Discriminating Pathological Voice From Healthy Voice Using Cepstral Peak Prominence Smoothed Distribution in Sustained Vowel

Discriminating pathological voice from healthy voice using Cepstral Peak Prominence Smoothed distribution in sustained vowel

Antonella Castellana[1], Alessio Carullo[1], Simone Corbellini[1], Arianna Astolfi[2]

Abstract

This paper deals with Cepstral Peak Prominence Smoothed (CPPS) distribution and its descriptive statistics as possible indicators of vocal health status. 41 voluntary patients and 35 control subjects participated in the experiment: all of them followed the same protocol, which includes three repetitions of the sustained vowel /a/ simultaneously acquired with a microphone in air and a contact sensor, the perceptual assessment of voice quality and the videolaringoscopy examination. The fifth percentile and the standard deviation of CPPS distribution were the parameters included in the best logistic regression models for the microphone in air and the contact sensor, respectively. The selected CPPS parameters had a strong to good discrimination power: an Area Under Curve of 0.95 and 0.87 has been found for the microphone in air and for the contact sensor, respectively. For each CPPS parameter, the repeatability has been also estimated and the Monte Carlo method has been implemented for the uncertainty evaluation of the discrimination threshold. Furthermore, preliminary recommendations for better accuracy and repeatability of future studies are provided: analyses on the main CPPS influence quantities and on the effect of the frequency content of the signal spectrum on the CPPS parameters have been provided.

Index Terms

Cepstral analyses, human voice, biomedical measurement, acoustic devices, reproducibility of results, Monte Carlo methods, uncertainty

I. Introduction

Traditionally, voice quality has been assesed using subjective tests, in which experts listen to live or recorded vocal signals and perceptually rate them. In order to overcome the subjectivity and the expensiveness of such methods and with the aim to find a less time-consuming tool, researchers started to analyze voice signals and to extract several parameters as indexes of different aspects of voice and voice-related issues.

Authors are with the Politecnico di Torino ([1]Dipartimento di Elettronica e Telecomunicazioni, [2]Dipartimento Energia), corso Duca degli Abruzzi, 24 - 10129 Torino (Italy); phone: +39 011 0904202, fax: +39 011 0904216, e–mail: antonella.castellana@polito.it

A first field of study deals with voice acoustic analysis as objective tool to assess voice disorders thanks to its non-invasiveness, low cost and ease of application [1]. The numerical output provided is relatively easy to communicate to all stakeholders, $e.g.$ voice clinicians, patients, third-party payers and physicians [2], and allows tracking of vocal behavior. Such analysis is thus appealing not only for diagnosis, but also for dysphonia prevention and dysphonia treatment.

Another object of study about the analysis of voice signals is related to the recent spread of innovative digital technologies that has caused the need of evaluating speech quality in telephone systems, $e.g.$ speech quality as one of the parameter for the service quality provided to the users by operators. Therefore, several non-intrusive tools have also been implemented to predict the speech quality in a telephone conversation, such as algorithms that use clipping statistics [3], digital watermarking [4], GSM encoders [5] and optimized multi-sine signals [6], but also In-Service Nonintrusive Measurement Device [7]. Moreover, techniques for the discrimination between speech and voice-band data transmission in telephone systems have been explored [8].

A further voice-related field is based on the investigation of vocal signals with the aim to study illnesses that are not directly linked to the vocal apparatus but for which the voice quality is an effective indicator. For example, monotonous sounding speech indicates depression and suicidal individuals often use toneless sounds while speech [9]. Furthermore, analysis techniques have been developed in order to detect snoring sounds during sleep [10] and to study the obstructive sleep apnea [11].

About the first field of study, that is the vocal signal analysis as a detector of vocal disorders, many algorithms and methods have been implemented (see Buder for an overview [12]), even though most of them suffer from a lack of metrological characterization. In this paper, the authors describe a method to obtain an objective analysis of dysphonia that takes the main uncertainty contributions into account and allows the main influence quantities to be identified.

The first investigated parameters were those in the time domain, $e.g.$ jitter and shimmer, whose main limitations have been highlighted in the existing literature. Since they depend on the accurate identification of cycle boundaries, that is where a cycle of vocal-fold vibration starts and finishes, they become unreliable with highly perturbed signals [13]. Furthermore, the good performance of the speech task, $i.e.$ a vowel produced with steady pitch and loudness, is very important for the computation of such parameters, since any changes in the signal could be read as increases in vocal perturbation [14]. To overcome such limitations, spectral- and cepstral-

based measures are currently considered: they can be applied also to continuous speech that is able to represent everyday speaking patterns [15]. In particular, cepstral parameters have been defined the most promises indexes of dysphonia severity. They are evaluated in the cepstrum domain, that is a log power spectrum of a log power spectrum [16]: while the first power spectrum shows the frequency distribution of the signal energy, the second spectrum indicates how periodic the harmonic components in the spectrum are. Two cepstral parameters have been defined, namely the Cepstral Peak Prominence (CPP) and its smoothed version (CPPS). CPP is a measure (in dB) of the cepstral peak amplitude, normalized for overall signal amplitude through a linear regression line estimated relating quefrency to cepstral magnitude [17]. CPPS considers two smoothing steps before calculating the cepstral peak prominence [16]. The meta-analysis on correlation coefficients between acoustic measurements and perceptual evaluation of voice quality by Maryn $et\ al.$ [18] highlighted the relevance of CPPS: they found that CPPS satisfied the meta-analytic criteria in sustained vowels as well as in continuous speech. CPPS has also resulted well correlated with perceptual judgement of overall grade of dysphonia and different types of voice quality [19]-[20]. Additionally, significantly different CPPS values between dysphonic and control group have been found in the vowel /a/ [21]. Despite the attention given to the parameter, in the existing literature there is a lack of investigation on CPPS diagnostic precision. Such analysis has been performed for the Acoustic Voice Quality Index (AVQI), which is a multivariate construct that includes CPPS and other four acoustic metrics [22]. All the above-mentioned studies used cepstrum software packages to estimate CPPS, which only provide the mean of CPPS values and in some cases the standard deviation: the most popular packages are Praat [23], SpeechTool [24] and the Analysis of Dysphonia in Speech and Voice module [25] of Multi-Speech from KayPENTAX (Montvale, NJ). These programs process signals acquired with microphones in air only.

In recent years, the diffusion of in-field long-term monitorings instead of in-clinic short-term measurements has been providing distributional parameters that are able to characterize the vocal behavior [26]. Proper devices for such voice monitoring have been developed: the NCVS dosimeter [27], the VoxLog [28], the Ambulatory Phonation Monitor [29], the Voice Care [30]-[33] and a smartphone-based platform [34]. The main advantage of these devices is the use of a contact sensor for the acquisition of the voice signal: it has a very limited sensitivity with respect to background noise levels and it does not impair the subject activity. A recent work by Mehta $et\ al.$ [35] evaluated CPP from vowels acquired with a microphone in air and an accelerometer

sensor using a commercially available program. They found that CPP measures from the two sensors were highly correlated, without significant differences between healthy and pathological voice.

The present paper investigates CPPS distributions in sustained vowel /a/ and their descriptive statistics as discriminators between healthy and unhealthy voices. Descriptive statistics different than the mean have been considered as possible candidate that could exhibit higher discrimination power. Signals acquired with two types of microphones have been included in the analysis, that are a headworn microphone and a contact Electret Condenser Microphone (ECM). A first uncertainty contribution that has been taken into account is related to the repeatability of a subject in performing the speech task. This contribution, which has been estimated as the intra-speaker variability of CPPS parameters in repeated sessions, has been used to assess the uncertainty of the threshold values between healthy and unhealthy voices by means of the Monte Carlo method. Preliminary results have been discussed in [36], while the present paper reports updated outcomes and the results of further investigations. The main influence quantities of the estimated cepstral parameters have been identified, which are the fundamental frequency of the vocalization and the broadband noise superimposed to the signal, providing recommendations for improving the accuracy of future studies. In addition, the reliability of CPPS estimation with respect to the frequency content of the vocal spectrum has been evaluated, which is mainly dependent on the bandwidth of the measuring chain used to acquire the vocal signal.

II. METHOD

A. Subjects

Fourty-one voluntary patients, 30 females and 11 males, participated in this study (age range: 20-77 years; mean: 51 years; standard deviation SD: 18.1 years). Thirty-five healthy adults with normal voices, 12 females and 23 males, were also included in the experiment (age range: 21-58 years; mean: 29 years; SD: 11.1 years). A clinical protocol that included a careful case history, auditory-perceptual measures, and videostroboscopy, was followed for all the participants, who were all native Italian speakers. Table I shows the otolaryngologic diagnoses in the patient group.

B. Procedure

The protocol was designed in order to avoid each step affecting the following one. The relevant steps of the procedure can be summarized as follows:

TABLE I

DIAGNOSES FOR THE PATIENT GROUP.

Organic dysphonia	**Patients**
Cyst	8
Edema	10
Sulcus vocalis	3
Polyp	4
Chronic laryngitis	4
Vocal fold hypostenia	3
Vocal fold paresis	2
Vocal fold nodul	2
Neurological disorder	3
Post-surgery dysphonia	2

1) each participant was asked to vocalize the sustained vowel /a/ on a comfortable pitch and loudness until he/she had need to breathe again, while he/she worn a headworn microphone and a contact microphone simultaneously;
2) participants repeated the previous task other two times, waiting few seconds of silence between the repetitions
3) two otolaryngologists performed the clinical practice that included a careful case history, auditory-perceptual measures (GIRBAS scale) and the videolaringoscopy examination.

The vowel /a/ was selected as speech material due to its large use in acoustic analysis of voice and the duration of each phonation was always longer than 2 s, as recommended in [38].

C. Equipment for recording procedure

The voice recordings were performed in a quiet room, where the A-weighted equivalent background noise level was measured with a calibrated class-1 sound level meter (NTi Audio XL2) over a period of 5 minutes in four different days, obtaining the average value of 50.0 dB (SD = 2.0 dB). Before performing the tasks described in steps (1) and (2), subjects worn the two microphones, that were:

- an omni-directional headworn microphone Mipro MU-55HN, which was placed at a distance of about 2.5 cm from the lips' edges of the talker, slightly to the side of the mouth. The microphone, which exhibits a flatness of 3 dB in the range from 40 Hz to 20 kHz, was

TABLE II

NUMBER OF SUBJECTS WHO UNDERTOOK THE EXPERIMENTS WITH THE DIFFERENT DEVICES MIPRO MU-55HN HEADWORN MICROPHONE AND ECM AE38 CONTACT MICROPHONE. NUMBER OF PATIENTS AND CONTROLS AND FEMALES (F) AND MALES (M) ARE ALSO REPORTED.

	Mipro MU-55HN			**ECM AE38**		
	F	M	Overall	F	M	Overall
Patients	30	11	41	28	6	34
Controls	12	23	35	12	23	35
Overall	42	34	76	40	29	69

connected to a bodypack transmitter ACT-30T, which transmits to a wireless system Mipro ACT 311. The output signal of this system was recorded with an handy recorder ZOOM H1 (Zoom Corp., Tokyo, Japan), that use a sample rate of $44.1\ \mathrm{kSa/s}$ and 16 bit of resolution;

- an Electret Condenser Microphone (ECM AE38 [Alan Electronics GmbH (Dreieich, Germany)]), which was fixed at the jugular notch of each talker by means of a surgical band. The microphone senses the skin vibrations induced by the vocal-fold activity and it was connected to the handy recorder ROLAND R05 (Roland Corp., Milano, Italy), that samples the signal at a rate of $44.1\ \mathrm{kSa/s}$ using 16 bit of resolution.

Table II shows the details related to the subjects who performed the experimental task with the two microphones.

D. Data processing

Data were transferred from the handy recorders to a Personal Computer in order to be post-processed. First, the phonation interval from 1 s to 6 s has been selected for each sustained vowel, using the software Adobe Audition (version 3.0). Then, a specific MATLAB (R2014b, version 8.4) script, developed by the authors, has been used to estimate the Cepstral Peak Prominence Smoothed (CPPS) following the procedure described by Hillenbrand [16]. The selected signal was down-sampled to $22050\ \mathrm{Sa/s}$ and CPPS has been estimated every 2 ms (frame) using a 1024-point (46 ms) analysis window. For each window, the Fast Fourier Transform (FFT) algorithm has been implemented twice in order to obtain the spectrum amplitude at the first step and then the cepstrum from it. Before extracting the cepstral peak, a two smoothing steps procedure has been performed as follows: the smoothing in time averages cepstra using a time-

window of 14 ms (7 frames) and then the smoothing in cepstrum averages cepstral-magnitude across quefrency with a seven-bin window. On the smoothed cepstrum, a regression line has been estimated in the quefrency vs cepstral magnitude domain without considering the first millisecond, as suggested in [17]. Quefrencies below 1 ms are more affected by the spectral envelope, which varies slowly, than by the spectrum periodicity [37], so they have not been considered in the regression line evaluation. The Cepstral Peak Prominence Smoothed (CPPS) has been calculated as the difference in dB between the peak in the cepstrum and the value on the regression line at the same quefrency. The cepstral peak has been searched in the range from 3.3 ms to 16.7 ms, since the quefrency corresponding to the cepstral peak is the reciprocal of the fundamental frequency and the respective values of 60 Hz and 300 Hz match the usual range of fundamental frequency in adults.

A time series of 2500 CPPS values (5000 ms/2 ms) has been obtained for each speech sample, which is treated as a distribution. Examples of CPPS distributions for pathological and healthy voices can be found in [36]. For each CPPS distribution, the following descriptive statistics have been calculated: mean ($CPPS_{\mathrm{mean}}$), median ($CPPS_{\mathrm{median}}$), mode ($CPPS_{\mathrm{mode}}$), 5th percentile ($CPPS_{5\mathrm{prc}}$) and 95th percentile ($CPPS_{95\mathrm{prc}}$) as measures of location of the distribution; standard deviation ($CPPS_{\mathrm{std}}$) and the interval between the maximum and the minimum value ($CPPS_{\mathrm{range}}$) as measures of its variance, kurtosis ($CPPS_{\mathrm{kurt}}$) and skewness ($CPPS_{\mathrm{skew}}$) for the characterization of distribution shape.

E. Analyses

1) CPPS parameters in healthy and unhealthy voices: the two-tailed Mann-Whitney U-test [39] has been used to investigate statistical differences between each coupled list of descriptive statistics related to the patient group and the control subjects. It is a non-parametric test that refers to independent samples: the null hypothesis (H0) states that MD = 0, where MD is the median of the population of the differences between the sample data for patients and controls. When the null hypothesis is accepted, the two lists of values seem to come from the same population, i.e. it is not possible to distinguish healthy and unhealthy samples. The one-sample Kolmogorov-Smirnov test has been performed to verify that data in each list are not normally distributed, with the exception for the kurtosis values of CPPS distributions ($CPPS_{\mathrm{kurt}}$) from patients. Such result allows the use of a non-parametric test for the analysis. The two above-mentioned tests have been performed using a MATLAB script.

2) Best logistic regression model: with the aim of investigating the effectiveness of the descriptive statistics for CPPS distribution as discriminators between dysphonic and healthy voices, a binary classification approach has been followed: a dichotomous variable, which has been coded as 0 or 1, has been given to each individual value of the descriptive statistics for CPPS distribution depending on the absence or the presence of dysphonia, respectively. The absence or the presence of the voice problem has been determined by the outcome of the videolaringoscopy examination. Then, a single-variable logistic regression model has been performed for each descriptive statistic and the best model was selected based on the highest Mc Fadden's R^2 and Area Under Curve (AUC) [40]. The Mc Fadden's R^2 characterizes the predictive power of a logistic regression model, while the area under the Receiver Operating Characteristic (ROC) curve describes the classification accuracy of the model. Area Under Curve (AUC) ranges from 0.5 to 1.0: an AUC near to 1 indicates a strong model's ability to separate those subjects with vocal disorders from those who have a healthy voice, while an AUC close to 0.5 means that the model has a poor capability to discriminate between the two groups.

Furthermore, the best threshold for the classification of healthy and pathological voices has been selected, observing a graph where sensitivity and specificity versus each possible threshold are plotted. Sensitivity is the true positive rate, i.e. the quota of people with voice problems who are correctly classified as positive. Specificity is the true negative rate, that is the percentage of subjects with healthy normal voice who are correctly identified as negative. The authors privileged a greater true positive rate (sensitivity) in selecting the best threshold, instead of taking the usual threshold that corresponds to the crossing point of sensitivity and specificity curves. All the analyses related to the logistic regression model has been performed using the statistical program RStudio (Version 0.99.489).

3) Intra-speaker variability: the repeatability of the descriptive statistics for CPPS distribution that have been included in the empirical fitted models has been investigated. Sixty-one subjects performed correctly the second task described in paragraph II-B, while wearing both the head-worn microphone and the ECM. For these participants, CPPS distributions have been calculated in the three repetitions of the sustained vowel /a/.

4) Monte Carlo method: the uncertainty estimation of the threshold values obtained for each logistic model has been assessed using the Monte Carlo method. First, the best fitting distribution for the lists of CPPS parameters that were included in the models has been determined throught the Maximum Likelihood Estimation algorithm in MATLAB. This analysis has been performed

for both healthy and pathological voices, including CPPS parameters from the three repetitions of the vowel for each subject. Then, 1000 trials of the Monte Carlo method have been repeated by randomly sampling 50 values from each fitted distribution. For each trial the best threshold of the logistic model has been determined, setting the equality between the sensitivity and the specificity obtained from the ROC analysis.

5) Influence quantities: the effects of fundamental frequency and broadband noise as influence quantities of the CPPS have been investigated by feeding the script that estimates the CPPS statistics with synthesized signals with well known characteristics. A set of vowels /a/ with the fundamental frequency in the range of 80 Hz to 260 Hz (frequency step of 20 Hz) has been synthetically generated using the software Sopran [41] with a sampling rate of 22050 Sa/s. The selected frequencies cover both the typical female and male fundamental frequency range in sustained vowels of adults [42]. For each fundamental frequency, a 2 s long vowel has been created setting the first eight formants as pass-band filters with a Q factor of 20 and center frequencies of 580 Hz, 1.7 kHz, 2.9 kHz, 4.3 kHz, 5.4 kHz, 6.5 kHz, 7.7 kHz, 9.0 kHz. The Signal-to-Noise Ratio (SNR) of this set of vowels is of about 100 dB, which is mainly related to the quantization noise. Other two sets of vowels with the same frequency characteristics have been created adding two levels of random noise using MATLAB noise generator. A mean zero white Gaussian noise has been superimposed to the vowel signals setting the standard deviation in order to obtain SNR of 40 dB and 20 dB. For each fundamental frequency, CPPS distributions have been estimated by processing the 1 s long middle part of the vowel signal.

6) Frequency content of the spectrum: the 4 s middle part of a sustained vowel /a/ acquired with the headworn microphone from a control subject have been used in order to investigate the behavior of CPPS distributions and their statistics with different frequency contents. Starting from the full spectrum bandwidth of the signal, that is of about 11 kHz, a 500 Hz frequency content has been cut away at a time and CPPS computation has been repeated for each step. This operation has been done down to a bandwidth of 1 kHz.

III. Results

A. Microphone in air

The p-values obtained from the Two-tailed Mann-Whitney U-test of the lists of descriptive statistics related to the two groups of subjects were lower than 0.05, with the exception of skewness and kurtosis. These outcomes mean that the null hypothesis is rejected for most of

CPPS parameters: CPPS distributions are significantly different in location, with an average value of 15.2 dB and 18.2 dB for $CPPS_{\mathrm{mean}}$ in patients and controls, respectively, and in variance, with an average value of 1.9 dB and 1.3 dB for $CPPS_{\mathrm{std}}$ in pathological and healthy voices, respectively.

Assuming the presence/absence of voice disorders as dependent variable, the best logistic regression model between healthy and unhealthy voice includes $CPPS_{5\mathrm{prc}}$ as independent variable. The following formula defines the best empirical fitted model:

$$P(Unhealty) = \frac{e^{(28.8-1.93\cdot CPPS_{5\mathrm{prc}})}}{1+e^{(28.8-1.93\cdot CPPS_{5\mathrm{prc}})}} \tag{1}$$

where $P(Unhealthy)$ is the probability of having unhealthy voice, which ranges from zero to one. The negative coefficient of $CPPS_{5\mathrm{prc}}$ shows that the probability to have unhealthy voice decreases as the $CPPS_{5\mathrm{prc}}$ increases. A Mc Fadden's R^2 equal to 0.62 and an AUC of 0.95 of the model highlight that there is a clear separation between patients and controls: Fig. 1 shows the fitted values obtained for each subject and most of patients are in the upper part of the graph, where the probability of having unhealthy voice is near to one, while most of controls have lower scores, near to zero. The best classification threshold was $P(Unhealthy) = 0.44$, that corresponds to 15.0 dB in terms of $CPPS_{5\mathrm{prc}}$, with a sensitivity equal to 0.90 and a specificity of 0.94. As shown in Fig. 1, the four patients that are wrongly classified by the model have been judged with the lowest overall grade G of dysphonia.

The results on the repeatability of $CPPS_{5\mathrm{prc}}$ are summarized in Fig. 2. For each subject, it shows the average values and the relative experimental standard deviations of the CPPS parameter in the three repetitions of the vowel /a/ acquired with the headworn microphone. Among the patient group, a clear separation between the first two grades G of dysphonia is not highlighted in the figure, while the three patients with G=3 show $CPPS_{5\mathrm{prc}}$ lower than $8\,\mathrm{dB}$. The average of the standard deviations of the $CPPS_{5\mathrm{prc}}$ is equal to $0.8\,\mathrm{dB}$ for the patient group and $0.5\,\mathrm{dB}$ for the control group.

Fig. 2 also shows the threshold uncertainty, that is represented as a gray area around the $CPPS_{5\mathrm{prc}}$ threshold. The probability density functions of the best-fitted distributions of $CPPS_{5\mathrm{prc}}$ in pathological and healthy voices (bimodal and normal, respectively) have been used in a Monte Carlo simulation based on 1000 trials [36]. The output was a 95% confidence interval of the threshold equal to $0.7\,\mathrm{dB}$, which constitutes the width of the gray area in the figure 2.

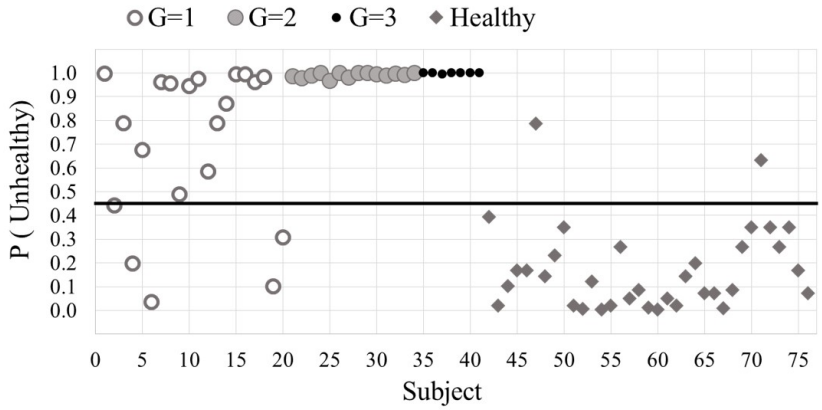

Fig. 1. Fitted values of the best logistic regression model, in terms of probability of having unhealthy voice, for vocalizations acquired with the headworn microphone Mipro MU-55HN. Circle points indicate the patient group (empty circles for the patients having a overall grade G of dysphonia equal to 1, gray circles for G=2 and black points for G=3); diamond points represent the control group. The bold line indicates the threshold value (0.44), which best separates patients and control subjects.

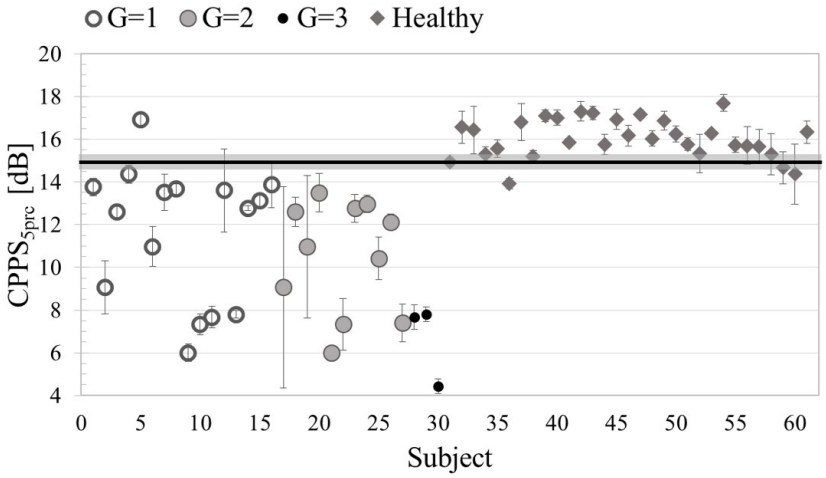

Fig. 2. Averaged values of $CPPS_{5\mathrm{prc}}$ in the three repetitions of the vowel for each subject, acquired with the headworn microphone Mipro MU-55HN. Circle points indicate the patient group with different grades of dysphonia; diamond points represent the control group. Bars indicate the experimental standard deviation for each subject. The bold line indicates the threshold value (15.0 dB) and the gray area corresponds to its 95% confidence interval.

B. Contact microphone

According to the outputs of the Two-tailed Mann-Whitney U-test, the lists of descriptive statistics for CPPS distributions related to the groups of patients and controls, who were recorded with the ECM, were significantly different in $CPPS_{\mathrm{mean}}$, $CPPS_{\mathrm{median}}$, $CPPS_{\mathrm{std}}$, $CPPS_{\mathrm{range}}$ and $CPPS_{5\mathrm{prc}}$ (p-values < 0.05). As a consequence, CPPS distributions resulted significantly different in location, e.g. the average $CPPS_{\mathrm{mean}}$ was equal to 18.0 dB for patients and 19.7 dB for controls, and in variance, e.g. the average $CPPS_{\mathrm{std}}$ was equal to 1.7 dB and 0.9 dB for patients and controls, respectively.

The following formula describes the best empirical fitted logistic model for vowels acquired with ECM, which uses $CPPS_{\mathrm{std}}$ as independent variable:

$$P(Unhealty) = \frac{e^{(-6.33+5.50\cdot CPPS_{\mathrm{std}})}}{1+e^{(-6.33+5.50\cdot CPPS_{\mathrm{std}})}} \tag{2}$$

where $P(Unhealthy)$ is the probability of having unhealthy voice, which ranges from zero to one. The positive coefficient of $CPPS_{\mathrm{std}}$ shows that the probability to have unhealthy voice increases as $CPPS_{\mathrm{std}}$ increases. The empirical model has a moderate discrimination power with a Mc Fadden's R^2 equal to 0.38 and an AUC of 0.87: Fig. 3 shows that the fitted values of the two groups are not clearly separated. The best classification threshold is $P(Unhealthy) = 0.43$, that corresponds to 1.1 dB in terms of $CPPS_{\mathrm{std}}$, with a sensitivity of 0.79 and a specificity of 0.69. Fig. 3 also shows that six out of seven patients that are wrongly classified by the model have been perceptually rated with the lowest overall grade G of dysphonia.

For each subject, the average values and the relative experimental standard deviations of $CPPS_{\mathrm{std}}$ in the three repetitions of the vowel /a/ acquired with the ECM are reported in Fig. 4. One should note that patients rated with G=1 have lower $CPPS_{\mathrm{std}}$ than those with G=2 and G=3. The average of the standard deviations of the $CPPS_{\mathrm{std}}$ is equal to 0.3 dB for the patient group and 0.2 dB for the control group.

The same numerical procedure described in III-A has been implemented in order to estimate the threshold uncertainty, where a bimodal and a lognormal probability density functions have been used for pathological and healthy voices, respectively. The output was a 95% confidence interval of 0.2 dB. This interval is represented as a gray area around the $CPPS_{\mathrm{std}}$ threshold in Fig.4.

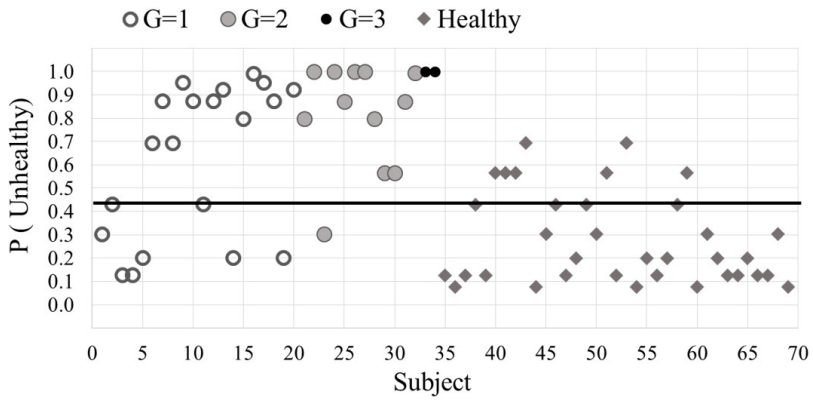

Fig. 3. The same of Fig. 1, for samples acquired with the contact microphone ECM AE38. The bold line indicates the selected threshold value, that is 0.43, which best separates patients and control subjects.

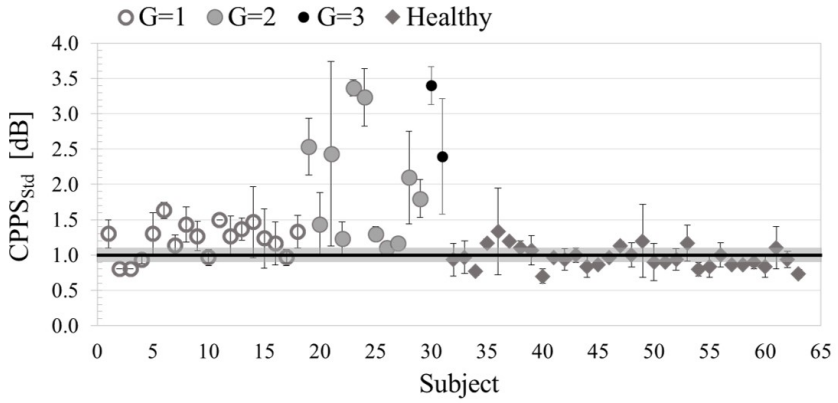

Fig. 4. Averaged values of $CPPS_{\mathrm{std}}$ in the three repetitions of the vowel for each subject, acquired with the contact microphone ECM AE38. Circle points indicate the patient group with different grades of dysphonia; diamond points represent the control group. Bars indicate the experimental standard deviation for each subject. The bold line indicates the threshold value (1.1 dB) and the gray area corresponds to its 95% confidence interval.

C. Influence quantities: fundamental frequency and noise

Fig. 5 shows the behavior of $CPPS_{5\mathrm{prc}}$ and $CPPS_{\mathrm{std}}$ corresponding to the sets of vowels /a/ that have been synthesized according to the procedure described in the section II-E5.

The estimated $CPPS_{5\mathrm{prc}}$ (red lines) shows a non monotonic behavior as the fundamental frequency increases for all of the three synthesized SNR levels. The standard deviation of the parameter $CPPS_{5\mathrm{prc}}$ in the investigated frequency range resulted in 1.3 dB, 1.6 dB and 1.3 dB

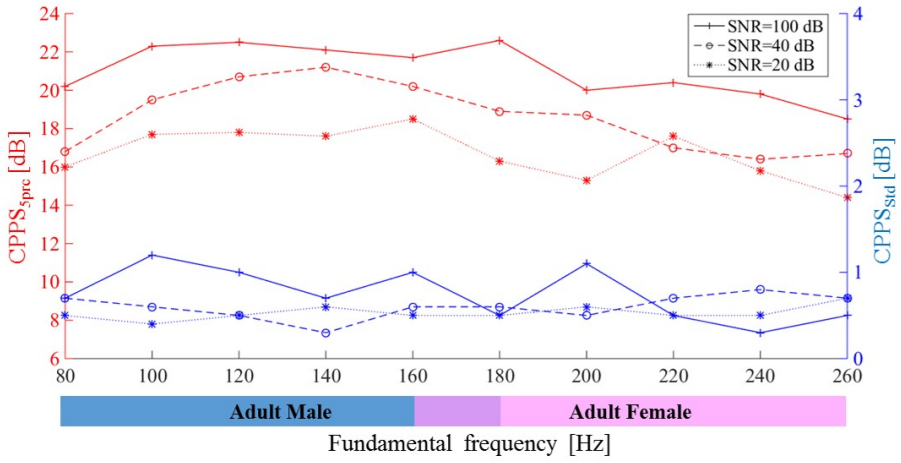

Fig. 5. Behavior of $CPPS_{5\mathrm{prc}}$ (red lines) and $CPPS_{\mathrm{std}}$ (blue lines) vs fundamental frequency, for three SNR levels (100 dB, 40 dB and 20 dB).

for SNR values equals to 100 dB, 40 dB and 20 dB, respectively. Hence the $CPPS_{5\mathrm{prc}}$ shows a moderate dependence on the fundamental frequency, which is of the same order of magnitude of the estimated uncertainty of the discrimination threshold between healthy and unhealthy voices. However, the estimated standard deviation refers to a frequency range that includes both male and females voices, then lower variability is obtained by separating the two frequency ranges. In addition, it is possible to strongly reduce the observed variability by limiting the field of use of the fundamental frequency: from a practical point of view, this could be implemented by providing a reference frequency to the subject before he/she produces the sustained vowel. With respect to the SNR level, the three $CPPS_{5\mathrm{prc}}$ curves are clearly separated: the one related to the highest SNR (100 dB) is above the other two curves, with an average value of 20.6 dB, while the one related to the noisiest signal (SNR of 20 dB) exhibits an average value of 16.3 dB. These findings confirm that the amplitude of the cepstral peak is dependent on the depth of the valleys between adjacent harmonics: higher the noise content in the spectrum shorter the height of the peak amplitude in the cepstrum [43]-[44].

The parameter $CPPS_{\mathrm{std}}$ (blue lines) vs the fundamental frequency is seemingly flat for the signals with SNR of 40 dB and 20 dB, while it exhibits an up-down trend when SNR is equal to 100 dB. Furthermore, $CPPS_{\mathrm{std}}$ tends to rise as SNR increases: its average value in the investigated frequency range is 0.7 dB (standard deviation 0.3 dB) for $SNR = 100$ dB, 0.6 dB

(s.d. 0.1 dB) for $SNR = 40$ dB and 0.5 dB (s.d. 0.1 dB) for $SNR = 20$ dB. This outcome proves that CPPS distributions have a higher variation when negligible noise is superimposed to the vocal signal.

One should note that the obtained values for the parameters $CPPS_{\mathrm{5prc}}$ and $CPPS_{\mathrm{std}}$ correspond to a healthy voice, since the former is higher than the identified threshold of 15.0 dB and the latter is lower than the threshold of 1.1 dB. This result, which is valid regardless of the effects of the investigated influence quantities, confirms the effectiveness of the proposed method, since synthesized vowels correspond to really healthy voices.

A further consideration can be made that is related to the differences of $CPPS_{\mathrm{5prc}}$ and $CPPS_{\mathrm{std}}$ between female and male typical fundamental frequency ranges. As shown in Fig. 5, adult male range is typically assumed from 80 Hz to 180 Hz, while adult female fundamental frequency is in the range from 160 Hz to 260 Hz. As highlighted before, $CPPS_{\mathrm{5prc}}$ curves have a slight downtrend as fundamental frequency increases. This seems confirmed by the results reported in the upper part of Fig. 6, since for the three investigated SNR levels the average of $CPPS_{\mathrm{5prc}}$ is higher in the male range than in the female one. However, there is no significant difference between the two mean values of genders, since the standard deviations corresponding to the two frequency ranges overlap. The bottom part of Fig. 6 shows the behavior of $CPPS_{\mathrm{std}}$ in male and female fundamental frequency ranges: also in this case, no significant differences have been found, even though the average $CPPS_{\mathrm{std}}$ is higher in the male range than in the female one for $SNR = 100$ dB, while the opposite behavior is observed for the other two SNR levels.

D. Frequency content of the spectrum

Fig. 7 shows how $CPPS_{\mathrm{5prc}}$ (red line) and $CPPS_{\mathrm{std}}$ (blue line) change when they are estimated from a healthy vowel /a/ whose spectrum has different frequency contents, starting from 11 kHz down to 1 kHz. Both the parameters have small variations between 11 kHz and 5 kHz, then $CPPS_{\mathrm{5prc}}$ increases reaching its maximum value for a frequency content of 3 kHz and it decreases again down to 1 kHz. The spectrum magnitude of the vowel under analysis, which is reported in the upper part of Fig. 7, highlights that the harmonic components between 5 kHz and 11 kHz have a limited energy content. In other words, these components contribute to the overall periodicity of the spectrum in a negligible way, so $CPPS_{\mathrm{5prc}}$ keeps quite constant down to 5 kHz (the dotted black vertical line helps in reading the graphs). If instead the frequency content of the

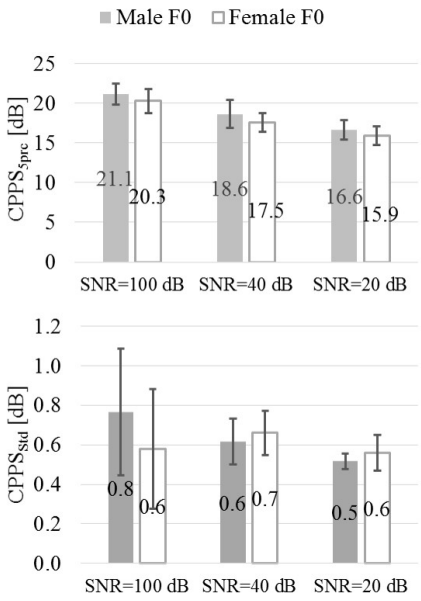

Fig. 6. Average values of $CPPS_{5\mathrm{prc}}$ (upper part) and $CPPS_{\mathrm{std}}$ (bottom part) in male and female frequency ranges; bars indicate the confidence interval obtained with a coverage factor $k = 2$.

spectrum is limited to 3 kHz, sharp and clear harmonic components are deleted, which have an important role in the definition of the spectrum periodicity: for this reason $CPPS_{5\mathrm{prc}}$ increases between 5 kHz and 3 kHz. Eventually, the parameter $CPPS_{5\mathrm{prc}}$ decreases between 3 kHz and 1 kHz because of the limited number of harmonic components included in the spectrum.

Differently from $CPPS_{5\mathrm{prc}}$, $CPPS_{\mathrm{std}}$ has a downward trend between 5.5 kHz and 3 kHz and it tends to have an up-down trend around a constant value again where the spectrum has a frequency content lower than 3 kHz. The reasons of such a change of behavior can be found in the previous observations about the spectrum periodicity. Fig. 7 also shows the frequency content of the signals acquired with the headworn microphone (MIC) and the ECM, which are respectively 10 kHz (vertical red dashed line) and 3.5 kHz (vertical blue dashed line). As we can observe in the graph at the bottom of the figure, the $CPPS_{5\mathrm{prc}}$ has been estimated where its behavior with the frequency content of the signal is almost stable, while $CPPS_{\mathrm{std}}$, which is calculated from the ECM signal, has been estimated in the region of its high variability with respect to the frequency content.

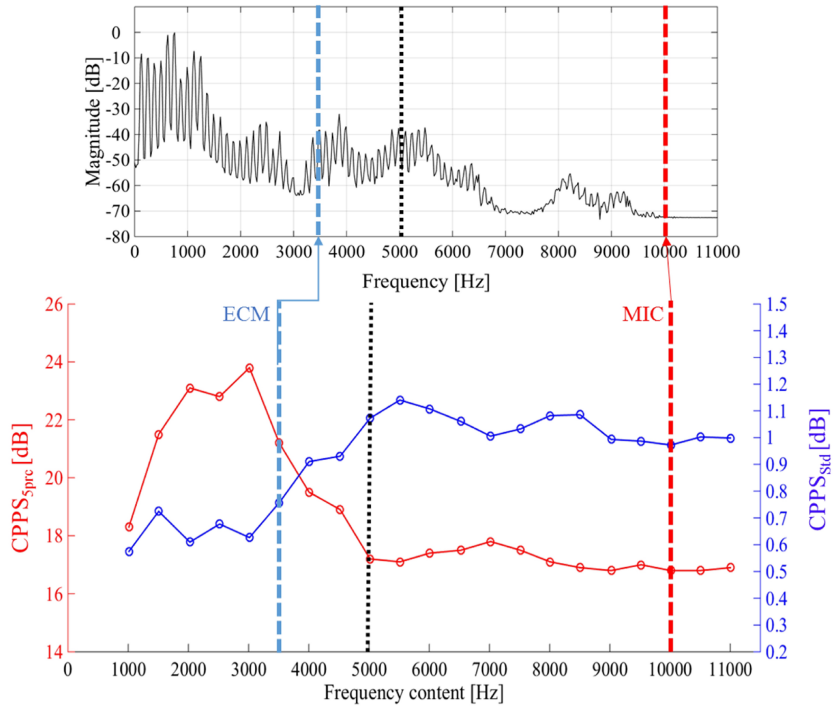

Fig. 7. (Bottom part) - Behavior of $CPPS_{5\mathrm{prc}}$ (red line) and $CPPS_{\mathrm{std}}$ (blue line) vs frequency content of the spectrum. (Upper part) - Spectrum magnitude of the vowel under investigation, acquired with the headworn microphone. Vertical dashed lines correspond to the frequency content of signals acquired with the ECM (blue line) and with the headworn microphone (red line). Vertical dotted black lines helps in reading the graphs.

IV. Conclusions

This paper investigates individual distributions of Cepstral Peak Prominence Smoothed (CPPS) and their descriptive statistics as possible indicators of vocal health. CPPS distributions have been obtained from sustained vowels /a/ vocalized by a group of patients and a group of controls and acquired with a microphone in air and a contact sensor (ECM). Regarding the speech material acquired with the microphone in air, the fifth percentile ($CPPS_{5\mathrm{prc}}$) resulted the best descriptive statistic for CPPS distributions that is able to discriminate healthy and unhealthy voices. The respective empirical logistic model shows a strong discrimination power ($AUC = 0.95$) and a discrimination threshold of $CPPS_{5\mathrm{prc}}$=15.0 dB, with lower values indicating unhealthy status of voice. Concerning the sustained vowels acquired with the ECM, instead, the standard deviation ($CPPS_{\mathrm{std}}$) was the best parameter that separates the two groups. The respective empirical logistic model has a good discrimination power, with AUC of 0.87, and a discrimination threshold of

$CPPS_{\mathrm{std}}$=1.1 dB, with larger values for pathological voice. Differently from the results by Mehta $etal.$ [30], the proposed method is able to discriminate healthy and unhealthy voice from both the microphone in air and a contact microphone. As expected, the intra-speaker variability of the two CPPS parameters was larger in the patients group than in the control one: its respective values were 0.8 dB and 0.5 dB for $CPPS_{5\mathrm{prc}}$ and 0.3 dB and 0.2 dB for $CPPS_{\mathrm{std}}$. This result highlights the limited capability of patients in the vocal production.

The uncertainty of the discrimination threshold for the two parameters $CPPS_{5\mathrm{prc}}$ and $CPPS_{\mathrm{std}}$ has been also estimated: the 95% confidence intervals were 0.7 dB and 0.2 dB, respectively, thus showing that its contribution is negligible with respect to the variability of each subject.

With the aim of providing guidelines that make the estimated CPPS parameters reliable, an analysis of the main CPPS influence quantities has been performed. The obtained outcomes highlighted that the fundamental frequency and the SNR level of the acquired signals could significantly affect the discrimination between healthy and pathological voices. For this reason, it is important to limit the field of use of the fundamental frequency, $e.g.$ providing a reference tone to the subject before he/she performs the speech task, and to avoid large difference in the SNR level during the experimental campaign.

Further investigations have been made in order to estimate the effect of the frequency content of the signal spectrum on the CPPS parameters. As the result of this analysis, it can be stated that a reliable estimation of the parameters $CPPS_{5\mathrm{prc}}$ and $CPPS_{\mathrm{std}}$ is obtained provided that the frequency content of the spectrum is not lower than 5 kHz. This justifies the lower discrimination power obtained for the contact microphone that showed a frequency content of about 3.5 kHz.

REFERENCES

[1] V. Parsa, D.G. Jamieson, *Acoustic discrimination of pathological voice: sustained vowels versus continuous speech*, J. Speech Lang. Hear. Res., vol. 44, pp. 327-339, 2001.

[2] L.G. Portney, M.P. Watkins, *Foundations of Clinical Research: Applications to Practice*, 2 Ed., Upper Saddle River, Prentice-Hall, 2000.

[3] L. Ding, A. Radwan, M.S. El-Hennawey and R.A. Goubran, *Measurement of the Effects of Temporal Clipping on Speech Quality*, IEEE Tr. on IM, vol. 55(4), pp. 1197-1203, 2006.

[4] L. Cai, R. Tu, and J. Zhao, *Speech Quality Evaluation: A new Application of Digital Watermarking*, IEEE Tr. on IM, vol. 56(1), pp. 45-55, 2007.

[5] F. Babich, R. Passini, E. Valentinuzzi and F. Vatta, *Development of a Test-Bench for Objective Speech Quality Measurements under Different Simulated Fading Conditions*, in Proc. IEEE I2MTC, Venezia, Italy, May 24-26, 1999, pp. 601-606.

[6] D.L. Carn, D. Grimaldi, *Voice quality measurement in telecommunication networks by optimized multi-sine signals*, Measurement, vol. 41, pp. 266-273, 2008.

[7] M. Bertocco, P. Paglierani, E. Rizzi, *On the Use of In-service, Nonintrusive Measurement Devices in the Performance Analysis of Telephone-Type Networks*, in Proc. IEEE I2MTC, Baltimore, MD (USA), May 1-4, 2000, pp. 86-89.

[8] L. Benetazzo, M. Bertocco, P. Paglierani, E. Rizzi, *Speech/Voice-Band Data Classification for Data Traffic Measurements in Telephone-Type Systems*, IEEE Tr. on IM, 49(2), pp. 413-417, 2000.

[9] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, *A review of depression and suicide risk assessment using speech analysis*, Speech Communication, vol. 71, pp. 10-49, 2015.

[10] W.D. Duckitt, S.K Tuomi, T.R. Niesler, *Automatic detection, segmentation and assessment of snoring from ambient acoustic data*, Physiol Meas., vol. 27(10), pp. 1047-56, 2006.

[11] D.L. Herath, U.R. Abeyratne, C. Hukins, *Hidden Markov modelling of intra-snore episode behavior of acoustic characteristics of obstructive sleep apnea patients*, Physiol Meas., vol. 36(12), pp. 2379-404, 2015.

[12] E.H. Buder, *Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990*, In: R.D. Kent, M.J. Ball, eds. Voice Quality Measurement, San Diego, CA: Singular Publishing Group, 2000, pp. 119-244.

[13] S. Bielamowicz, J. Kreiman, B.R. Gerratt, M.S. Dauer and G.S. Berke, *Comparison of voice analysis systems for perturbation measurement*, Journal of Speech and Hearing Research, vol. 39, pp. 126-134, 1996.

[14] Y. Zhang and J.J. Jiang, *Acoustic analyses of sustained and running voices from patients with laryngeal pathologies*, J. Voice, vol. 22, pp. 1-9, 2008.

[15] S.Y. Lowell, R.H. Colton, R.T. Kelley, Y.C. Hahn, *Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker*, J. Voice, vol. 25, pp. 223-232, 2011.

[16] J. Hillenbrand, R.A. Houde, *Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech*, J. Speech Hear. Res., vol. 39, no. 2, pp. 311-21, 1996.

[17] J. Hillenbrand, R.A. Cleveland, R.L. Erickson, *Acoustic correlates of breathy vocal quality*, J. Speech Hear. Res., vol. 37, pp. 769-778,1994.

[18] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, P. Corthals, *Acoustic measurement of overall voice quality: a meta-analysis*, J. Acoust. Soc. Am., vol. 126, pp. 2619-2634, 2009.

[19] Y.D. Heman-Ackah, R.J. Heuer, D.D. Michael, R. Ostrowski, M. Horman, M.M. Baroody, J. Hillenbrand, R.T. Sataloff, *Cepstral peak prominence: a more reliable measure of dysphonia*, Ann. Otol. Rhinol. Laryngol., vol.112, pp. 324-333, 2003.

[20] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, T. Haderlein, *Vowel- and text-based cepstral analysis of chronic hoarseness*, J. Voice, vol. 26, pp. 416-424, 2012.

[21] L.F. Brinca, P.F. Batista, A.I. Tavares, I.C. Goncalves, M.L. Moreno, *Use of Cepstral Analyses for Differentiating Normal From Dysphonic Voices: A Comparative Study of Connected Speech Versus Sustained Vowel in European Portuguese Female Speakers*, J. Voice, vol. 28, no. 3, pp. 282-286, 2014.

[22] Y. Maryn, M. De Bodt, N. Roy, *The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders*, J. Commun. Disord.,vol. 43, pp.161-174, 2010.

[23] P. Boersma and D. Weenink, Institute of Phonetic Sciences, University of Amsterdam, The Netherlands, http:// www.praat.org/. (last view: 21/06/2017).

[24] J.M. Hillenbrand, J.M. Hillenbrand Homepage, *[Online] Available at: http://homepages.wmich.edu/ hillenbr/* (last view: 21/06/2017).

[25] S.N. Awan, N. Roy, M.E. Jette, G.S. Meltzner, R.E. Hillman, *Quantifying dysphonia severity using a sepctral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V*, Clin. Linguist. Phon., vol. 24, pp.742-758, 2010.

[26] M. Ghassemi, J.H. Van Stan, D.D. Mehta, M. Zanartu, H.A. Cheyne, R.E. Hillman, J.V. Guttag, *Learning to Detect Vocal*

Hyperfunction From Ambulatory Neck-Surface Acceleration Features: Initial Results for Vocal Fold Nodules, IEEE Tr. on Biomedical Engineering, vol. 61, no. 6, pp. 1668-1675, 2014.

[27] P.S. Popolo, J.G. Svec, and I.R. Titze, *Adaptation of a Pocket PC for Use as a Wearable Voice Dosimeter*, Journal of Speech Language and Hearing Research, vol. 48, pp. 780-791, 2005.

[28] VoxLog portable voice meter *[online] available: http://www.sonvox.com/index.html.* (last view: 31/10/2016).

[29] H.A. Cheyne, H.M. Hanson, R.P. Genereux, K.N. Stevens and R.E. Hillman, *Development and Testing of a Portable Vocal Accumulator*, J. of Speech Lang. and Hear Research, vol. 46, pp. 1457-1467, 2003.

[30] A. Carullo, A. Penna, A. Vallan, A. Astolfi, and P. Bottalico, *A portable analyzer for vocal signal monitoring*, in Proc. IEEE I2MTC, Graz, Austria, May 13-16, 2012, pp. 2206-2211.

[31] A. Carullo, A. Vallan, and A. Astolfi, *Design Issues for a Portable Vocal Analyzer*, IEEE Tr. on IM, vol. 62(5), pp. 1084-1093, 2013.

[32] A. Carullo, A. Vallan, and A. Astolfi, *A Low-Cost Platform for Voice Monitoring*, in Proc. IEEE I2MTC, Minneapolis, MN (USA), May 6-9, 2013, pp. 67-72.

[33] A. Carullo, A. Vallan, A. Astolfi, G.E. Puglisi, L. Pavese, *Validation of calibration procedures and uncertainty estimation of contact-microphone based vocal analyzers*, Measurement, vol. 74, pp. 130-142, 2015.

[34] D.D. Mehta, M. Zaartu, S.W. Feng, H.A Cheyne, R.E. Hillman, *Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform*, IEEE Tr. on Biomedical Engineering, vol. 59, no. 11, pp. 3090-3096, 2012.

[35] D.D. Mehta, J.H. Van Stan, R.E. Hillman, *Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer*, IEEE/ACM Trans Audio Speech Lang Process., vol. 24, no. 4, pp. 659-668, 2016.

[36] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, M. Spadola Bisetti and J. Colombini, *Cepstral Peak Prominence Smoothed distribution as discriminator of vocal health in sustained vowel*, in Proc. IEEE I2MTC, Torino, Italy, May 22-25, 2017, pp. 552-557.

[37] G. de Krom, *A cepstrum-based technique for determing a harmonics to noise ratio in speech signals*, J. S. Hear. Res., vol. 36, pp. 254-266, 1993.

[38] R.F. Coleman, *Sources of variation in phonetogram*, J. Voice, vol. 7, pp. 1-14, 1993.

[39] J.D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, Taylor Francis, 2003, pp. 215-223.

[40] D. Hosmer, S. Lemeshow, R. Sturdivant, *Applied Logistic Regression*, third edition, Wiley, 2013.

[41] Sopran, Tolvan Data *[online] available: http://www.tolvan.com/index.php?page=/sopran/sopran.php*(last view:17/06/2017)

[42] H. Goy, D.N. Fernandes, M.K. Pichora-Fuller, and P. van Lieshout, *Normative Voice Data for Younger and Older Adults*, J. Voice, vol. 27, no. 5, pp. 545-555, 2013.

[43] J. Murphy, *On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals*, J. Acoust. Soc. Am., vol. 120, no. 5, pp. 2896-2907, 2006.

[44] R. Fraile, J.I. Godino-Llorente, *Cepstral peak prominence: A comprehensive analysis*, Biomedical Signal Processing and Control, vol. 14, pp. 42-54, 2014.