applied sciences

MDPI

Article

A Transparent Decision Support Tool in Screening for Laryngeal Disorders Using Voice and Query Data

Jonas Minelga [1], Antanas Verikas [1,2,*], Evaldas Vaiciukynas [1], Adas Gelzinis [1] and Marija Bacauskiene [1]

1 Department of Electric Power Systems, Kaunas University of Technology, Studentu 50, LT-51368 Kaunas, Lithuania; jonasmin@gmail.com (J.M.); evaldas.vaiciukynas@ktu.lt (E.V.); adas.gelzinis@ktu.lt (A.G.); marija.bacauskiene@ktu.lt (M.B.)

2 Centre for Applied Intelligent Systems Research, Halmstad University, Kristian IV:s väg 3, P.O. Box 823, S-30118 Halmstad, Sweden

* Correspondence: antanas.verikas@hh.se; Tel.: +46-72-977-3515

Received: 4 September 2017; Accepted: 20 October 2017; Published: 24 October 2017

Abstract: The aim of this study is a transparent tool for analysis of voice (sustained phonation */a/*) and query data capable of providing support in screening for laryngeal disorders. In this work, screening is concerned with identification of potentially pathological cases by classifying subject's data into *'healthy'* and *'pathological'* classes as well as visual exploration of data and automatic decisions. A set of association rules and a decision tree, techniques lending themselves for exploration, were generated for pathology detection. Data pairwise similarities, estimated in a novel way, were mapped onto a 2D metric space for visual inspection and analysis. Accurate identification of pathological cases was observed on unseen subjects using the most discriminative query parameter and six audio parameters routinely used by otolaryngologists in a clinical practice: equal error rate (EER) of 11.1% was achieved using association rules and 10.2% using the decision tree. The EER was further reduced to 9.5% by combining results from these two classifiers. The developed solution can be a useful tool for Otolaryngology departments in diagnostics, education and exploratory tasks.

Keywords: decision tree; *t*-SNE visualization; association rules; pathological voice

1. Introduction

Laryngeal disorders are relatively common, affecting ∼5% according to [1] or 6.2% according to [2] of the general population, and are encountered in varying degrees of severity. Worldwide, larynx-related cancer causes about 200,000 annual deaths. This number keeps growing, while decreasing numbers of deaths related to many other types of cancer are observed. Therefore, efforts targeting preventive laryngeal health-care are required.

Information obtained from both invasive and non-invasive measurements is used in laryngology for diagnostics and monitoring of treatment outcomes. Indirect laryngoscopy and video laryngostroboscopy are probably the most informative modalities in diagnostics for laryngeal disorders [3]. However, sophisticated tools and an invasive course of actions required for obtaining such data impede wide scale monitoring of human larynx using such instrumentation.

On the other hand, voice signals are obtained non-invasively and computer-based analysis of voice data is used increasingly in monitoring of treatment outcomes and screening for laryngeal disorders [4–11]. Several measures computed from voice data are already widely used to quantify dysphonia changes and characterise outcomes of therapeutic and surgical treatment of laryngeal diseases [8–11]. It was demonstrated that even a voice signal, transmitted through a telephone channel, contains much information for the task of laryngeal pathology detection [12].

Questionnaire data, routinely collected by laryngologist, is another source of easily and non-invasively obtained information. As shown in [13], query data-based detection of disordered larynx consistently outperforms detection based on voice data, and fusion of these modalities improves the performance further. It was demonstrated that query data can provide useful information for pathology detection, which is not present in voice data [14]. In [13], large sets of voice and query features were used in a random forest [15] aiming to achieve low classification error in a two-class (healthy/pathological) classification task. By contrast, in this work, our focus is on small feature sets, transparent simple classification techniques and data exploration based on the suggested novel way to assess the similarity of observations.

Various voice analysis tools, such as LingWaves (version 3.0, WEVOSYS medical technology GmbH, Baunach, Germany), Computerized Speech Lab (PENTAX Medical, Tokyo, Japan), and Dr. Speech (version 4, Tiger DRS, Seattle, Washington, USA), are usually used by laryngologists in clinical practice. Dr. Speech is very popular due to its low price and good documentation, but it is not really clear how to exploit full potential of analysis results available from the software [16]. To the best of our knowledge, there is no tool capable of exploiting both voice and query data in screening for laryngeal disorders. Some aspects of using query data in screening for laryngeal pathologies were explored from the medical viewpoint in [17,18].

The aim of this study is to create a user-friendly tool for identification of potentially pathological cases based on transparent analysis techniques and capable of exploiting both voice and query data. Currently, the tool is oriented towards experts working at departments of otolaryngology, but, in the near future, the developed tool could run on a smart phone, including voice recording and analysis, and be much more versatile.

As a voice modality, sustained phonation of vowel /*a*/ has been used in this study, since sustained phonation is simple, lessens variance in sustained vowels and increases reliability of computed acoustic features [11,19]. In addition, sustained phonation is rather unaffected by aspects related to different languages. Research concerning suitability of different vowels [20,21] often concludes that vowel /*a*/ results in the lowest EER in laryngeal pathology detection.

2. Voice and Query Data

A detailed description of voice and query data used in this study can be found in [17,18]. It is worth mentioning that, when labelling the data (healthy/pathological), the initial diagnosis was based on the visual appearance of larynx from the video laryngostroboscopy and direct microlaryngoscopy. The final diagnosis was confirmed by histological examination of laryngeal specimens taken during endolaryngeal microsurgical intervention. Three laryngologists were involved in the labelling process. Low inter-rater variability and high reliability was observed. In the following two subsections, we present a short description of the data, in order to facilitate understanding the content of this article.

2.1. Voice Data

Voice recordings were made using an acoustic cardioid microphone AKG Perception 220 (AKG Acoustics, Vienna, Austria), having a frequency range from 20 Hz to 20 kHz. A sound-proof booth was used as a place for recordings where the microphone was placed at a 10-cm distance from the subject's mouth (all participating subjects were seated with a headrest) with about 90º microphone-to-mouth angle. The audio format used was wav (dual-channel PCM, 16 bit samples at 44 kHz rate), resulting in the Nyquist frequency F_{max} = 22 kHz. The same database as in [22] was used in this study.

The "Voice" database contains 273 subjects (163 normal and 110 pathological voices), varying in sex and ranging from 19 to 85 years in age. The normal voice group consisted of 51 men and 112 women while there were 42 men and 68 women in the group of pathological voices. Thus, in total, 180 women and 93 men participated in the study. Subjects in the *normal* voice group had no history of chronic voice disorders, had no complaints regarding their voice and considered it as normal. The voice of these subjects was assessed by laryngologists as normal. Video laryngostroboscopy did not reveal any

pathological alterations in the larynx. The *pathological* voice group included patients with mass lesions of vocal folds and unilateral vocal fold paralysis.

2.2. Query Data

The questionnaire statements identified by Bacauskiene et al. [23] were used in this study. In [23], these statements were deemed as being the most important for solving the voice pathology detection tasks. The "Query" database contained 596 subjects (106 healthy males, 221 healthy females, 118 pathological males, and 151 pathological females) ranging from 17 to 86 years in age. All subjects from the "Voice" database are present in the "Query" database. The questionnaire data were obtained from subjects' responses to the set of questions, presented in Table 1. A set of 26 questions may seem too big for daily clinical work. However, it was shown that this set of 26 questions can be reduced to nine without substantial decrease in pathology detection accuracy: the equal error rate (EER) increased from 6.18 to 7.72% when using nine questions instead of 26 [13].

Table 1. Query items.

#	Question Content	Units (or Scale) of Measurement
1	Subject's gender	{Male, Female}
2	Subject's age	discrete number
3	Average duration of intensive speech use	hours/day
4	Average duration of intensive speech use	days/week
5	Smoking	{Yes, No}
6	Smoking intensity	cigarettes/day
7	Smoking history	years
8	Maximum phonation time	seconds
9	SSA of voice function quality	visual analogue scale from 0 to 100
10	SSA of voice hoarseness	from 0 (*no*) to 100 (*severe hoarseness*)
11	Voice handicap progressing	grade from 1 to 4
12	SSA of daily experienced stress level	from 0 (*no*) to 100 (*very much stress*)
13	Frequency of singing	grade from 1 to 5
14	Frequency of talking/singing in a smoke-filled room	grade from 1 to 5
15	SSA of experienced discomfort due to voice disorder	from 0 (*no*) to 100 (*huge discomfort*)
16	SSA of "too weak voice"	from 0 (*no*) to 100 (*very clear*)
17	SSA of repetitive "loss of voice"	from 0 (*no*) to 100 (*very clear*)
18	SSA of reduced voice	from 0 (*no*) to 100 (*very distinctly*)
19	SSA of reduced ability to sing	from 0 (*no*) to 100 (*very distinctly*)
20	Frequency of voice cracks or aberrant voice	from 0 (*no*) to 100 (*very often*)
21	Level of vocal usage	level from 1 to 4
22	Speaking took extra effort (G_1)	from 0 (*no*) to 5 (*severe problem*)
23	Throat discomfort or pain after voice usage (G_2)	from 0 (*no*) to 5 (*severe problem*)
24	Voice weakens while talking, vocal fatigue (G_3)	from 0 (*no*) to 5 (*severe problem*)
25	Voice cracks or sounds different (G_4)	from 0 (*no*) to 5 (*severe problem*)
26	GFI [24] = $G_1 + G_2 + G_3 + G_4$	grade from 0 to 20

* SSA stands for subjective self-assessment and GFI for glottal function index.

As a quantization step in rule extraction through affinity analysis (see Section 3.1), answers to questions with a scale possessing more than seven unique values were transformed into quartiles. Zero values were considered as a separate category and did not undergo the transformation. Answers to questions # 22–25 (G_1–G_4) were replaced with a new response G_0, which indicates that the subject responded to at least one of the questions by a *no problem* answer:

$$G_0 = (G_1 = 0 \vee G_2 = 0 \vee G_3 = 0 \vee G_4 = 0). \quad (1)$$

If there were no answers to questions # 22–25 as "*no problem*" ($G_1 > 0$ and $G_2 > 0$ and $G_3 > 0$ and $G_4 > 0$), then the value of G_0 was set to 0.

3. Methodology

In this work, the pathology detection task is solved through classification of voice and query data into two classes, *healthy* and *pathological*. Since the tools are to be used by laryngologists in clinical practice, data classification and exploration algorithms applied should lend themselves to provide insights into automatic decisions and numerical data (be transparent).

The same applies to parameters characterizing subject's voice, sustained phonation of vowel */a/* in this case. Hundreds and thousands of parameters can be extracted for characterization of a subject's voice [22]. While the use of huge sets of parameters and advanced algorithms indicates how good detection performance can be achieved. Such tools are usually attributed to the "black box" category, since they provide very little insights into the data they analyse and the decisions they take. Therefore, to characterize a subject's voice, we used just six audio parameters widely adopted by laryngologists all over the world: fundamental frequency (F0), jitter, shimmer, normalized noise energy (NNE), harmonics-to-noise ratio (HNR), and signal-to-noise ratio (SNR). F0 measures how low or high the frequency of subject's voice is; jitter assesses the periodicity of vocal fold vibrations; shimmer reflects the cycle-to-cycle variability in amplitude of the voice; NNE quantifies the relative level of vocal noise (compared to the harmonics); HNR quantifies the amount of noise in the voice signal; and SNR compares the voice signal total energy to the energy of the aperiodic component of the signal. There are several variants of the parameters based on signal-analysis techniques applied to compute them. In this work, the utterance-based approach was used to compute the audio parameters. Information on how these parameters can be computed can be found in [12,25]. A set (of a similar size) of other voice parameters could probably provide higher classification accuracy. However, the aforementioned six parameters are already widely used to quantify dysphonia changes and characterise outcomes of therapeutic and surgical treatment of laryngeal diseases.

Two transparent techniques were developed for pathology detection: *association rules* for analysis of query data and a decision tree for exploring voice parameters. Decision trees are popular classification tools in various medical applications. Decision trees have also been used for voice pathology detection too [26,27]. The main novelty of this work is the suggested way of combined use of association rules and decision trees for voice pathology detection based voice and query data. A novel approach to data exploration is also suggested.

The C4.5 decision tree [28], applied in this work, was created using the six audio parameters augmented with the glottal function index (GFI) parameter from the query data. The GFI parameter was added due to its high frequency of participation in various association rules [13]. According to Bach et al. [24], the GFI is a reliable and reproducible symptom index, which is easy to administrate and exhibits good validity. The index is useful in the assessment of patients with glottal dysfunction. The Matlab (version R2012b, MathWorks, Natick, Massachusetts, USA) environment was used to create the decision tree. One can expect obtaining similar results using the Weka tool [29], trees.J48, which is a clone of the C4.5 decision tree learner.

Association rules, identifying the strongest co-occurrences of answers, were mined using an unsupervised Apriori algorithm [30]. A simple confidence-based classifier was derived from association rules for detection of laryngeal disorders. Both association rules and a decision tree are transparent algorithms, based on parameters routinely used by otolaryngologists, and can greatly contribute to better understanding of various conditions and treatment planning.

Ability to explore data and automatic decisions is of great value for an end user. We used the *t*-distributed stochastic neighbor embedding (*t*-SNE) algorithm for visualization of multidimensional data in this work. For deeper insights, the user is also provided with a decision path in the decision tree and class-conditional distributions of parameters used for pathology detection.

3.1. Association Rules

Association rules were obtained through affinity analysis identifying links between observations and/or between variables in the explored data set. The result of this analysis is a set of rules

of the `if <antecedent> then <consequent>` type, where *antecedent* (A) is a response or a set of co-occurring responses concerning questionnaire statements, while *consequent* (C) is the subject's diagnosis. The following measures were used, to assess importance or interestingness of the rule:

$$support(A \rightarrow C) = P(A \wedge C), \tag{2}$$

$$confidence(A \rightarrow C) = \frac{P(A \wedge C)}{P(A)}, \tag{3}$$

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{P(C)}, \tag{4}$$

where $P(condition)$ is the probability of the condition (fraction of observations having the condition), $\wedge$ is logical AND, *support* is the popularity of the rule (fraction of observations having both A and C), *confidence* describes the strength or purity of the rule (how often having A leads to C), and *lift* is a measure of surprise (the increased likelihood of C being found together with A).

Some studies [31,32] demonstrated how association rules can be used to construct a classifier. Here, we apply an extracted rule set for classification of query data using the weighted majority approach. A normalized certainty of rules triggered by a subject's responses is computed first:

$$Certainty = 100 \cdot \frac{\sum_{i=1}^{J} p_i - \sum_{i=1}^{K} h_i}{\sum_{i=1}^{J} p_i + \sum_{i=1}^{K} h_i}, \tag{5}$$

where J is the number of 'pathological' rules triggered, K is the number of 'healthy' rules triggered, p_i is the confidence of the triggered ith 'pathological' rule, and h_i is the confidence of the triggered ith 'healthy' rule.

Sign of the resulting certainty value determines the diagnosis: a positive value means pathological and a negative value means a healthy case.

3.2. Evaluation of Detection

Ten-fold cross validation was used to assess performance of the tool. To evaluate the goodness of detection, certainties were used as scores. For voice data-based detection, a score for the observation $\mathbf{x}$ is given by the probability of the dominant class at a terminal node the $\mathbf{x}$ is assigned to. For query data-based detection, the certainty obtained using Equation (5) was used as score.

We evaluated performance of a classifier using the following measures: (a) the detection error trade-off (DET) curve and the EER; (b) the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The DET, EER, ROC, and AUC measures were computed using an interpolated version of the ROC through a pool adjacent violators algorithm, namely the ROC convex hull method (see the BOSARIS toolkit [33]).

Various thresholds can be used to convert a soft decision into a hard one and the overall performance of a detector can be conveniently represented by the ROC or DET curve. Due to the logarithmic scale, the DET curves enable comparison of several models at a glance easier than the ROC curves [34]. Nevertheless, measurements of AUC from ROC can be valuable for various comparisons.

A convenient way to compare the accuracy of models with different DET (ROC) curves is the equilibrium point, usually known as EER. EER is the point where the DET (ROC) curve intersects the diagonal and: (a) miss rate (false positive rate) = false alarm rate (false negative rate), for DET; or (b) sensitivity (true positive rate) = specificity (true negative rate), for ROC.

3.3. Exploring Analysis Results

3.3.1. Data Visualization by the *t*-SNE Algorithm

Many algorithms have been developed for mapping data from a multidimensional space to a low-dimensional space. The Classical Multidimensional Scaling [35] or Principal Coordinates Analysis (PCoA) is a popular technique for performing such tasks. PCoA is a linear technique computing a series of eigenvalues and eigenvectors.

In this work, we are interested in nonlinear mapping techniques possessing greater flexibility than linear ones. Amongst numerous nonlinear mapping techniques [36], we selected the *t*-SNE algorithm [37]. The *t*-SNE algorithm has shown an excellent performance on both real world and artificial data [37]. Ability to embed a new observation onto a beforehand generated map is an appealing property of the algorithm.

The *t*-SNE represents similarities (proximities) of observations as conditional probabilities. The proximity of $\mathbf{x}_j$ to $\mathbf{x}_i$ in a high-dimensional space (in our case this space would be given by the six audio parameters augmented with the GFI parameter from the query data) is defined as the conditional probability $p_{j|i}$ that $\mathbf{x}_j$ is chosen by $\mathbf{x}_i$ as its neighbour when neighbours are chosen in proportion to their probability density defined by a Gaussian centered on $\mathbf{x}_i$ [37]:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma^2)}, \tag{6}$$

where $p_{i|i}$ are set to zero and σ is a parameter selected experimentally, the width of the Gaussian centered on an observation. In dense regions, smaller values of σ are more appropriate.

A similar conditional probability $q_{j|i}$ is calculated in the low-dimensional space: $\mathbf{y}_i$ and $\mathbf{y}_j$, counterparts of $\mathbf{x}_i$ and $\mathbf{x}_j$. In this work, the low-dimensional space is given by the first and the second *t*-SNE coordinates, a two-dimensional space used for data visualization. The joint probability $q_{ji} = q_{ij}$ is defined as:

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}, \tag{7}$$

where $q_{i|i}$ are set to zero. The mapping sought is obtained by minimizing the Kullback–Leibler divergence between the joint probability distributions P and Q.

In Equations (6) and (7), a distance is computed to assess similarity between observations. However, in multidimensional spaces, the distance-based similarity of observations suffers from a too big influence of irrelevant noisy variables. Therefore, we use a decision tree and association rules designed for data classification, to assess the similarity of observations in this work.

3.3.2. Assessing the Similarity of Observations

When using a decision tree, assessment of similarity between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is based on measuring "distance" between two leaf nodes of the decision tree occupied by the observations. We suggest assessing the similarity of observations $\mathbf{x}_i$ and $\mathbf{x}_j$ by the following equation:

$$p_{ij}^t = 1/(e^{w \cdot g_{ij}}), \tag{8}$$

where w is a parameter, and g_{ij} is the number of tree branches between the two leaf nodes occupied by $\mathbf{x}_i$ and $\mathbf{x}_j$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ occupy the same leaf node, then $g = 0$ and $p_{ij} = 1$. The parameter w governs the influence of the "distance" (the number of tree branches) between two leaf nodes occupied by the observations on the similarity values.

Association rules-based similarity between observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed according to Equation (9):

$$p_{ij}^r = \frac{2\sum_{n\in\mathcal{U}\cap\mathcal{V}} p_{in}}{\sum_{n=1}^{N_\mathcal{U}} p_{in} + \sum_{n=1}^{N_\mathcal{V}} p_{jn}}, \tag{9}$$

where $\mathcal{U}$ and $\mathcal{V}$ are the sets of rules activated by the observations $\mathbf{x}_i$ and $\mathbf{x}_j$, correspondingly, $N_\mathcal{U}$ and $N_\mathcal{V}$ are the number of rules in those sets, and p_{in} is the certainty of the nth rule activated by $\mathbf{x}_i$.

3.3.3. Visualizing Data Distributions and Decisions

Distributions of the six audio parameters and all query parameters are created and visualised for healthy and pathological cases as separate probability density functions (PDFs) using the Epanechnikov kernel smoothing, as recommended in [38]. Figure 1 presents examples of PDFs for the six audio parameters. Probability density functions allow obtaining very useful preliminary understanding based on parameters otolaryngology specialists are familiar with. These distributions evolve over time with inclusion of new subjects (patients and controls), which provides doctors with information on trends and allows comparing different groups of patients.

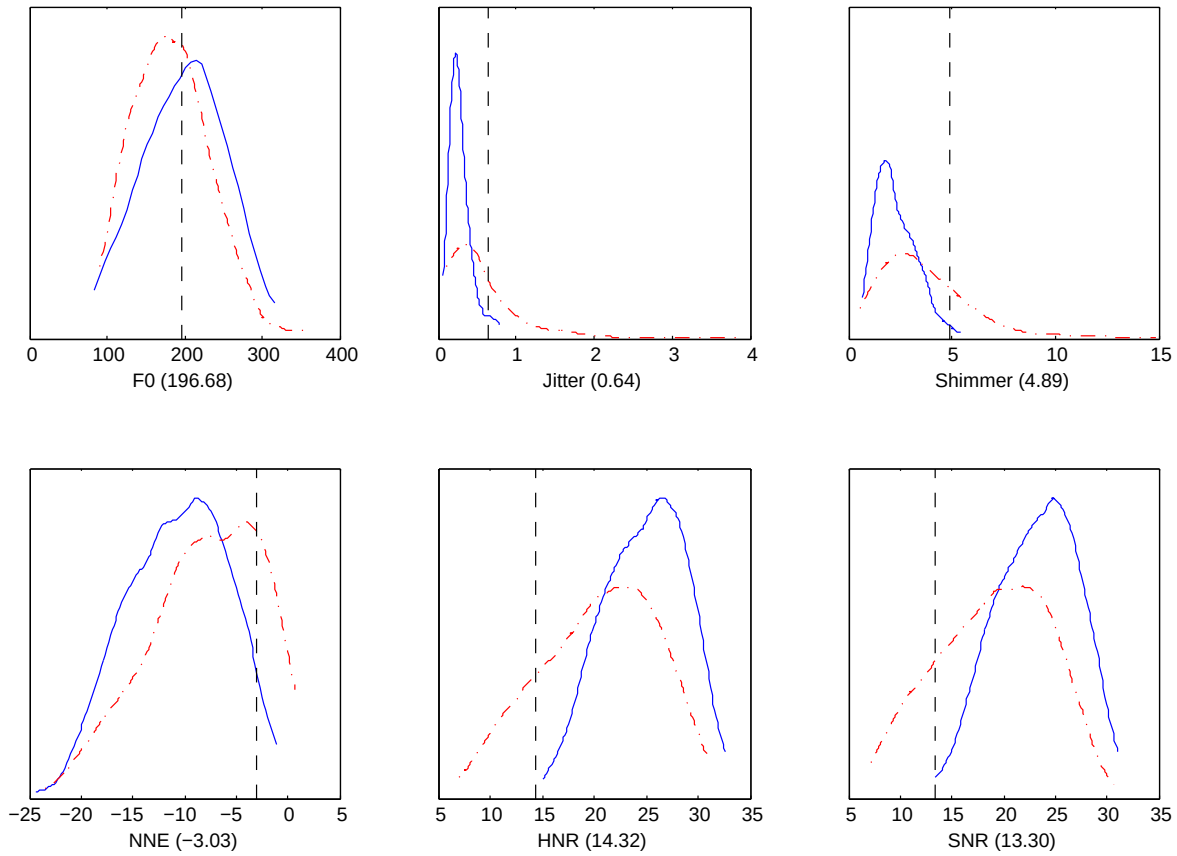

Figure 1. Probability (y-axis) density functions of audio parameters of *healthy* (blue) and *pathological* (red, dashed-dotted) cases. A dashed line and numbers in the parentheses correspond to a parameter value of a subject randomly selected from the database used to train the developed tools.

To facilitate reasoning, a decision tree is also provided for the user. A fragment of such a tree is exemplified in Figure 2. The decision path (highlighted in Figure 2) helps with exploring the patient diagnosis more thoroughly.

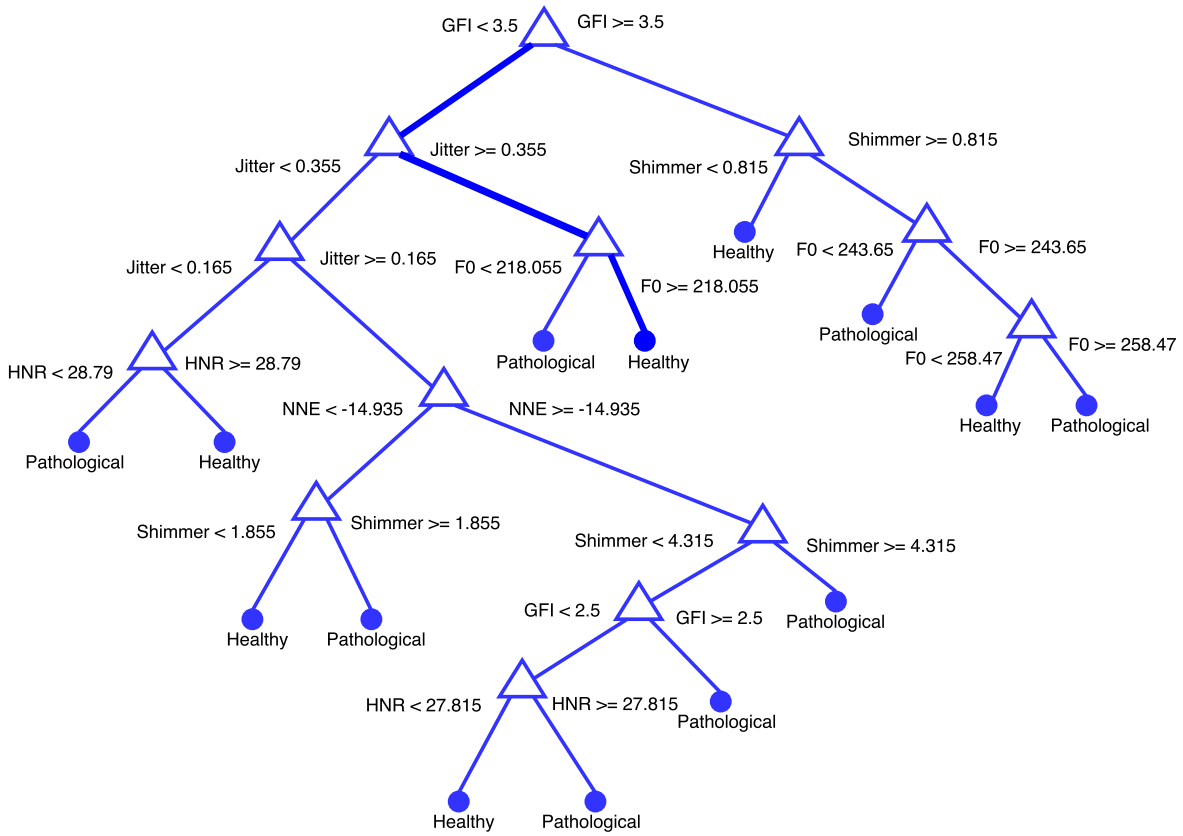

Figure 2. A fragment of the decision tree created using the six audio parameters and the GFI (Glottal function index) parameter.

4. Experiments and Results

The pathology detection techniques were trained and validated using voice data from [22] and questionnaire data from [13]. Values of all parameters used in the algorithms were determined experimentally by the cross-validation using these data sets. The techniques were also tested using data collected from unseen subjects during routine clinical work after the designing process was completed. The graphical user interface was created using Matlab. The following issues were studied:

1. Pathology detection accuracy using data collected during routine clinical work from unseen subjects.
2. Effectiveness of the association rules extracted from the query data and resulting insights gained from the rules.
3. Exploration of audio and query data, and automatic decisions using the developed visualization and analysis techniques.

4.1. Pathology Detection Accuracy

Apart from the ten-fold cross validation tests carried out in the designing phase of the tool, the goodness of detection was also assessed using voice and query data from 45 unseen subjects (nine normal and 36 pathological). The normal group consisted of four men and five woman while there were 13 men and 23 women in the group of pathological voices. This set of data was collected after the designing process was completed. Figures 3 and 4 summarize results of these additional tests, where DET and ROC curves as well as EER and AUC values are presented. The EER points on the DET curves are given by the intersection points of the curves with the diagonal shown by the dashed line (see Figure 3).

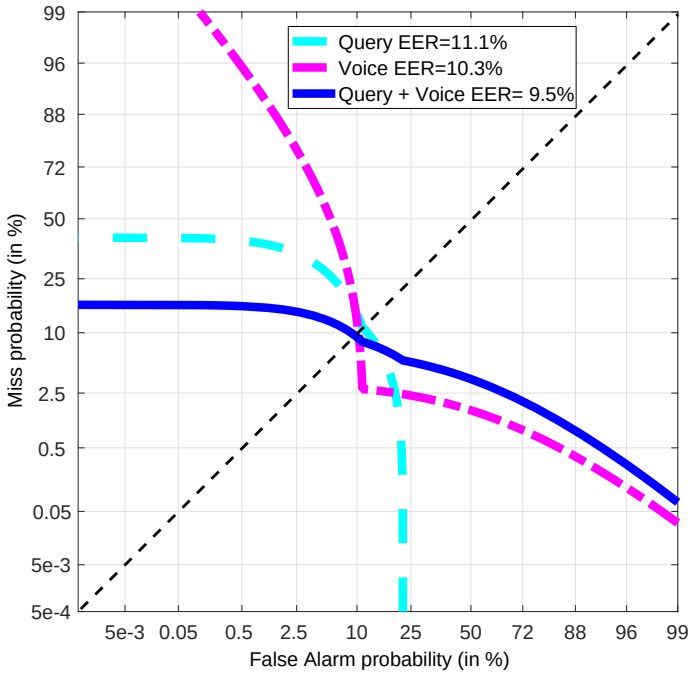

Figure 3. DET (Detection error trade-off) curves and EER (Equal error rate) for unseen voice and query data.

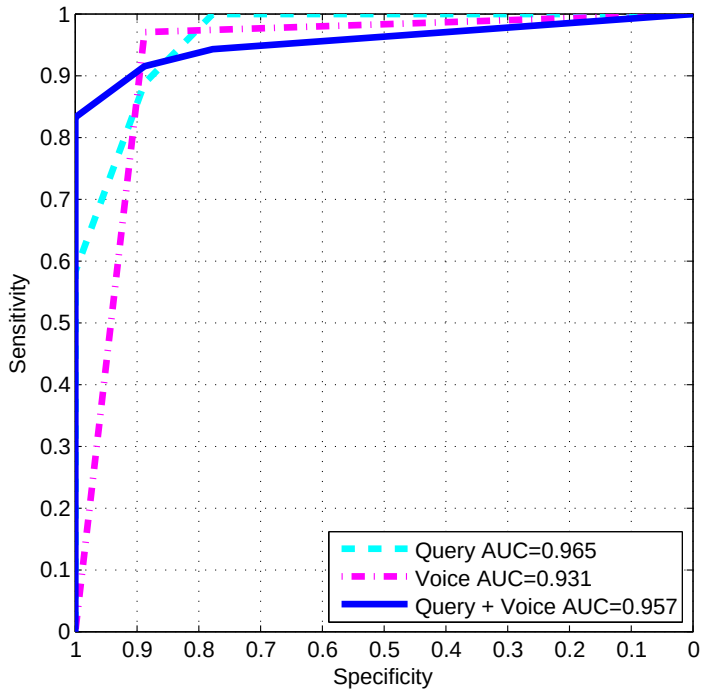

Figure 4. ROC (Receiver operating characteristic) curves and AUC (Area under the curve) for unseen voice and query data.

The obtained EER values are very encouraging bearing in mind the simplicity of the techniques: EER of 11.1% using association rules, 10.3% using the decision tree and 9.5% when combining (via weighted averaging) results obtained from both the rules and the tree. The combination weights were set to be proportional to the detection accuracy obtained in the designing process based on the ten-fold cross validation. As expected, a combination of the two data modalities improves pathology

detection accuracy. As can be seen from the DET and ROC curves, the combined classifier is not only the most accurate classifier around the EER operating point, but also shows the lowest false alarm probability (or highest specificity) near the low miss probability (or high sensitivity) mode of operation, which can be considered as an appealing property when screening for disorders in preventive health-care.

The ten-fold cross validation performed on the data sets described in Sections 2.1 and 2.2 resulted in EER of 17.3% for the voice data, 12.7% for the query data and 11.8% when combining results obtained from the two modalities via weighted averaging. The higher error rates obtained for these much larger data sets are expected due to larger diversity of the data. A larger fraction of mild laryngeal pathology cases were present in these larger data sets compared to the set of data collected after the designing process was completed. Cases of mild laryngeal pathology are usually more difficult to distinguish from the normal (healthy) ones compared to the cases of more severe pathology.

4.2. Association Rules

Generated association rules were filtered to retain only rules having diagnosis in the consequent part and: $support > 0.28$ and $confidence > 0.9$ for a *healthy* subject; and $support > 0.16$ and $confidence > 0.9$ for a *pathological* subject. These threshold values were selected experimentally. The gravity of each response, or absolute frequency of each retained antecedent component belonging to the pathological or healthy rules, is provided in Table 2. In Table 2, G_0 is given by Equation (1), MPT stands for maximum phonation time, VAS for visual analog scale, and all of the other abbreviations are short names of the corresponding questions summarised in Table 1. The 'healthy rules' turned out to be longer, but redundant, where the most persistent components showed the highest gravity. The 'pathological rules' were found to be less complex and their components were slightly more diverse.

Table 2. Rule coverage for mined antecedent items. Absolute frequency (gravity), as the number of rules the specific response participates in, is in "Healthy" and "Pathological" columns.

Question #	Name & Value	Healthy	Pathological
22–25	$G_0 = 0$	11	0
11	H = 2	11	0
6	C = 0	6	1
7	Y = 0	6	0
4	U = 7	3	2
8	MPT = [2,12]	0	4
21	L = 3	4	0
9	VAS = (63,100]	0	2
22–25	$G_0 = 0$	0	1
20	X = (73,100]	0	1
19	S = (75.8,100]	0	1
18	R = (60,100]	0	1
16	W = (60,100]	0	1
15	D = (65,100]	0	1
1	Gender = F	1	0

4.3. Exploring Data and Decisions

Data similarity values, computed using Equations (8) and (9), were averaged and collected into a similarity/proximity matrix. The proximity matrix was then mapped onto the 2D space using the *t*-SNE algorithm. The perplexity parameter of the *t*-SNE algorithm was chosen empirically and set to 50. The algorithm converged after 320 iterations. The resulting *t*-SNE visualization is shown in Figure 5, where 'triangles' stand for healthy and 'squares' for pathological cases. Filled circles denote two selected cases, assigned to the 'healthy' class by the classifier, used for tracking purposes. The light-filled circle stands for one subject and the dark-filled one for another subject.

A 2D map created by the *t*-SNE algorithm facilitates selection of specific cases for deeper studies and comparison. The map allows spotting misclassified, incorrectly labelled observations or observations mapped very closely, but coming from different classes and requiring deeper analysis. The determined class label and classification certainty obtained from association rules and decision tree are linked to each observation. The observation labelled by the light-filled circle in Figure 5 illustrates one such case. The observation was assigned to the 'healthy' class by the classifier, but was mapped in the area mainly occupied by 'pathological' cases. Such observations are candidates for deeper exploration and the map lends itself for such analysis. By clicking on a selected observation on the 2D map, the user accesses information linked to the selected individual: the six voice parameters (F0, Jitter, Shimmer, NNE, HNR and SNR) and questionnaire data. The observation denoted by the dark-filled circle in Figure 5 represents a correctly classified 'healthy' subject.

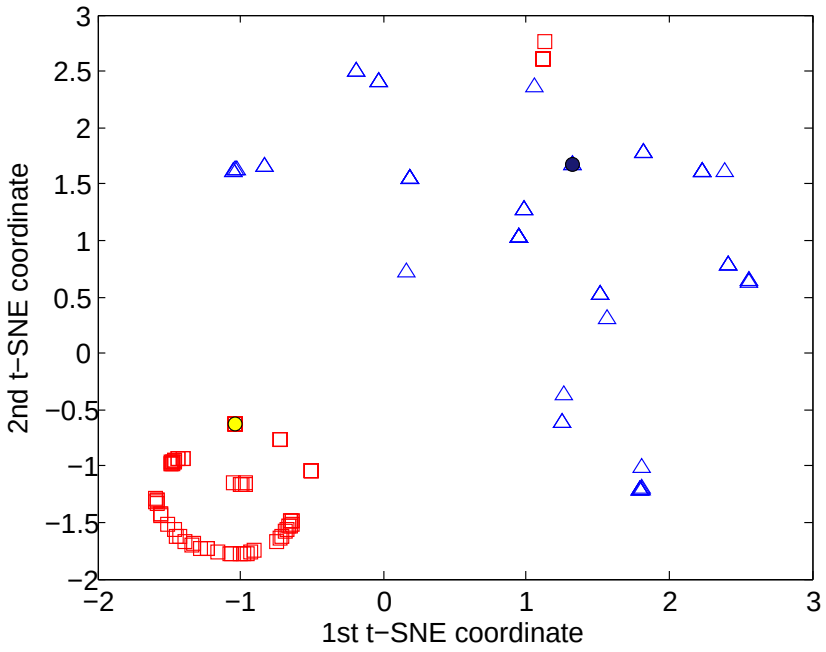

Figure 5. Visualization of observations, represented by proximity values, by the *t*-SNE algorithm.

The class-conditional probability density functions of audio and query parameters are of great help in such kind of analysis. Figure 6 presents examples of class-conditional probability density functions, estimated using the Epanechnikov kernel smoothing, of query parameters. Density functions of the age parameter, see the 1st row 2nd column in Figure 6, illustrate deficiency in data collection pointing out the lack of healthy subjects in a certain age group. The class-conditional density functions also show how a big difference can be expected between values of a certain parameter computed using subjects from the *healthy* and *pathological* classes. More than one mode in the distribution hints at the existence of clusters in the data.

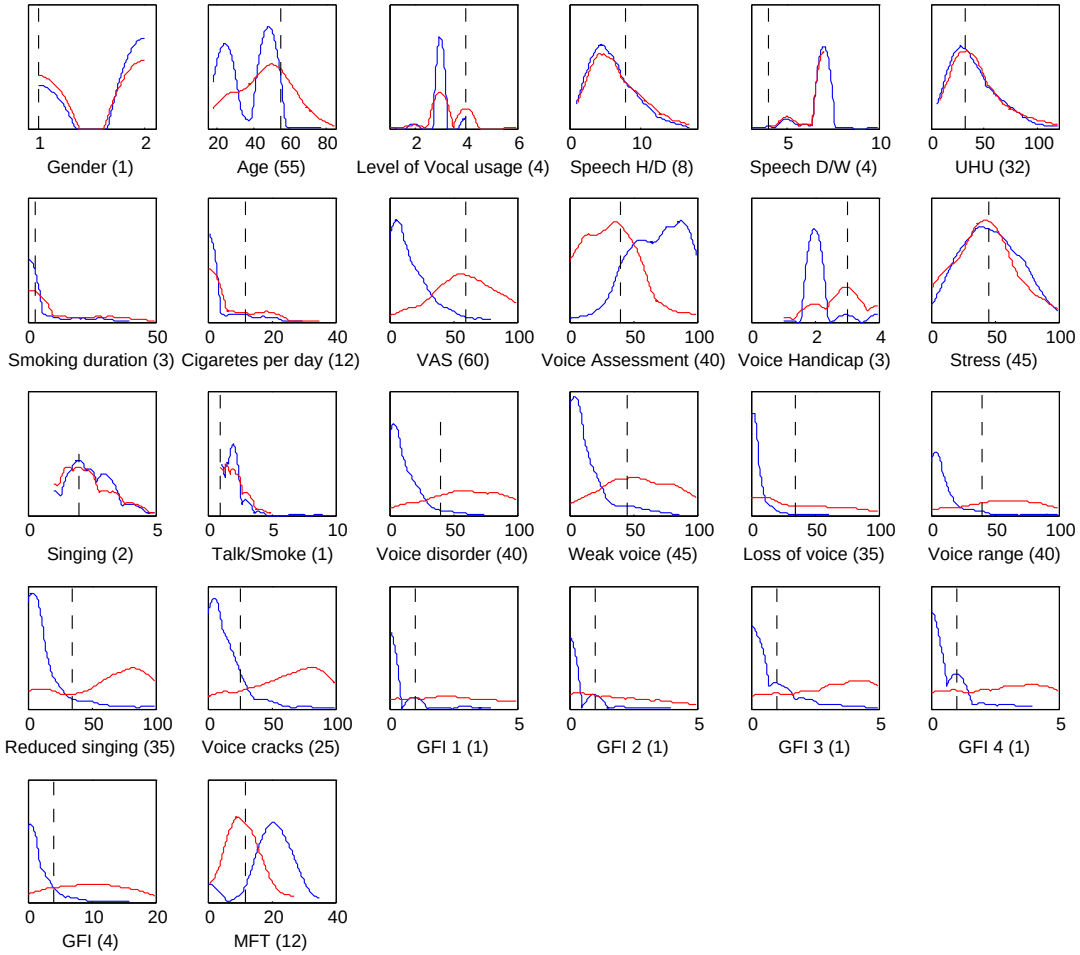

Figure 6. Probability (*y*-axis) density functions of query parameters for *healthy* (blue) and *pathological* cases (red). A dashed line and numbers in the parentheses correspond to a parameter value of a subject randomly selected from the database used to train the developed tools.

5. Discussion

The main goal of this study was to equip otolaryngologists with a transparent technique for analysis of voice and query data capable of providing support in screening for laryngeal disorders. In this work, screening concerns pathology detection through classification of subject's data into '*healthy*' and '*pathological*' classes as well as exploration of data and automatic decisions. To the best of our knowledge, none of the existing techniques exhibit such properties.

To achieve transparency, automatic decisions are obtained from a set of association rules extracted from query data and a decision tree designed using the most discriminative query parameter and six audio parameters routinely used by otolaryngologists in a clinical practice. Accurate pathology detection was observed on unseen subjects: EER of 11.1% was achieved using association rules and 10.2% using the decision tree. The EER was further reduced to 9.5% by combining results from these two classifiers. The obtained EER rates are lower compared to those achieved using sophisticated 'black box' techniques based on voice data only [22].

Besides a class label, the classification techniques also provide the level of confidence in the decision they suggest. The low EER achieved by the association rules-based classifier indicates high preventive health-care potential available in questionnaire data. These findings indicate that relevantly phrased queries can serve as very useful sensors in characterizing subject's health and are of great value in the education process of otolaryngology students.

The suggested novel way to assess pairwise similarity of observations allows mitigating the influence of irrelevant variables, since only variables contributing to a decision are used. Therefore, visualizations of multidimensional data represented by such pairwise similarities are more robust to noise compared to visualizations based on ordinary distance measures e.g., Euclidian distance. The *t*-SNE algorithm proved to be a useful tool for obtaining 2D data maps from data similarities. Such 2D maps are very helpful in comparative studies and providing deeper insights into data representing various groups of subjects. Ability to relate basic voice and query parameters to a specific diagnosis as well as to parameters of similar (in terms of position on a 2D map) subjects helps laryngologists to make associations and generalization over different cases. A 2D map is also of great help in identifying erroneously labelled data and other unexpected deviations in both voice and query data.

Probability density functions, created using the Epanechnikov kernel smoothing method, provide additional information, when evaluating patients with respect to audio or query data parameters. Visualization of probability density functions of the audio parameters, such as Jitter, Shimmer, HNR and SNR, was appreciated by otolaryngology specialists, since they use these parameters in their daily work. Probability density functions of query data provide additional information about a patient (in a statistical sense) and serve as valuable learning material for otolaryngology students. Decision trees provide additional insights concerning main voice parameters and are very useful for teaching/learning purposes.

6. Conclusions

1. The low EER achieved by the association rules-based classifier indicates high preventive health-care potential available in questionnaire data.
2. Relevantly phrased queries can serve as very useful sensors in characterizing subject's health and are of great value in the education process of otolaryngology students.
3. The suggested novel way to assess pairwise similarity of observations allows mitigating the influence of irrelevant variables, since only variables contributing to a decision are used.
4. A 2D map is of great help in comparative studies, providing deeper insights into data representing various groups of subjects, making associations and generalization over different cases, identifying erroneously labelled data and other unexpected deviations in both voice and query data.
5. Probability density functions provide additional information about a patient (in a statistical sense) and serve as valuable learning material for otolaryngology students.

Acknowledgments: This research was funded by a grant (No. MIP-075/2015) from the Research Council of Lithuania. The data set was collected by the Department of Otolaryngology, Lithuanian University of Health Sciences.

Author Contributions: Antanas Verikas and Evaldas Vaiciukynas conceived and designed the experiments; Jonas Minelga performed the experiments; Antanas Verikas, Marija Bacauskiene and Adas Gelzinis analyzed the data; and all authors contributed to writing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the curve
DET	Detection error trade-off
EER	Equal error rate
F0	Fundamental frequency
GFI	Glottal function index
HNR	Harmonics to noise ratio
MPT	Maximum phonation time
NNE	Normalized noise energy
ROC	Receiver operating characteristic
SNR	Signal-to-noise ratio

SSA Subjective self-assessment
t-SNE *t*-Distributed stochastic neighbor embedding
VAS Visual analog scale

References

1. Behrbohm, H.; Kaschke, O.; Nawka, T.; Swift, A. *Ear, Nose and Throat Diseases: With Head and Neck Surgery*, 3rd ed.; Thieme Medica: Stuttgart, Germany, 2009; pp. 293–329.
2. Nelson Roy, R.M.M.; Thibeault, S.; Parsa, R.A.; Gray, S.D.; Smith, E.M. Prevalence of voice disorders in teachers and the general population. *J. Speech Lang. Hear. Res.* **2004**, *47*, 281–293.
3. Verikas, A.; Gelzinis, A.; Bacauskiene, M.; Hallander, M.; Uloza, V.; Kaseta, M. Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders. *Artif. Intell. Med.* **2010**, *49*, 43–50.
4. Linder, R.; Albers, A.E.; Hess, M.; Poppl, S.J.; Schonweiler, R. Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *J. Voice* **2008**, *22*, 155–163.
5. Maier, A.; Haderlein, T.; Stelzle, F.; Noth, E.; Nkenke, E.; Rosanowski, F.; Schutzenberger, A.; Schuster, M. Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer. *EURASIP J. Audio Speech Music Process.* **2010**, *2010*, 926951.
6. Godino-Llorente, J.I.; Fraile, R.; Saenz-Lechon, N.; Osma-Ruiz, V.; Gomez-Vilda, P. Automatic detection of voice impairments from text-dependent running speech. *Biomed. Signal Process. Control* **2009**, *4*, 176–182.
7. Muhammad, G.; Mesallam, T.A.; Malki, K.H.; Farahat, M.; Mahmood, A.; Alsulaiman, M. Multidirectional Regression (MDR)-Based Features for Automatic Voice Disorder Detection. *J. Voice* **2012**, *26*, 817.
8. Horii, Y. Jitter and shimmer differences among sustained vowel phonations. *J. Speech Hear. Res.* **1982**, *25*, 12–14, doi:10.1044/jshr.2501.12.
9. Maryn, Y.; Corthals, P.; De Bodt, M.; Van Cauwenberge, P.; Deliyski, D. Perturbation Measures of Voice: A Comparative Study between Multi-Dimensional Voice Program and Praat. *Folia Phoniatr. Logop.* **2009**, *61*, 217–226.
10. Maryn, Y.; Bodt, M.D.; Barsties, B.; Roy, N. The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages. *Eur. Arch. Otorhinolaryngol.* **2013**, *271*, 1609–1619, doi:10.1007/s00405-013-2730-7.
11. Zhang, Y.; Jiang, J.J. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J. Voice* **2008**, *22*, 1–9.
12. Moran, R.J.; Reilly, R.B.; de Chazal, P.; Lacy, P.D. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 468–477.
13. Vaiciukynas, E.; Verikas, A.; Gelzinis, A.; Bacauskiene, M.; Minelga, J.; Hallander, M.; Padervinskis, E.; Uloza, V. Fusing voice and query data for non-invasive detection of laryngeal disorders. *Expert Syst. Appl.* **2015**, *42*, 8418–8426.
14. Verikas, A.; Bacauskiene, M.; Gelzinis, A.; Vaiciukynas, E. Questionnaire-versus voice-based, screening for laryngeal disorders. *Expert Syst. Appl.* **2012**, *39*, 6254–6262.
15. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
16. Smits, I.; Ceuppens, P.; De Bodt, M.S. A Comparative Study of Acoustic Voice Measurements by Means of Dr. Speech and Computerized Speech Lab. *J. Voice* **2005**, *19*, 187–196.
17. Vegiene, A. The Value of Voice Multidimensional Assessment in Screening of Laryngeal Disorders. Ph.D. Thesis, Lithuanian University of Health Sciences, Kaunas, Lithuania, 2014.
18. Padervinskis, E. The Value of Automatic Voice Categorization Systems Based on Acoustic Voice Parameters and Questionnaire Data in the Screening of Voice Disorders. Ph.D. Thesis, Lithuanian University of Health Sciences, Kaunas, Lithuania, 2016.
19. Wormald, R.N.; Moran, R.J.; Reilly, R.B.; Lacy, P.D. Performance of an automated, remote system to detect vocal fold paralysis. *Ann. Otol. Rhinol. Laryngol.* **2008**, *117*, 834–838.
20. Henriquez, P.; Alonso, J.B.; Ferrer, M.A.; Travieso, C.M.; Godino-Llorente, J.I.; Diaz-de Maria, F. Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1186–1195.

21. Martínez, D.; Lleida, E.; Ortega, A.; Miguel, A.; Villalba, J. Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit. In *Advances in Speech and Language Technologies for Iberian Languages*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 328, pp. 99–109.
22. Verikas, A.; Gelzinis, A.; Vaiciukynas, E.; Bacauskiene, M.; Minelga, J.; Hallander, M.; Uloza, V.; Padervinskis, E. Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone. *Med. Eng. Phys.* **2015**, *37*, 210–218.
23. Bacauskiene, M.; Verikas, A.; Gelzinis, A.; Vegiene, A. Random forests based monitoring of human larynx using questionnaire data. *Expert Syst. Appl.* **2012**, *39*, 5506–5512.
24. Bach, K.; Belafsky, P.; Wasylik, K.; Postma, G.; Koufman, J. Validity and Reliability of the Glottal Function Index. *Arch. Otolaryngol. Head Neck Surg.* **2005**, *131*, 961–964.
25. Gelzinis, A.; Verikas, A.; Bacauskiene, M. Automated speech analysis applied to laryngeal disease categorization. *Comput. Methods Progr. Biomed.* **2008**, *91*, 36–47.
26. Byeon, H. The Risk Factors of Laryngeal Pathology in Korean Adults Using a Decision Tree Model. *J. Voice* **2015**, *29*, 59–64.
27. Cordeiro, H.; Fonseca, J.; Meneses, C. Spectral envelope and periodic component in classification trees for pathological voice diagnostic. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, Chicago, IL, USA, 27–31 August 2014; pp. 4607–4610.
28. Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 Algorithms in Data Mining. *Knowl. Inf. Syst.* **2007**, *14*, 1–37.
29. Smith, T.C.; Frank, E. *Statistical Genomics: Methods and Protocols*; Springer: New York, NY, USA, 2016; pp. 353–378.
30. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile, 12–15 September 1994; pp. 487–499.
31. Tunç, B.; Dağ, H. Generating Classification Association Rules with Modified Apriori Algorithm. In Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED), Madrid, Spain, 15–17 February 2006; pp. 384–387.
32. Palanisamy, S. Association Rule Based Classification. Master's Thesis, Worcester Polytechnic Institute, Worcester, MA, USA, 2006.
33. Brummer, N.; de Villiers, E. The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. *arXiv* 2013, arXiv:1304.2865.
34. Saenz-Lechon, N.; Godino-Llorente, J.I.; Osma-Ruiz, V.; Gomez-Vilda, P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomed. Signal Process. Control* **2006**, *1*, 120–128.
35. Borg, I.; Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed.; Springer: New York, NY, USA, 2005.
36. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer: New York, NY, USA, 2007.
37. Van der Maaten, L.; Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
38. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*, 1st ed.; Monographs on Statistics and Applied Probability; Chapman and Hall: London, UK, 1986.