Biomedical Signal Processing and Control 42 (2018) 134–144

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A methodology for voice classification based on the personalized fundamental frequency estimation

Laura Verde[a], Giuseppe De Pietro[b], Giovanna Sannino[b,*]

[a] Department of Engineering, University of Naples "Parthenope", Centro Direzionale, Isola C4, Naples, Italy
[b] National Research Council of Italy, Institute of High Performance Computing and Networking (CNR-ICAR), Via Pietro Castellino, 111 Naples, Italy

ARTICLE INFO

Article history:
Received 30 May 2017
Received in revised form
21 December 2017
Accepted 17 January 2018

Keywords:
Voice disorders
Signal processing
Fundamental frequency analysis
m-Health system

ABSTRACT

Nowadays, the incidence of voice disorders is increasing rapidly, with about a third of the population suffering from dysphonia at some point in their lives. Dysphonia is a disorder that alters vocal quality and can impair and reduce the quality of life. The structural or functional alteration of the phonatory apparatus, unhealthy lifestyles or an excessive use of the vocal cords for work activities (e.g. teaching) can cause voice disorders. Unfortunately, people who suffer from dysphonia often underestimate its symptoms and therefore delay consulting a speech therapist for accurate voice assessment and treatment. Voice disorder evaluation involves a series of tests, including an acoustic analysis. This quantifies the measurements of voice quality through the evaluation of certain characteristic parameters, for example the fundamental frequency (F_0). In this paper, a personalized methodology for the estimation of the F_0 is presented. The personalization is accomplished by taking into account two of the main factors that influence the F_0, the gender and age of the subject. The estimation of the F_0 is crucial for the classification of the voice signal, because the discrimination of a healthy voice from a pathological one is achieved by evaluating the inclusion of the F_0 value within the healthy range. To evaluate the presented methodology, we have carried out a set of tests by using some voice signals selected from an available database in order to compare the classification ability of the proposed methodology with other algorithms existing in the literature. The numerical results obtained show that the proposed methodology provides a good accuracy, sensitivity, and specificity, respectively of over 77%, 72% and 81%, values better than those achieved by the most frequently other used and cited fundamental frequency estimation algorithms. Additionally, a statistical analysis to evaluate whether or not a statistically significant difference exists between the accuracy, sensitivity and specificity has been carried out. The outcome of the ANOVA tests and of the *t*-tests confirms that there is a significant difference between the proposed methodology and the other algorithms. Finally, the presented methodology could be embedded in a portable and simple m-health application that could be useful for the monitoring of the state of vocal health and the prevention of voice disorders.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Dysphonia is a voice disorder that affects about 29% (considering a lifetime prevalence) of the population, impacting on their social, psychological and professional life [1], with in general a prevalence in the population aged 60 years or more [2]. It is an alteration of the sound structure of the voice that can be understood as a reduction in the acoustic energy, or a variation of the melodic component or tonal harmonic structure.

There are several potential causes of voice disorders. Although they may be hereditary, some voice disorders are caused by disease, for example by a common infectious illness (e.g. sinusitis or bronchitis), a chronic medical disease (e.g. sarcoidosis, hypothyroidism or rheumatoid arthritis), an inflammatory condition resulting, for example, from smoking or a neurological condition (e.g. vocal cord paralysis or Parkinson's disease). Usually the etiology of the voice disorder can be related to an incorrect use of the vocal cords, or an inappropriate speaking or breathing technique, possibly leading to voice alterations [3,4].

People who routinely use their voice in their jobs are particularly prone to vocal disorders. As reported in many studies [5],

* Corresponding author.
E-mail addresses: laura.verde@uniparthenope.it (L. Verde), giuseppe.depietro@icar.cnr.it (G. De Pietro), giovanna.sannino@icar.cnr.it (G. Sannino).

https://doi.org/10.1016/j.bspc.2018.01.007
1746-8094/© 2018 Elsevier Ltd. All rights reserved.

professional voice users have a greater possibility of contracting dysphonia, aphonia, edema, polyps or nodules, compared to non-vocal professionals (51.2% vs 27.4%). Teachers represent the main category of voice users seeking medical help for voice problems. Between 20% and 80% of teachers have reported various vocal symptoms, yet only 13.5% are aware of the risks of dysphonia during education or training activities. In fact, many teachers underestimate their vocal symptoms and do not necessarily take appropriate countermeasures.

In addition to its impact on health and quality of life, severely limiting communication at work and affecting all social aspects of daily activities, dysphonia results in the sufferer making frequent medical appointments with a consequent loss of productivity due to absenteeism which may undermine work functions. Teachers, for example, miss a significantly high number of working days throughout their career due to vocal problems. A European study [6] reports that one out of five teachers (19.2%) reported missing at least one day of work because of voice-related dysfunction. As reported in [7], approximately 10% of the U.S population has experienced dysphonia and it is estimated that the social cost incurred by problems such as unemployment caused by dysphonia amounts to $2.5 billion each year for the teaching profession alone.

One of the most prevalent voice disorders is Reinke's edema, a chronic, diffuse, inflammatory disease of the vocal folds [8]. As demonstrated in the study detailed in [4], Reinke's edema is the third most common cause of dysphonia with an incidence of 14%, and appears almost exclusively in chronic smokers older than 40. Its main impact is on women [9], causing the voice to become hoarse and low in pitch. Several studies [10,11] have confirmed the widespread prevalence of this disorder, a prevalence that has boosted the interest of the medical partner involved in this project, encouraging us to focus this study on voices suffering from this disease.

The lack of awareness of dysphonia, and of the causes of these disorders, is a potential barrier to appropriate treatment meaning that in some circumstances it is not possible to achieve a complete resolution [12].

Based on these considerations, this paper presents a personalized methodology, an improvement of the currently existing one, described in [13], for the performance of a smart, easy and quick screening of the voice to discriminate between a possibly pathological and a healthy voice that could be embedded in a smart mobile phone.

Currently, the proposed methodology evaluates only the Fundamental Frequency (F_0) of the speech signal, the main parameter estimated to evaluate voice disorders, variations of which are indicative of the patient's state of health, characterizing quali-quantitatively a specific vocal dysfunction.

The innovation of the methodology presented is the personalization of the algorithm to calculate the F_0, that takes into account two of the main factors that influence this parameter, the gender and age of the subject [14–16]. In more detail, we have finalized a method, based on the exhaustive search algorithm [17], a more accurate evaluation of the F_0. More accurate means that, considering this evaluation of the F_0, we obtain have had the opportunity to classify in a more accurate way (so obtaining fewer false positives and false negatives compared with other classifiers) the speech signal as healthy or pathological.

2. Background

Voice alteration is rarely associated with disease symptoms but often with a temporary alteration due to voice usage (e.g., by teachers). Teachers in general, for example, tend to underestimate the seriousness of the condition. Clinical voice alteration analysis, performed by an otorhinolaryngological expert, is an important instrument for early detection. It is based on a visual inspection of the vocal tract by means of a laryngoscopy associated with vocal signal analysis using an appropriate software, like PRAAT [18,19], a software system that takes its name from the imperative form of "praaten" ("to speak" in Dutch), or the Multi-Dimensional Voice Program (MDVP) [20].

These analyses useful for the clinical evaluation of the voice are required by the SIFEL Protocol (Società Italiana di Foniatria e Logopedia) [21], a protocol containing the guidelines for the clinical and instrumental investigations essential for the evaluation of voice disorders proposed by the Italian Society of Phoniatrics and Logopedics.

The SIFEL protocol provides a series of different examinations to evaluate the presence of voice pathologies, such as the anamnestic evaluation to analyze the behavioral characteristics and vocal attitudes of the subject and the familial history of the disorder, the laryngovideostroboscopic examination to detect any physiological or morphological laryngeal alteration, the acoustic analysis, and the subjective self-assessment of the voice.

While the laryngoscopy requires a subjective identification of problems in the larynx and vocal folds, resulting in a qualitative assessment of these structures, the acoustic analysis is used to give an objective quantification of the health of the speech signal through an evaluation of characteristic parameters calculated from a recording of the vowel /a/ of five seconds in length. The acoustic analysis is a useful instrument to evaluate the presence of a possible laryngeal alteration that can cause a deviation in vocal quality. The association between laryngeal functionality and acoustic measures has been shown in several studies in literature [22–24].

The first important parameter calculated for the acoustic analysis is the F_0 [25,26], defined as the frequency of the opening and closing of the glottis. The rate of vibration of the vocal folds is an important index of laryngeal function. Any abnormality of the larynx can alter the speech production system, resulting in a deterioration of voice quality. Unfortunately, this parameter is influenced by several conditions due to physiological and non-physiological factors, including most significantly, the age and the gender of the subject [14–16]. In fact, if, on the one hand, the voice production and vocal health are influenced by people's lifestyle including factors such as smoking, an incorrect diet or an excessive alcohol intake, on the other hand, the main differences are linked to the anatomical differences in laryngeal systems between men and women during their life. Although the pre-puberty laryngeal systems of males and females are quite similar, in adulthood gender differences in laryngeal geometry affect the voice production. In fact, the female vocal folds are shorter and thinner than male ones and these differences contribute to the different F_0 between women and men. Additionally, the higher F_0 in women increases their risk of developing voice disorders because a higher F_0 results in a greater number of vocal fold oscillations and collisions with a resulting higher stress on the thinner vocal folds.

3. Fundamental frequency estimation methodologies: state of the art

As specified in the SIFEL protocol, the F_0 is the first and the most important parameter that has to be evaluated during the acoustic analysis. Moreover, F_0 retrieval is at the basis of the other parameters calculated in the acoustic analysis and most noise estimation methods [27,28]. Such methods evaluate the noise that characterizes the pathological voices caused for example, by an incomplete closure of the vocal folds with corresponding random variations of the speech signal. A critical issue is that there is no standard algorithm to calculate the F_0 of a voice signal. It is important to note that the more accurate the computation of the F_0, and thus its estimation, the more reliable is the voice analysis, otherwise the classification of the voice signal as healthy or pathological. Unfortunately, currently there is no F_0 extraction method that operates consistently for pathological voices [29]. This is due to the more serious and complex irregularities of the vocal fold vibrations in pathological voices compared to healthy ones.

In literature there are several different methods proposed to process the voice signal, like *Spectral Analysis* [30], the *Hilbert-Huang transform* [31–33], the *Robust Algorithm for Pitch Tracking (RAPT)* [34], the *Dynamic Programming Projected Phase-Slope Algorithm (DYPSA)* [35], the *Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram (STRAIGHT)* method [36] and the extraction based on the *Autocorrelation Function* of the speech signal [37]. This latter algorithm is commonly used in a large set of software systems. For example, it is implemented in the PRAAT system [18] and in the MDVP tool [38], two of the standard systems for voice analysis currently on the market.

The *Average Magnitude Difference function (AMDF)* [39] is an alternative to the autocorrelation function as a means to estimate the F_0. In AMDF, the pitch period is calculated from the location of the minimum values of difference function between the input speech signal and its shifted version.

The *Subharmonic-to-Harmonic Ratio Procedure (SHRP)* is another method for the extraction of speech signal features. It is an algorithm that estimates the F_0 in the frequency domain calculating the Subharmonic-to-Harmonic Ratio (SHR) [40], that is the amplitude ratio between sub-harmonics and harmonics using, respectively, a logarithmic frequency scale and a spectrum shifting technique. Instead, the *Sawtooth Waveform Inspired Pitch Estimator (SWIPE)* [41] estimates the pitch as the F_0 of the sawtooth waveform whose spectrum best matches the spectrum of the input signal in the frequency domain.

4. Proposed methodology

The methodology here presented for the F_0 estimation is an optimization and personalization of the *Yin algorithm* [42], and an improvement of exiting one described in [13].

As previously described, in Section 3, one of the most commonly used algorithms is the autocorrelation method. Given a discrete time signal $x(t)$ defined for all t and sampled at sampling frequency F_s (for example 50 kHz, that is the F_s of the signals considered in our experimental phase, as specified in Section 5), the autocorrelation function $r_x(T)$ is generally defined by the following equation:

$$r_x(T) = \lim_{L\to\infty} \frac{1}{2L+1} \sum_{t=-L}^{L} x_t - x_{(t+T)} \tag{1}$$

where L is the section length being analyzed, x_t are the samples of input speech and $x_{(t+T)}$ are the samples time shifted with T delay.

The search for the highest value of the autocorrelation function in the region of interest is used to estimate the F_0 [37,43]. Unfortunately, using the autocorrelation method, the F_0 estimation is affected by errors caused by the noise.

In the proposed methodology, the estimation error is gradually reduced by a set of computational actions, as described below.

Starting from a recording of a speech signal, like the one shown in Fig. 1, the signal is divided into N windows of 10 ms long, as illustrated in Fig. 2, where vertical dotted lines represent the partition of the signal into windows.

Due to physiological and intentional causes, a speech signal is a non-periodic signal. We can consider it as a periodic function with a period T, defining the speech signal x_t (t is the time variable) invariant for a time shift of T, or otherwise:

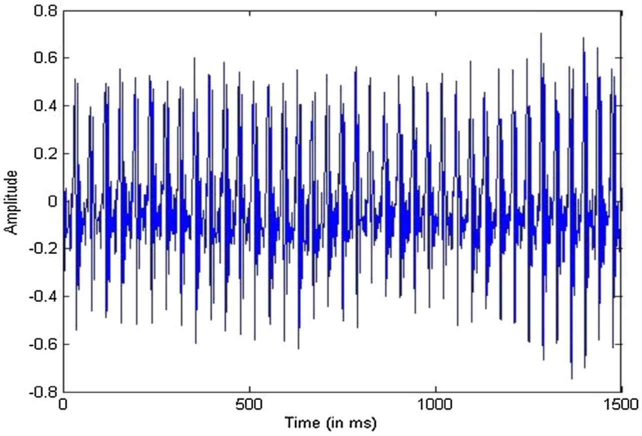


Fig. 1. An example of a speech signal.

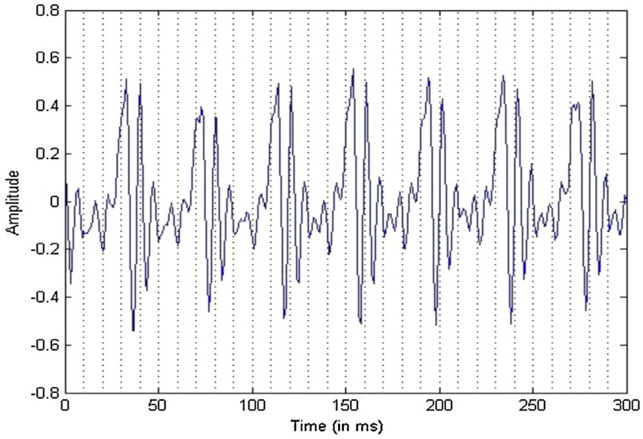


Fig. 2. The speech signal divided in windows.

$$x_t - x_{(t+T)} = 0, \quad \forall t \tag{2}$$

The formula in Eq. (2) is equivalent, if the square is applied and averaged over a window of size W (10 ms), as reported in Eq. (3):

$$\sum_{j=t+1}^{t+W} (x_j - x_{j+T})^2 = 0 \tag{3}$$

The difference function $d_t(\tau)$ is calculated to find the unknown period T, defined as follows:

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2 \tag{4}$$

The speech signal assumes the shape illustrated in Fig. 3 and the τ values for which this difference function $d_t(\tau)$ is minimized, otherwise equal to zero, are searched.

The squared sum defined in Eq. (4) may be expanded and the difference function $d_t(\tau)$ may be expressed in terms of the autocorrelation $r_t(\tau)$, as illustrated in the following equation:

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \tag{5}$$

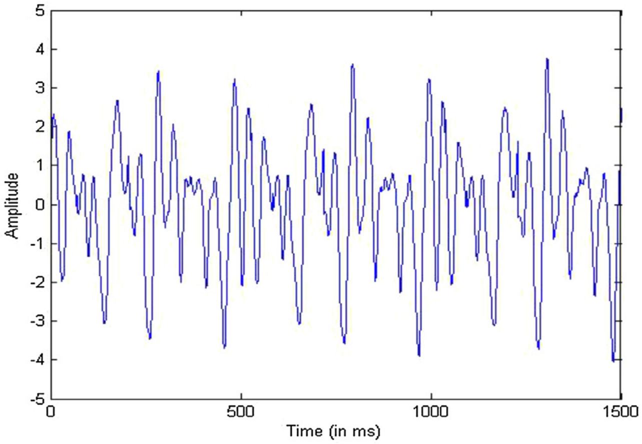


Fig. 3. The speech signal after the difference function.

where the first two terms are defined as energy terms. The autocorrelation function $r_t(\tau)$ of the signal is defined as:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{6}$$

Comparing Eqs. (5) and (6) we can observe that the second term $r_{t+\tau}(0)$ indicated in Eq. (5) varies with τ. This means that if the speech signal is periodic, the energy terms will be equal to zero. Thus, finding the τ values that minimize $d_t(\tau)$ corresponds with selecting the τ values that maximize $r_t(\tau)$.

However, since the speech signal is not a periodic signal, using the difference function $d_t(\tau)$ provides an error rate lower than autocorrelation $r_t(\tau)$, as demonstrated in [42]. For this reason, it is better to select the τ values by minimizing $d_t(\tau)$ instead of using $r_t(\tau)$.

Nevertheless, the difference function is sensitive to amplitude changes of the signal. The Cumulative Mean Normalized Difference Function (CMNDF) is introduced to solve this problem, because this function is more resilient to such amplitude changes. This function is calculated as illustrated in Eq. (7), and the signal changes as shown in Fig. 4.

$$d'_t(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ \dfrac{d_t(\tau)}{\dfrac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)} & \text{otherwise} \end{cases} \tag{7}$$

The τ values that minimize $d'_t(\tau)$, expressed in Eq. (7), and that are smaller than a threshold value, here called β, are taken as T_i.

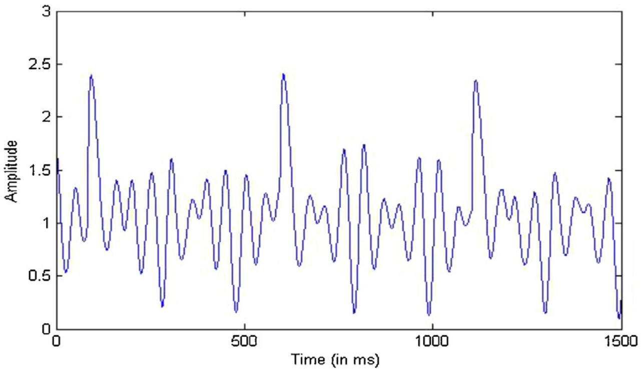


Fig. 4. The speech signal after the cumulative difference function.

Table 1
Datasets for the supervised learning.

Gender	Set	# of voices.	%
Female	Training	51	40.16%
	Validation	38	29.92%
	Testing	38	29.92%
Male	Training	5	38.46%
	Validation	4	30.77%
	Testing	4	30.77%

Otherwise, if the τ values do not satisfy these conditions, T_i is equal to −1.

$$T_i = \begin{cases} \tau & \text{if } \tau < \beta \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

In the methodology presented, the β-value represents the element of innovation. In fact, this threshold is *personalized* for each subject, taking into account both the gender and age of the specific subject, while in the *Yin algorithm* [42] the β-value is fixed and is equal to 0.1. In detail, a unique threshold value has been obtained for each single age, both for females and males. This personalized process is described in detail in Section 4.1.

To increase the accuracy of the F_0 estimation, all the local minima, one for each window, are refined by parabolic interpolation. Next, the F_0 is calculated as the average value of the fundamental frequencies of all windows of the speech signal, considering the inverse of T_i as the F_0 of the ith window, as shown in Eq. (9).

$$F_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i} \tag{9}$$

where N is the number of the windows.

In accordance with the SIFEL protocol, each speech signal on which the analysis is performed is 5 s in length. Therefore, considering windows of a length of 10 ms, N is equal to 500.

4.1. β-Value optimization

The β-value optimization problem has been addressed to personalize the methodology directed at calculating the F_0. The personalization is aimed at reducing the errors in the classification of the voice signal, in addition to maximizing the classification accuracy. The threshold β is fundamental to detect the periods T_i able to minimize the CMNDF in Eq. (8). The periods T_i are used in Eq. (9) to estimate the F_0. This means that personalizing the β-value corresponds to personalizing the F_0 estimation.

To resolve the β-value optimization problem, we developed a supervised learning algorithm that we performed over a database composed of 140 subjects, that is described in Section 5.

To personalize the threshold value for each gender, we considered two different datasets, one composed only of 127 female voices (both healthy and pathological), and the other composed only of 13 male voices (both healthy and pathological).

We randomly organized each database, and then divided each of them into three sets, i.e. a training, testing and validation set, as shown in Table 1.

The training sets were used during the supervised learning phases (one phase for each gender), where a set of β-threshold value ranges were found for each age by carrying out an exhaustive search phase [17], i.e. by investigating all the classification performances on the training sets obtained with all possible β-threshold values from 0.01 to 1, with a step of 0.01.

In this phase, each subject, even if of the same age as others in the database, is treated independently. In Fig. 5 we have reported the intermediate results matrix that contains one line for each subject. Each line, i.e. item *i*, of the matrix is composed of: the age of the subject ($i_{i,1}$), the numerosity *n* of the subjects ($i_{i,2}$) that at this step is always equal to 1 because each subject is treated independently, and the classification result (from $i_{i,3}$ to $i_{i,102}$). This latter corresponds to 0 if, by using the β-threshold value contained in the first line of that column, we obtain a F_0 estimation that is responsible for an incorrect classification of the voice. Otherwise, it corresponds to 1 if the threshold value leads to a F_0 estimation that is responsible for a correct classification of the voice.

For sake of legibility, in Fig. 5 we have reported only a section of the intermediate results matrix. It should be noted that for a specific age, we can find different β-threshold value ranges that allow the correct classification, and that not all ages are covered by the subjects included in the training set. Additionally, a case can also occur in which, for two or more subjects of the same age, the exhaustive search phase has extracted different threshold-ranges (see Fig. 6 for subjects of 20 years old).

To solve these inconsistencies, the algorithm selects the threshold-ranges in common among the subjects of the same age, as shown in Fig. 7, by updating the numerosity *n* of the subjects ($i_{i,2}$).

At the end of the exhaustive search phase, the generalization ability of the best threshold-range is assessed by making use of the validation set.

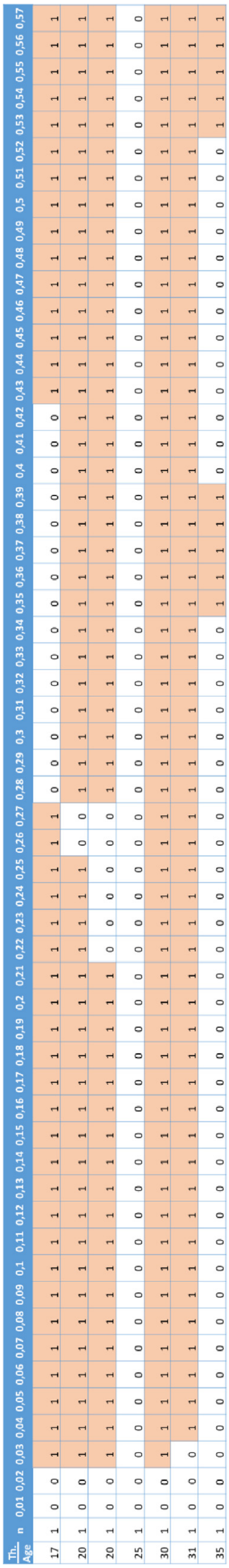
In detail, the algorithm evaluates the classification ability of the threshold-ranges found in the previous phase, and, as shown in Fig. 8, the algorithm selects the sub-set of threshold-ranges able to correctly classify both the subjects of the training set and the subjects included in the validation set.

In the case of subjects with ages not considered in the training set, the classification ability is performed on the threshold-range [0,1], while in the case of several threshold-ranges for subjects of the same age, the algorithm selects as the best threshold-range the first one, that is the threshold-range with the smallest value. For example, if we look at Fig. 8, for subjects of 20 years old the algorithm selected the first threshold-range, from 0.03 to 0.21. Finally, in any case in which there is no matching between the threshold-ranges selected in the previous phase by using the training set and the threshold-ranges selected by using the validation set, the algorithm selects the range of the validation set as the best threshold-range.

Finally, the β threshold value is calculated as the midpoint of the selected best threshold-range for each age. For any ages not covered by either testing or the validation set, the β threshold value is obtained by using the spline interpolation method [44].

At the end of the optimization problem analysis, the β threshold value found is tested over the items of the testing set. It should be noted that these database subjects were never considered in the two previous steps performed to solve the optimization problem. Consequently, the use of this third set allows an estimation of how well this β threshold value can perform over previously unseen voices.

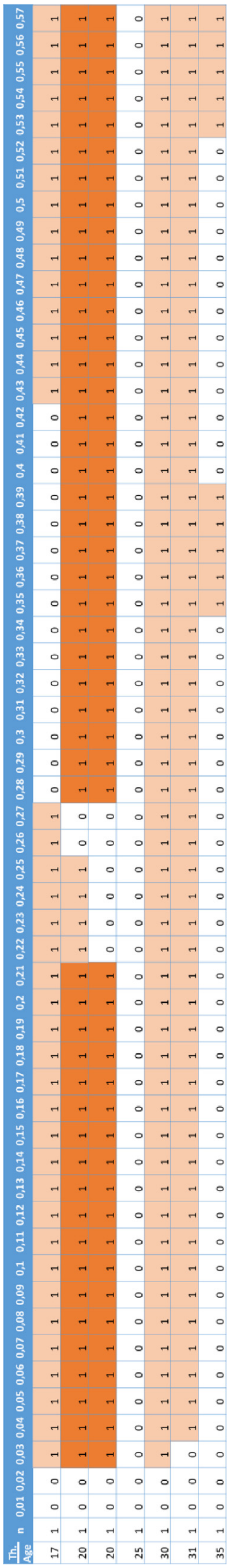
All classification results obtained by using the personalized β threshold values on the whole dataset are reported in Table 3.



Th. Age	n	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19	0,2	0,21	0,22	0,23	0,24	0,25	0,26	0,27	0,28	0,29
17	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
20	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
20	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Th. Age	n	0,3	0,31	0,32	0,33	0,34	0,35	0,36	0,37	0,38	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5	0,51	0,52	0,53	0,54	0,55	0,56	0,57
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Fig. 5. Exhaustive search phase – Intermediate results matrix. This matrix contains one item for each subject. In the first column the age of the subject is reported, and the second column the numerosity *n* of the subjects that at this step is always equal to 1 because in this phase each subject, even if of the same age as others in the database, is treated independently. Finally, in the other columns, the classification results obtained for each subject by using the β-threshold value indicated in the corresponding column (first line highlighted in blue) are reported: 0 if by using the β-threshold value we obtain a F_0 estimation that is responsible for an incorrect classification of the voice; 1 if the threshold value leads to a F_0 estimation that is responsible for a correct classification of the voice. Finally the cases for each age that allow the correct classification are highlighted in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)



Th. Age	n	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19
17	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Th. Age	n	0,2	0,21	0,22	0,23	0,24	0,25	0,26	0,27	0,28	0,29	0,3	0,31	0,32	0,33	0,34	0,35	0,36	0,37	0,38
17	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

Th. Age	n	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5	0,51	0,52	0,53	0,54	0,55	0,56	0,57
17	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Fig. 6. Exhaustive search phase – Intermediate results matrix. This matrix contains one item for each subject. In the first column the age of the subject is reported, and the second column the numerosity n of the subjects that at this step is always equal to 1 because in this phase each subject, even if of the same age as others in the database, is treated independently. Finally, in the other columns, the classification results obtained for each subject by using the β-threshold value indicated in the corresponding column (first line highlighted in blue) are reported: 0 if by using the β-threshold value we obtain a F_0 estimation that is responsible for an incorrect classification of the voice; 1 if the threshold value leads to a F_0 estimation that is responsible for a correct classification of the voice. It should be noted that for a specific age, we can find different threshold-ranges that allow the correct classification. A case can also occur in which, for two or more subjects of the same age, the exhaustive search phase has extracted different threshold-ranges (see subjects of 20 years old). To solve these inconsistencies, the algorithm selects the threshold-ranges in common among the subjects of the same age. These cases are highlighted in dark orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

5. Testing phase

The β-value optimization problem and the experimental phase were carried out using the available on-line database, the "Saarbrucken Voice Database" (SVD) [45].

The SVD is a collection of voice recordings from more than 2000 people with different pathologies, created by the Institute of Phonetics of Saarland University. It contains recordings of sustained /a/, /i/ and /u/ vowels. All the recordings are sampled at 50 kHz and their resolution is 16-bit, recorded in a mono-channel WAV format to preserve the fidelity of the signal.

Starting from the SVD database, we built a new dataset composed of both healthy and pathological voices. It includes 140 recordings containing the sustained phonation of the vowel sound /a/. The use of the vowel /a/ in speech analysis is required by the SIFEL protocol [21]. It has also been demonstrated that there is a strong correlation between the ElectroGlottoGraphic (EGG) parameters and the acoustic features of the vowel /a/ [46].

The dataset consists of 73 healthy voices and 67 voices of subjects suffering from Reinke's edema, a disease of interest for the medical partner due to its high contemporary incidence. The partition of the considered voices is schematized as follows:

$$
\text{140 Voices}\begin{cases}
\textit{healthy} & \text{73 voices}\begin{cases}\textit{female} & \text{66 voices}\\ \textit{male} & \text{7 voices}\end{cases}\\[2em]
\textit{pathological} & \text{67 voices}\begin{cases}\textit{female} & \text{61 voices}\\ \textit{male} & \text{6 voices}\end{cases}
\end{cases}
$$

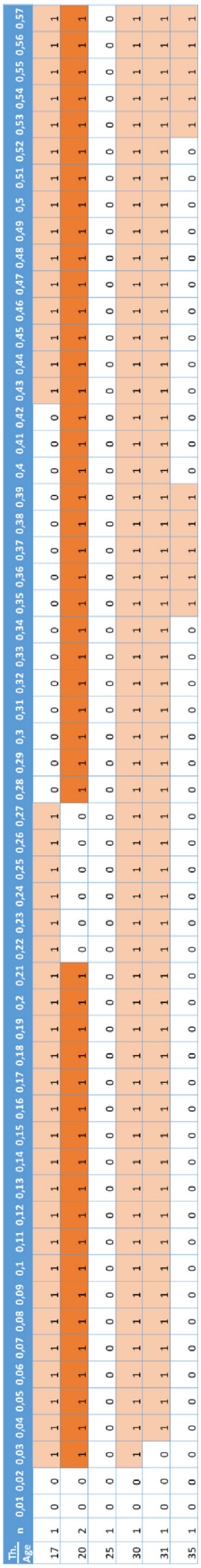
More details of voice signals used in this study are indicated in Table 2, in which we have reported the number (No) of selected voices for each age range and gender, the percentage calculated for each group and for the complete dataset.

The low number of male samples is related to the higher incidence of Reinke's edema in female subjects than males, as indicated in several studies reported in [8,9,47]. For this reason, the dataset contains more female voices than male ones. It is important to remember that we have used all the available voices suffering from Reinke's edema from the Saarbruken Voice Database, that currently represents the only available database on line of voice disorders. Unfortunately, this database contains only 6 male voices suffering from Reinke's edema.

5.1. Numerical results

To validate and estimate the goodness of the proposed methodology, we compared the realized algorithm, in terms of classification performance, with the most commonly used and cited fundamental frequency estimation algorithms, as reported in [48].

In detail, we computed the F_0 values by using, in addition to our proposed methodology, AMDF [39], SWIPE [41], SHRP [40], Yin [42], RAPT [34], STRAIGHT [36] and PRAAT [18]. In our tests, we used PRAAT, freely available, instead of MDVP. There are few differences between the implementations of the two systems and, in particular, as reported in [20,49], the F_0 measurements for the two



Th./Age	n	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19
17	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	2	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Th./Age	0,2	0,21	0,22	0,23	0,24	0,25	0,26	0,27	0,28	0,29	0,3	0,31	0,32	0,33	0,34	0,35	0,36	0,37	0,38
17	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
20	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

Th./Age	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5	0,51	0,52	0,53	0,54	0,55	0,56	0,57
17	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Fig. 7. Exhaustive search phase – Intermediate results matrix. This matrix contains one item for each subject. In the first column the age of the subject is reported, and the second column the numerosity n of the subjects that at this step is always equal to 1 because in this phase each subject, even if of the same age as others in the database, is treated independently. Finally, in the other columns, the classification results obtained for each subject by using the β-threshold value indicated in the corresponding column (first line highlighted in blue) are reported: 0 if by using the β-threshold value we obtain a F_0 estimation that is responsible for an incorrect classification of the voice; 1 if the threshold value leads to a F_0 estimation that is responsible for a correct classification of the voice. The thresholds-ranges selected by using the training set are highlighted in orange and dark-orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 2
Details of the voice signals used in this study.

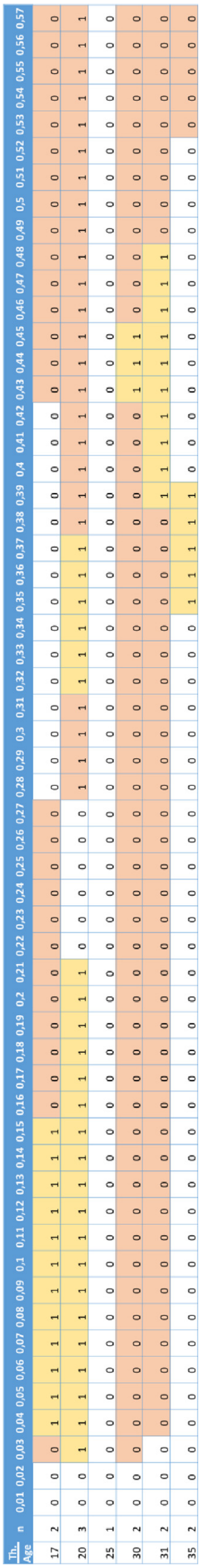
Category	Gender	Age group	No.	% for each group	% on complete dataset
Healthy	Female	17–29	20	30.30%	14.29%
		30–39	10	15.15%	7.14%
		40–49	10	15.15%	7.14%
		50–59	13	19.70%	9.29%
		60+	13	19.70%	9.29%
Healthy	Male	17–29	3	42.86%	2.14%
		30–39	1	14.29%	0.71%
		40–49	1	14.29%	0.71%
		50–59	–	–	–
		60+	2	28.57%	1.43%
Pathological	Female	17–29	–	–	–
		30–39	7	11.48%	5.00%
		40–49	23	37.70%	16.43%
		50–59	13	21.31%	9.29%
		60+	18	29.51%	12.86%
Pathological	Male	17–29	–	–	–
		30–39	–	–	–
		40–49	–	–	–
		50–59	2	33.33%	1.43%
		60+	4	66.67%	2.86%
Total	***Female***	17–60+	*127*	*90.71*	
	Male	17–60+	*13*	*9.29*	

programs are similar, calculated using an algorithm based on the autocorrelation method.

To classify if a voice is pathological, otherwise if it is affected by Reinke's edema, each F_0 was computed by using one of the algorithms listed above, and then the F_0 value was compared with a fixed normal (or healthy) range of values.

As previously mentioned, the F_0 value is influenced by several factors, such as the gender, age, lifestyle and, consequently, defining a healthy range of values of F_0 to discriminate between pathological and healthy voices is very difficult. Currently, in the scientific literature there is no standard range of variability against which the F_0 can be evaluated [23]. However, there are several statistical studies aimed at illustrating the typical values of speech F_0 in healthy people, both for men and women, as demonstrated in [50]. Of course, the mean values change slightly with age. For men, a typical F_0 value is 124, and any decrease in this is most dramatic during puberty continuing with successive deceleration until about 35 years of age until at about 55 years of age, it begins to rise again [51,52]. For women, a typical F_0 value is 211 and it is stationary up to the age of menopause, when it decreases to reach a minimum that is about 15 Hz lower at around 70 years of age [52,53].

Based on these studies and thanks to the support of the group of specialists of the School of Otorhinolaryngology of the "University Magna Graecia" of Catanzaro (Italy), who are also involved in the project, we have selected as normal or healthy the range 189–260 Hz for female voices and 104–158 Hz for male voices. This range was calculated by considering the mean value of the F_0 and the average F_0 variation (standard deviation) according to the main studies of clinical interest existing in literature [50,54–56] that report results for adult male and female subjects.



Th. Age	n	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19
17	2	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
20	3	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Th. Age	n	0,2	0,21	0,22	0,23	0,24	0,25	0,26	0,27	0,28	0,29	0,3	0,31	0,32	0,33	0,34	0,35	0,36	0,37	0,38
17	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	3	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

Th. Age	n	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5	0,51	0,52	0,53	0,54	0,55	0,56	0,57
17	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
31	2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
35	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 8. Validation phase results – this matrix contains one item for each subject of the Validation data-set. This matrix contains one item for each subject. In the first column the age of the subject is reported, and the second column the numerosity *n* of the subjects that at this step is always equal to 1 because in this phase each subject, even if of the same age as others in the database, is treated independently. Finally, in the other columns, the classification results obtained for each subject by using the β-threshold value indicated in the corresponding column (first line highlighted in blue) are reported: 0 if by using the β-threshold value we obtain a F_0 estimation that is responsible for an incorrect classification of the voice; 1 if the threshold value leads to a F_0 estimation that is responsible for a correct classification of the voice. The best threshold-ranges for each age that allow the correct classification for the subjects included both in the training set and validation set are highlighted in yellow; the threshold-ranges selected in the exhaustive search phase are highlighted in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Considering these ranges normal or healthy, we have applied the following IF/THEN rules to classify a voice signal as healthy or pathological.

$$\text{Female voice is classified as} \begin{cases} \textit{healthy} & \text{if } (189 \le F_0 \le 280) \\ \textit{pathological} & \text{otherwise} \end{cases}$$

$$\text{Male voice is classified as} \begin{cases} \textit{healthy} & \text{if } (104 \le F_0 \le 158) \\ \textit{pathological} & \text{otherwise} \end{cases}$$

The results were classified according to the following definitions:

- True positive (TP): the algorithm detected a pathology when a pathology was present;
- True negative (TN): the algorithm detected a healthy voice when a healthy voice was present;
- False positive (FP): the algorithm detected a pathology when a healthy voice was present; and
- False negative (FN): the algorithm detected a healthy voice when a pathology was present.

The classification performances were measured in terms of:

- Sensitivity: TP/(TP + FN);
- Specificity: TN/(TN + FP);
- Positive Predictive Value (PPV): TP/(TP + FP);
- Negative Predictive Value (NPV): TN/(TN + FN);
- Accuracy: (TP + TN)/(TP + TN + FP + FN).

Table 3 shows the results in terms of the accuracy, specificity, sensitivity, and positive and negative predictive percentages. In Table 3, the best value achieved is reported in bold for each parameter. The results in this table show that the percentage of correct classifications provided by the proposed methodology is the highest for all the considered parameters.

This means that our algorithm provides a good estimation of the F_0, enabling users to distinguish between healthy and pathological voices. The table shows sensitivity and specificity values, respectively, of over 72% and 81%, which indicates that the proposed methodology has a lower number of false negatives and false positives (i.e. pathological voices erroneously classified as non-pathological or healthy voices erroneously classified as pathological) in comparison with the other algorithms.

Due to the good classification performances of the proposed methodology, we have embedded it into an easy and portable tool, an app, able to empower people with information about dysphonia and its related causes, and to provide a reliable real-time screening of their voice. Currently, we have realized a prototypal app released to operate on Android-based Operating System (OS) devices [13]. The app is able to acquire the user's vocal signal by using the microphone of the mobile device, to elaborate this signal (the sound of the vowel /a/ voiced for 5 s) by calculating in real time the F_0 using the proposed methodology, and to classify the voice as healthy or pathological. In preliminary tests, the algorithm performance was satisfactory. The execution time of the algorithm is about 50–70 ms,

Table 3
Average performance of the compared classifiers.

Algorithms	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
Proposed method	**71.64**	**80.82**	**77.42**	**75.64**	**76.43**
PRAAT [18]	68.66	72.60	69.70	71.62	70.71
AMDF [39]	64.18	72.60	68.25	68.83	68.57
SWIPE [41]	59.70	75.34	68.97	67.50	67.86
SHRP [40]	61.19	73.97	68.33	67.50	67.86
YIN [42]	70.15	73.97	71.21	72.97	72.97
RAPT [34]	68.66	73.97	70.77	72.00	71.43
STRAIGHT [36]	61.19	73.97	68.33	67.50	67.86

Table 4
The ANOVA table for accuracy.

Source	SS	df	MS	F	Prob > F
Columns	0.63888	7	0.09127	590	3.93783e−309
Error	0.12252	792	0.00015		
Total	0.7614	799			

Table 5
The ANOVA table for sensitivity.

Source	SS	df	MS	F	Prob > F
Columns	1.52433	7	0.21776	219.52	1.12011e−180
Error	0.78565	792	0.00099		
Total	2.30998	799			

Table 6
The ANOVA table for specificity.

Source	SS	df	MS	F	Prob > F
Columns	0.48324	7	0.06903	80.77	2.36188e−88
Error	0.67691	792	0.00085		
Total	1.16015	799			

the length of time depending on the data processor embedded in the mobile device, while the app occupies little memory space on the device (about 16 MB), each stored file of each vocalization being about 70 KB.

5.2. The statistical analysis

We carried out a statistical analysis to evaluate whether or not a statistically significant difference exists between the accuracy, sensitivity and specificity. We used ANOVA, a well-known collection of statistical models based on the analysis of variance.

We carried out a one-way ANOVA, because there is one measurement variable, the F_0, and one nominal variable.

The null hypothesis is that there is no significant difference in the classification among the considered algorithms. In other words, the H_0 is that the F_0 values calculated by using each algorithm are the same for the same subject, and the alternative hypothesis is that the F_0 values are not the same.

All statistical analysis steps were performed by using MATLAB.

We chose a level of significance, α, equal to 0.05, which represents the maximum admissible probability of incorrectly rejecting a given null hypothesis, assessing the statistical equivalence when this is true. When the p-value, the probability (Prob) computed by the test, is greater than or equal to 0.05 the null hypothesis is true, and there is no statistically significant difference among the algorithms. Otherwise, if the p-value is lower than this α value, the null hypothesis is rejected, which means that a statistically significant difference among the algorithms exists.

The results of the one-way ANOVA for accuracy, sensitivity and specificity are reported in Tables 4–6, respectively. Please note that SS is the Sum of Squares, df indicates the degrees of freedom, MS is the mean square, that is calculated by dividing the Sum of Squares by its degrees of freedom, and the F value is the ratio of the mean squares.

The outcome of the ANOVA tests is that the resulting p-values are very low and therefore there is a statistically significant difference. At this point, we performed an exhaustive set of unpaired t-tests with a level of significance $\alpha = 0.05$ between each pair of the six algorithms. We report the results in Tables 7–9.

For each pair of algorithms (x, y) we report two pieces of information: the p-value obtained by performing the t-test between algorithms x and y, together with a letter to indicate if there is a significant statistical difference (Y) or not (N). We performed only one test for each pair of algorithms, with no comparison performed between an algorithm x and itself.

The tables confirm that there is a significant difference between the proposed methodology and the other algorithms.

6. Conclusions and future work

In this paper a new personalized methodology for voice screening has been proposed. This represents the first and the most important analysis performed to evaluate the patient's vocal state of health. It is useful to identify possible variations in the signal, indices of probable voice disorders, allowing the calculation of several other parameters provided by the SIFEL protocol, using objective data capable of characterizing quali-quantitatively a specific vocal dysfunction.

The proposed methodology gives a very good estimation of the fundamental frequency of the voice signal, with the highest classification performance. The results achieved during the experimental phases have demonstrated that our methodology has the best accuracy in comparison with other F_0 estimation algorithms.

A good estimation of the F_0 is the starting point to estimate other parameters provided by the acoustic analysis for an appropriate evaluation of the patient's vocal state of health. It is useful, for example, for the analysis of the pitch variability through the evaluation of characteristic parameters such as jitter, shimmer, and the Harmonic to Noise Ratio (HNR). All these additional parameters will be aims of our future work.

Currently, this methodology has been applied only to a specific disease, the Reinke's edema, due to its high incidence, a fact which has increasing the interest of the medical partner involved in the study. Indeed, it is the third most common cause of dysphonia in the world. However, the methodology could be extended also to other diseases. In this regard, we plan to perform in the future new experiments in order to test the methodology on a larger number of subjects, and to evaluate the application of this methodology to other diseases. To achieve this aim, we are involved in the design of an experimental phase, in which, in collaboration with the medical staff involved in the project, we will conduct a pilot study to collect a larger number of subjects recruited at different ages. This experimental phase will give us the possibility of investigating the generalizability of the results obtained in this preliminary study, across different age groups, but also the possibility of creating a

Table 7
The *t*-test table for accuracy.

	Prop. Meth.	PRAAT [18]	AMDF [39]	SWIPE [41]	SHRP [40]	YIN [42]	RAPT [34]	STRAIGHT [36]
Prop. Meth.	–	7.81E−84 Y	1.20E−99 Y	3.94E−110 Y	1.23E−110 Y	1.48E−59 Y	9.81E−74 Y	1.9E−109 Y
PRAAT	–	–	9.34E−27 Y	8.89E−42 Y	2.72E−42 Y	1.04E−14 Y	1.49E−05 Y	2.50E−41 Y
AMDF	–	–	–	0.0003 Y	0.0002 Y	9.50E−47 Y	3.49E−38 Y	0.0003 Y
SWIPE	–	–	–	–	0.95 N	2.19E−60 Y	4.23E−53 Y	0.95 N
SHRP	–	–	–	–	–	6.97E−61 Y	1.22E−53 Y	1 N
YIN	–	–	–	–	–	–	4.25E−05 Y	7.16E−60 Y
RAPT	–	–	–	–	–	–	–	1.46E−52 Y
STRAIGHT	–	–	–	–	–	–	–	–

Table 8
The *t*-test table for sensitivity.

	Prop. Meth.	PRAAT [18]	AMDF [39]	SWIPE [41]	SHRP [40]	YIN [42]	RAPT [34]	STRAIGHT [36]
Prop. Meth.	–	4.97E−10 Y	6.55E−40 Y	6.07E−67 Y	1.91E−59 Y	0.001 Y	3.59E−10 Y	1.33E−58 Y
PRAAT	–	–	1.63E−19 Y	6.23E−49 Y	4.53E−40 Y	0.001 Y	0.98 N	1.93E−39 Y
AMDF	–	–	–	9.31E−20 Y	7.13E−11 Y	3.35E−30 Y	5.92E−20 Y	1.01E−10 Y
SWIPE	–	–	–	–	0.0007 Y	5.34E−59 Y	1.07E−49 Y	0.001 Y
SHRP	–	–	–	–	–	8.82E−51 Y	8.72E−41 Y	0.96 N
YIN	–	–	–	–	–	–	0.0009 Y	5.26E−50 Y
RAPT	–	–	–	–	–	–	–	3.94E−40 Y
STRAIGHT	–	–	–	–	–	–	–	–

Table 9
The *t*-test table for specificity.

	Prop. Meth.	PRAAT [18]	AMDF [39]	SWIPE [41]	SHRP [40]	YIN [42]	RAPT [34]	STRAIGHT [36]
Prop. Meth.	–	1.68E−48 Y	6.97E−49 Y	3.12E−28 Y	1.15E−37 Y	1.15E−37 Y	8.87E−39 Y	1.15E−37 Y
PRAAT	–	–	0.90 N	3.15E−10 Y	0.001 Y	0.001 Y	0.001 Y	0.001 Y
AMDF	–	–	–	1.55E−10 Y	0.0007 Y	0.0007 Y	0.0008 Y	0.0007 Y
SWIPE	–	–	–	–	0.001 Y	0.001 Y	0.0009 Y	0.001 Y
SHRP	–	–	–	–	–	1 N	0.922 N	1 N
YIN	–	–	–	–	–	–	0.922 N	1 N
RAPT	–	–	–	–	–	–	–	0.922 N
STRAIGHT	–	–	–	–	–	–	–	–

new and more appropriate voice database. For these reasons, the presented results should be considered as preliminary.

Finally, we recall that there are numerous other conditions that may cause vocal disorders. Factors such as stress and the general health of the user may greatly affect the structures and quality of the vocal mechanism. These other conditions are not currently taken into consideration, so the classification provided by the methodology should not be considered as a diagnosis. However, in the case of a probable presence of dysphonia detected by the methodology, it is advisable for the user to undergo a proper and complete vocal examination performed by a qualified otorhinolaryngologist or speech-language pathologist.

Acknowledgments

The authors would like to acknowledge the project “Smart Health 2.0” – PON04A2_C for its support of this work. Additionally, the authors wish to thank Prof. Nicola Lombardo, of the Department of Otolaryngology-Head and Neck Surgery of the University “Magna Graecia” of Catanzaro (Italy) involved in the SmartHealth 2.0 project, for his invaluable contribution in the identification of the healthy range of values of F_0 used in this study.

References

[1] N. Roy, R.M. Merrill, S.D. Gray, E.M. Smith, Voice disorders in the general population: prevalence, risk factors, and occupational, Laryngoscope 115 (11) (2005) 1988–1995.

[2] L. de Araújo Pernambuco, A. Espelt, P.M.M. Balata, K.C. de Lima, Prevalence of voice disorders in the elderly: a systematic review of population-based studies, Eur. Arch. Oto-Rhino-Laryngol. 272 (10) (2015) 2601–2609.

[3] A. Remacle, C. Petitfils, C. Finck, D. Morsomme, Description of patients consulting the voice clinic regarding gender, age, occupational status, and diagnosis, Eur. Arch. Oto-Rhino-Laryngol. 274 (3) (2017) 1567–1576.

[4] R.H.G. Martins, H.A. do Amaral, E.L.M. Tavares, M.G. Martins, T.M. Gonçalves, N.H. Dias, Voice disorders: etiology and diagnosis, J. Voice 30 (6) (2016), 761-e1.

[5] L.C.C. Cutiva, I. Vogel, A. Burdorf, Voice disorders in teachers and their associations with work-related factors: a systematic review, J. Commun. Disord. 46 (2) (2013) 143–155.

[6] E. Van Houtte, S. Claeys, F. Wuyts, K. Van Lierde, The impact of voice disorders among teachers: vocal complaints, treatment-seeking behavior, knowledge of vocal care, and voice-related absenteeism, J. Voice 25 (5) (2011) 570–575.

[7] H. Byeon, Exploring potential risk factors for benign vocal fold mucosal disorders using weighted logistic regression, Int. J. Biosci. Biotechnol. 6 (4) (2014) 77–86.

[8] M. Dajer, F. Andrade, A. Montagnoli, J. Pereira, D. Tsuji, Vocal dynamic visual pattern for voice characterization, in: Journal of Physics: Conference Series, IOP Publishing, 2011, p. 012026.

[9] D. Marcotullio, G. Magliulo, T. Pezone, Reinke's edema and risk factors: clinical and histopathologic aspects, Am. J. Otolaryngol. 23 (2) (2002) 81–84.

[10] A. Assunção, I. Bassi, A. de Medeiros, C. de Souza Rodrigues, A. Gama, Occupational and individual risk factors for dysphonia in teachers, Occup. Med. 62 (7) (2012) 553–559.

[11] N. Zidar, N. Gale, A. Cardesa, L. Ortega, Larynx and hypopharynx, in: Pathology of the Head and Neck, Springer, 2016, pp. 333–386.

[12] S.R. Schwartz, S.M. Cohen, S.H. Dailey, R.M. Rosenfeld, E.S. Deutsch, M.B. Gillespie, E. Granieri, E.R. Hapner, C.E. Kimball, H.J. Krouse, J.S. McMurray, S. Medina, K. O'Brien, D.R. Ouellette, B.J. Messinger-Rapport, R.J. Stachler, S. Strode, D.M. Thompson, J.C. Stemple, J.P. Willging, T. Cowley, S. McCoy, P.G. Bernad, M.M. Patel, Clinical practice guideline: hoarseness (dysphonia), Otolaryngol.-Head Neck Surg. 141 (3) (2009) S1–S31.

[13] L. Verde, G. De Pietro, P. Veltri, G. Sannino, An m-health system for the estimation of voice disorders, in: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2015, pp. 1–6.

[14] E.J. Hunter, K. Tanner, M.E. Smith, Gender differences affecting vocal health of women in vocally demanding careers, Logoped. Phoniatr. Vocol. 36 (3) (2011) 128–136.

[15] E.T. Stathopoulos, J.E. Huber, J.E. Sussman, Changes in acoustic characteristics of the voice across the life span: measures from individuals 4–93 years of age, J. Speech Lang. Hear. Res. 54 (4) (2011) 1011–1021.

[16] M. Brockmann, M.J. Drinnan, C. Storck, P.N. Carding, Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task, J. Voice 25 (1) (2011) 44–53.

[17] A.K. Jain, R.P. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[18] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Proceedings of the Institute of Phonetic Sciences, vol. 17, Amsterdam (1993) 97–110.

[19] P. Boersma, V. Van Heuven, Speak and Unspeak With PRAAT, Glot International, 2001, pp. 341–347.

[20] O. Amir, M. Wolf, N. Amir, A clinical comparison between MDVP and PRAAT softwares: is there a difference? MAVEBA (2007) 37–40.

[21] A. Ricci Maccarini, E. Lucchini, La valutazione soggettiva ed oggettiva della disfonia: il protocollo sifel, Presented at the Relazione ufficiale al XXXVI Congresso Nazionale della Societa' Italiana di Foniatria e Logopedia (2002).

[22] L.W. Lopes, L.B. Simões, J.D. da Silva, D. da Silva Evangelista, A.C.d.N. E Ugulino, P.O.C. Silva, V.J.D. Vieira, Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses, J. Voice 31 (3) (2017), 382-e15.

[23] B. Barsties, M. De Bodt, Assessment of voice quality: current state-of-the-art, Auris Nasus Larynx 42 (3) (2015) 183–188.

[24] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, V. Saferis, Categorizing normal and pathological voices: automated and perceptual categorization, J. Voice 25 (6) (2011) 700–708.

[25] B. Yegnanarayana, K. Murty, Event-based instantaneous fundamental frequency estimation from speech signals, IEEE Trans. Audio Speech Lang. Process. 17 (4) (2009).

[26] W. De Colle, O. Schindler, Voce e computer: analisi acustica digitale del segnale verbale (il sistema CSL-MDVP), Omega, 2001.

[27] C. Manfredi, M. D'Aniello, P. Bruscaglioni, Acoustic measure of noise energy in vocal folds operated patients, in: 9th European Signal Processing Conference (EUSIPCO 1998), IEEE, 1998, pp. 1–4.

[28] B. Boyanov, S. Hadjitodorov, Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases, IEEE Eng. Med. Biol. Mag. 16 (4) (1997) 74–82.

[29] P. Mitev, S. Hadjitodorov, Fundamental frequency estimation of voice of patients with laryngeal disorders, Inf. Sci. 156 (1) (2003) 3–19.

[30] 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 25 (5) (2011) e223–e232.

[31] B. Jiao, Y. Zeng, Y. Mao, Pitch detection method based on Hilbert-Huang transform for speech signal, Comput. Eng. Appl. 1 (2015) 041.

[32] G. Schlotthauer, M.E. Torres, H.L. Rufiner, Pathological voice analysis and classification based on empirical mode decomposition, in: Development of Multimodal Interfaces: Active Listening and Synchrony, Springer, 2010, pp. 364–381.

[33] M.E. Torres, G. Schlotthauer, H. Rufiner, M. Jackson-Menaldi, Empirical mode decomposition. Spectral properties in normal and pathological voices, in: 4th European Conference of the International Federation for Medical and Biological Engineering, Springer, 2009, pp. 252–255.

[34] D. Talkin, A robust algorithm for pitch tracking (RAPT), in: Speech Coding and Synthesis, 1995.

[35] P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the dypsa algorithm, IEEE Trans. Audio Speech Lang. Process. 15 (1) (2007) 34–43.

[36] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, T. Irino, Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT, Interspeech (2005) 537–540.

[37] L. Tan, M. Karnjanadecha, Pitch detection algorithm: autocorrelation method and AMDF, Proceedings of the 3rd International Symposium on Communications and Information Technology, vol. 2 (2003) 551–556.

[38] M.M. Baki, G. Wood, M. Alston, P. Ratcliffe, G. Sandhu, J.S. Rubin, M.A. Birchall, Comparison between OperaVOX^{TM} and MDVP: preliminary results, Otolaryngol. Head Neck Surg. 149 (2 Suppl) (2013) P203–P204.

[39] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, H.J. Manley, Average magnitude difference function pitch extractor, IEEE Trans. Acoust. Speech Signal Process. 22 (5) (1974).

[40] X. Sun, Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, IEEE, 2002, pp. I-333.

[41] A. Camacho, Swipe: a sawtooth waveform inspired pitch estimator for speech and music (Ph.D. thesis), University of Florida, 2007.

[42] A. De Cheveigné, H. Kawahara, Yin, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111 (4) (2002) 1917–1930.

[43] L. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process. 25 (1) (1977) 24–33.

[44] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, C. De Boor, A Practical Guide to Splines, vol. 27, Springer-Verlag, New York, 1978.

[45] D. Martínez, E. Lleida, A. Ortega, A. Miguel, J. Villalba, Voice pathology detection on the Saarbruecken voice database with calibration and fusion of scores using multifocal toolkit, in: Advances in Speech and Language Technologies for Iberian Languages, Springer, 2012, pp. 99–109.

[46] A.A. Dibazar, T.W. Berger, S.S. Narayanan, Pathological voice assessment, in: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. EMBS'06, IEEE, 2006, pp. 1669–1673.

[47] L.W. Lopes, L.B. Simões, J.D. da Silva, D. da Silva Evangelista, A.C.d.N. E Ugulino, P.O.C. Silva, V.J.D. Vieira, Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses, J. Voice (2017).

[48] A. Tsanas, M. Zañartu, M.A. Little, P.E. McSharry, Robust fundamental frequency estimation in sustained vowels using ensembles? IEEE Trans. Audio Speech Lang. Process. (2011).

[49] O. Haldun, M.A. Kiliç, M.A. Şafak, Comparison of results in two acoustic analysis programs: PRAAT and MDVP, Turk. J. Med. Sci. 41 (5) (2011) 835–841.

[50] H. Traunmüller, A. Eriksson, The frequency range of the voice fundamental in the speech of male and female adults (manuscript), Department of Linguistics, University of Stockholm, 1993.

[51] H. Hollien, T. Shipp, Speaking fundamental frequency and chronologic age in males, J. Speech Lang. Hear. Res. 15 (1) (1972) 155–159.

[52] M. Pegoraro Krook, Speaking fundamental frequency characteristics of normal swedish subjects obtained by glottal frequency analysis, Folia Phoniatr. Logopaed. 40 (2) (1988) 82–90.

[53] M.L. Stoicheff, Speaking fundamental frequency characteristics of nonsmoking female adults, J. Speech Lang. Hear. Res. 24 (3) (1981) 437–441.

[54] A. Palumbo, B. Calabrese, P. Vizza, N. Lombardo, A. Garozzo, M. Cannataro, F. Amato, P. Veltri, A novel portable device for laryngeal pathologies analysis and classification, in: Advances in Biomedical Sensing, Measurements, Instrumentation and Systems, Springer, 2010, pp. 335–352.

[55] G. Williamson, Human Communication: A Linguistic Introduction, 2nd ed., Speech-Language Services, Billingham, 2006.

[56] M.P. Gelfer, V.A. Mikos, The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels, J. Voice 19 (4) (2005).