered by 5th order median filter. Filtered lags are used to calculate fundamental frequency according to the formula

$$f_0 = \frac{f_S}{L_{MED}}, \quad (4)$$

where f_S is sampling frequency and L_{MED} is lag filtered by median filter.

In addition to some static features the feature vector contains dynamic features. Dynamic features represent differences between selected feature values. These values are compared for adjacent frames. Dynamic features are also called delta or delta-delta coefficients according to the distance between compared frames. In general delta coefficient calculation is defined by formula

$$\Delta f_p(k) = f_p(k+1) - f_p(k-1), \quad (5)$$

where f_p is p-th feature and k is the index of current frame. In order to obtain second order differences (delta-deltas) the formula (5) would be used again in different form. Features are now replaced with delta features

$$\Delta\Delta f_p(k) = \Delta f_p(k+1) - \Delta f_p(k-1). \quad (6)$$

As for the Voice disorder diagnostician, delta coefficients are computed for log energy, spectral energy and MFCC. Delta-delta coefficients are computed for log energy and MFCC only.

IV. Voice Acquisition and Processing

The recording process takes place in specialized silent chambers located in University Hospital Královské Vinohrady. The procedure consists of these steps: (i) microphone calibration, (ii) recording and evaluation of examined data (patients read standardized text). Usage of a single microphone, having the same acoustic conditions and reading standardized text is beneficial for minimizing negative variability as much as possible. So that only a variability of voice can be analyzed. It means any vocal changes influenced by treatment can be distinguished. As for the necessary facility, omnidirectional DPA microphone with a flat frequency response is a key part. Because of low sound quality of embedded sound cards, external sound card is used for transferring sound to the computer.

A. *Voice Disorder Diagnostician*

Although the voice processing is nowadays very common in various research area, not many software tools exist for the purpose of diagnostics. Our application called Voice disorder diagnostician can record patients and display the results of current data immediate analysis. Graphical user interface was simplified to make its usage efficient and intuitive even for staff. Voice disorder diagnostician was programmed in Matlab 2018a so as to use some of its brand-new audio processing features and functions. Beta version of our software and the voice processing methods were both published in [12]. There are several major changes in comparison with older version.

When launching the application, the main menu is displayed to user. Language options can be change in the lower right corner. Default language is set to Czech. There are four options referring to specific application modes. User can choose from “Enter new patient data”, “Import patient data”, “Recording mode” and “Analysis mode”. Status bar is located below. It displays whether any patient was chosen or not. The recording and analysis modes are unavailable in this phase.

The first option opens a window, where physician can enter necessary information needed for further processing and analysis. With a respect to European union GDPR, the application does not process any sensitive data such as name or surname. Each patient has its own generated unique code. This code has no

relation with name of subject and so the patients cannot be identified this way. After entering patient's unique code, sex, age and additional information including diagnosis specification can be inserted. When confirmed, all data is saved to file and user is navigated back to the main menu. Status bar then shows the code of patient and buttons of recording and analysis modes become available. An alternative first step would be the second option Choose existing patient from database. Application opens an explorer window filtering files by desired extension. It allows user to choose a patient according to the patient's code. After importing a data file, status bar shows detailed information.

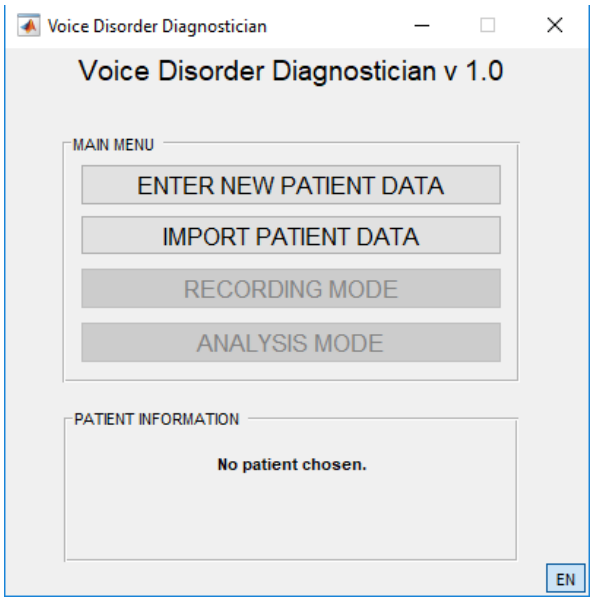

Figure 2 - Main menu.

Since the patient has been chosen recording made is available. New window opens after clicking corresponding button. Graph shows no values because no recording device has been set. After selecting a source, recording device immediately captures sound and sample values are displayed to make sure that device was selected correctly. Red button "not recording" indicates that recording process has not yet started. After clicking the red button, it turns green and informs user that recording of speech just started. Samples shown before are cleared and user can see only sample values corresponding to the sound currently captured. Recording session time is pre-set to five minutes but it can be stopped manually any time simply by clicking the green button again. When button color is red again, it confirms that recording process stopped and the whole captured speech is displayed. Recording mode screenshot is shown at figure 3.

Before saving an audio file, current phase of treatment has to be selected. Three treatment phases are distinguished. Phase zero (before surgery) – patient suffers from voice disorder and no treatment has been applied yet. Phase one (approximately 2 days after surgery) – patient undergone a surgery and he/she starts recovering. Phase two (one or 2 weeks later) – patient comes for usual check-up. Significant voice health improvement should be recognizable. The last step is to export audio recording to a file. This can be done by clicking the "save" button. Recording is automatically saved in predefined format with wav extension. The file name is a combination of patient's unique code and number of treatment phase. In case of incorrectly selected phase, error message warns that file already exists, and user should consider changing the treatment phase. User is notified if file was successfully saved. After recording is saved user is navigated back to the main menu.

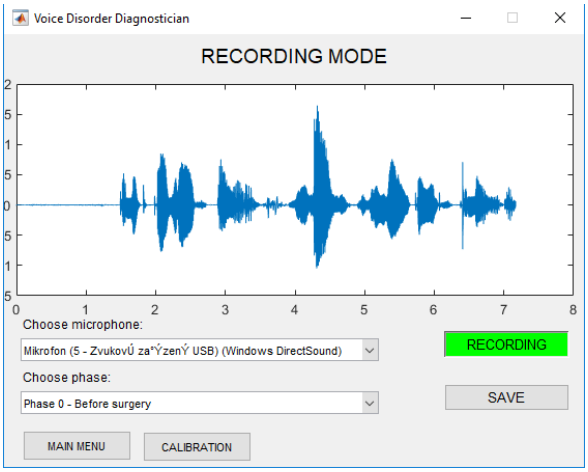

Figure 3 - Recording mode.

The last mode is reserved for detailed analysis of sound files. Analysis mode window is divided into two parts. This makes possible a comparison of two recordings. This should be usually used to compare treatment phases and see the differences. User selects first audio file from menu to import it. Imported file is processed by audio processing and feature extraction algorithm and results are shown in the first column. Then second file can be imported the same way. Graphs of second recording are located in the second column. Analysis mode displays a variety of characteristics divided into pages. Browsing the pages can be made by navigation buttons in the lower right corner. First page shows three graphs – audio file sample values (which is equal to sound pressure level if microphone is calibrated), log energy of signal and zero crossing rate. Second page shows pitch (fundamental frequency) graphs. One graph is the result of pitch detection algorithm (cross correlation method for detecting lag) and the other one shows histogram of detected frequencies. Third page shows spectrogram and displays mean values of so-called cepstral coefficients and spectral energy distribution in given frequency bins.

V. Results

As mentioned before, the whole recording process takes place in University Hospital Královské Vinohrady. Local silent chamber fulfills strict requirements for data acquisition. Experiment started during February 2019. Data collection process lasts about three weeks for each patient. Unfortunately, there are many voice disorder types and some of them are rare. In order to have statistically significant results the dataset should be extended. Despite having a lack of data, some key factors and features can be still discussed. Figures 4 and 5 shows the changes between patient's stage zero and stage one.

Patient spoke more silently and didn't emphasize too much. This corresponds to sample values and low energy. Zero crossing rate is higher, which means that more noise character is present in the speech. On the other hand, figure 5 confirms patient's voice health improvement. Researched subject speaks more precise and louder and so the energy rises. Zero crossing rate decreases because of the fact that the voice is clearer, and patient don't suffer from any other difficulties. Even the tempo also sometimes measured as speech velocity is slightly higher. Although the results seem promising, the amount of collected data is too low for classification and automatic voice disorder diagnostics.

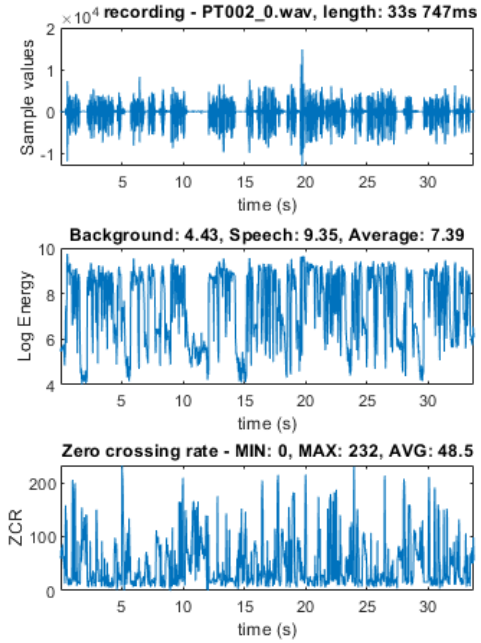

Figure 4 - Patient before surgery - stage zero.

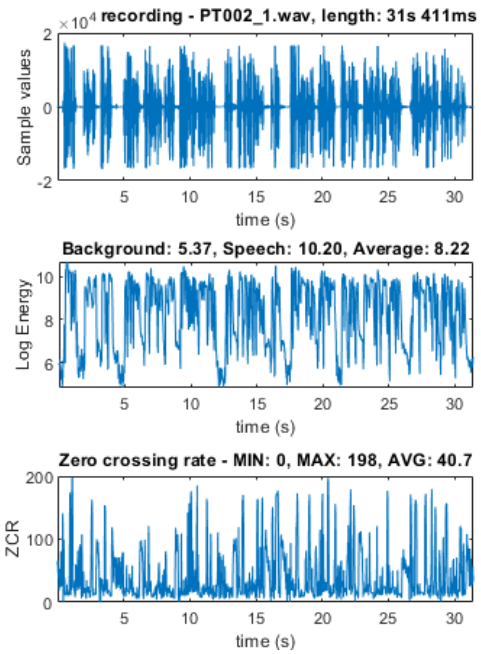

Figure 5 - The same patient after surgery – stage one.

VI. Conclusion

The proposed methodology shows modern sensors biomedical application. The main advantage of Voice disorder diagnostician is in objective description (without doctor's subjective opinion). Results from the testing stage in the hospital are promising and it improves significantly their current approach. Because the work is based on data from the rare disease, we are now in the stage of collecting data from patients to be able to validate the approach on a greater amount of cases. Further work will be aimed on speech velocity measurement and creation of a new application mode capable of automatic voice disorder diagnostics.

Acknowledgment

This research was supported by students SGS grant at Faculty of Electrical Engineering, University of Pardubice. This support is very gratefully acknowledged.

References

[1] LIN, Duo. Blood surface-enhanced Raman spectroscopy based on Ag and Au nanoparticles for nasopharyngeal cancer detection. *Laser physics* [online]. 2016, **26**(5), 055601 [cit. 2017-01-18]. DOI: 10.1088/1054-660X/26/5/055601. ISSN 1054660X.

[2] WANG, Jing. Label-free detection of serum proteins using surface-enhanced Raman spectroscopy for colorectal cancer screening. *Journal of biomedical optics* [online]. 2014, **19**(8), 087003-087003 [cit. 2017-01-18]. DOI: 10.1117/1.JBO.19.8.087003. ISSN 10833668.

[3] WANG, J. SERS spectroscopy and multivariate analysis of globulin in human blood. *Laser physics* [online]. 2014, **24**(6), 065602 [cit. 2017-01-18]. DOI: 10.1088/1054-660X/24/6/065602. ISSN 1054660X.

[4] REKHA, P. Raman Spectroscopic Characterization of Blood Plasma of Oral Cancer. In: *2013 IEEE 4TH INTERNATIONAL CONFERENCE ON PHOTONICS (ICP)* [online]. 2013, s. 135-137 [cit. 2017-01-18]. ISBN 1467360759. ISSN 23305665.

[5] FENG, Shangyuan. Gastric cancer detection based on blood plasma surface-enhanced Raman spectroscopy excited by polarized laser light. *Biosensors & bioelectronics* [online]. 2011, **26**(7), 3167-3174 [cit. 2017-01-18]. DOI: 10.1016/j.bios.2010.12.020. ISSN 09565663.

[6] HARRIS, Andrew. Raman spectroscopy in head and neck cancer. *Head & neck oncology* [online]. 2010, **2**(1), 26-26 [cit. 2017-01-26]. DOI: 10.1186/1758-3284-2-26. ISSN 17583284.

[7] Gómez Vilda, P., Rodellar Biarge, et al.: Characterizing neurological disease from voice quality analysis. Cognit. Comput. 5, pp. 399–425. (2013).

[8] Benzeghiba, M., De Mori, R.: Automatic speech recognition and speech variability: a re-view. In: Speech Commun, vol 49, pp. 763–786. (2007).

[9] Mehta Daryush D., Van Stan Jarrad H.: Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update. Frontiers in Bioengineering and Biotechnol-ogy. Vol. 3, (2015).

[10] Mehta, D. D., Zañartu, M.: Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. IEEE Trans. Biomed. Eng. 59, 3090–3096. (2012)

[11] Mehta, D. D., Zeitels, S. M.: High-speed videoendoscopic analysis of relationships between cepstral-based acoustic measures and voice production mechanisms in patients undergoing phonomicrosurgery. Ann. Otol. Rhinol. Laryngol. 121, 341–347. (2012).

[12] Jičínský M., Mareš J.: "Measurable changes of voice after voice disorder treatment," unpublished.