Artificial Intelligence in Medicine 49 (2010) 43–50

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders

Antanas Verikas [a,b,*], Adas Gelzinis [a], Marija Bacauskiene [a], Magnus Hållander [b], Virgilijus Uloza [c], Marius Kaseta [c]

[a] Department of Electrical & Control Equipment, Kaunas University of Technology, Studentu 50, LT-51368, Kaunas, Lithuania
[b] Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden
[c] Department of Otolaryngology, Kaunas University of Medicine, Eiveniu 2, LT-50009 Kaunas, Lithuania

ARTICLE INFO

Article history:
Received 16 September 2008
Received in revised form 19 January 2010
Accepted 16 February 2010

Keywords:
Classification committee
Support vector machine
Multiple feature sets
Variable selection
Larynx pathology

ABSTRACT

Objective: This paper is concerned with soft computing techniques for categorizing laryngeal disorders based on information extracted from an image of patient's vocal folds, a voice signal, and questionnaire data.

Methods: Multiple feature sets are exploited to characterize images and voice signals. To characterize colour, texture, and geometry of biological structures seen in colour images of vocal folds, eight feature sets are used. Twelve feature sets are used to obtain a comprehensive characterization of a voice signal (the sustained phonation of the vowel sound /a/). Answers to 14 questions constitute the questionnaire feature set. A committee of support vector machines is designed for categorizing the image, voice, and query data represented by the multiple feature sets into the *healthy*, *nodular* and *diffuse* classes. Five alternatives to aggregate separate SVMs into a committee are explored. Feature selection and classifier design are combined into the same learning process based on genetic search.

Results: Data of all the three modalities were available from 240 patients. Among those, 151 patients belong to the nodular class, 64 to the diffuse class and 25 to the healthy class. When using a single feature set to characterize each modality, the test set data classification accuracy of 75.0%, 72.1%, and 85.0% was obtained for the image, voice and questionnaire data, respectively. The use of multiple feature sets allowed to increase the accuracy to 89.5% and 87.7% for the image and voice data, respectively. The test set data classification accuracy of over 98.0% was obtained from a committee exploiting multiple feature sets from all the three modalities. The highest classification accuracy was achieved when using the SVM-based aggregation with hyper parameters of the SVM determined by genetic search. Bearing in mind the difficulty of the task, the obtained classification accuracy is rather encouraging.

Conclusions: Combination of both multiple feature sets characterizing a single modality and the three modalities allowed to substantially improve the classification accuracy if compared to the highest accuracy obtained from a single feature set and a single modality. In spite of the unbalanced data sets used, the error rates obtained for the three classes were rather similar.

1. Introduction

In clinical practice, the diagnostic procedure of laryngeal diseases is based on evaluation of patient's complaints, history, and data of instrumental as well as histological examination. During the last years a variety of techniques for examination of the larynx and objective measurements of voice quality have been developed [1,2]. Evaluation of larynx has improved significantly with the establishment of the computer tomography (CT) and magnetic resonance imaging (MRI), as the technologies provide insights into the endoscopically blind areas and reveal depth of tumour infiltration. The technologies may be beneficial in staging larynx carcinoma and planning the most appropriate surgical procedure [3–6]. Ultrasonography is useful in cases of larger laryngeal lesions and may have some role in screening unilateral vocal fold pathologies. At the same time, further fine-tuning of the technique may be necessary [7,8].

When resorting to automated characterization of human larynx, laryngeal images, voice signal and patient's questionnaire data can be considered as the main information sources for the characterization. Nowadays, automated analysis of voice is

[*] Corresponding author at: Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden. Tel.: +46 35 167140; fax: +46 35 216724.
E-mail addresses: antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), marija.bacauskiene@ktu.lt (M. Bacauskiene), magnus.hallander@hh.se (M. Hållander), virgilijus.uloza@kmuk.lt (V. Uloza), marius.kaseta@kmuk.lt (M. Kaseta).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.02.002

increasingly used for detecting and screening laryngeal pathologies [9–16]. It was demonstrated that even telephone-based voice may lend itself for screening laryngeal disorders [11]. According to Hadjitodorov and Mitev, depending on the disease and its stage, the following changes can be observed in the vocalized voice signal in pathological cases [10]:

i Significant cycle-to-cycle pitch and amplitude perturbations;
ii Decrease of the voice signal fundamental frequency and amplitude;
iii Dominance of the first harmonic in the signal spectrum;
iv Presence of a turbulent noise;
v Decrease or loss of the harmonics over 1 kHz and presence of sub-harmonics;
vi Pauses in the pitch period generation.

There were very few attempts to create systems for automated analysis of colour laryngeal images. In [17], a technique for automated categorization of manually marked suspect lesions into *healthy* and *diseased* classes was presented. The categorization is based on textural features extracted from co-occurrence matrices computed from manually marked areas of vocal fold images. The classification accuracy of 81.4% was reported when testing the system on a very small set of 35 images. A much larger set of laryngeal images has been used in studies presented in [18,19]. The algorithms developed exploit features of various types and do not require any manual marking.

Attempts to exploit the patient's questionnaire data for screening laryngeal disorders are even more scarce. The questionnaire data may carry information, which is not present in the acoustic or visual modalities. In [20], a genetic search and support vector machine (SVM) based technique to categorize the patient's questionnaire data was presented. The categorization results provide an indication on usefulness of the data for screening laryngeal pathologies.

The long-term goal of this work is a decision support system for diagnostics of laryngeal diseases. A voice signal, colour images of vocal folds, and questionnaire data are the information sources used in the analysis. This paper is concerned with exploiting the three information sources mentioned above for categorizing laryngeal diseases. An SVM is used as classifier to make the categorization. Variable selection and classifier design is integrated into the same learning process based on genetic search.

2. The data

The mixed gender local database has been used in this study. The medical task considered in this paper concerns the laryngeal colour images, voice signal, and the query data based automated categorization of laryngeal disorders into three decision classes: *healthy* and two *pathological* classes, namely *diffuse* and *nodular* mass lesions of vocal folds [18]. The pathological classes can be characterized as follows. A rather common, clinically discriminative group of laryngeal diseases was chosen for the analysis, i.e. mass lesions of vocal folds. Mass lesions of vocal folds could be categorized into six classes, namely, *polypus*, *papillomata*, *carcinoma*, *cysts*, *keratosis*, and *nodules*. This categorization is based on clinical signs and histological structure of the mass lesions of vocal folds. We distinguished two groups of mass lesions of vocal folds, i.e. *nodular* lesions (localized thickenings) – *nodules*, *polyps*, and *cysts*– and *diffuse* lesions—*papillomata*, hyperplastic laryngitis with *keratosis*, and *carcinoma*. Clinically, nodular lesions (localized thickenings) visually appear as single lesions of various sizes with a smooth, regular surface and distinct margins surrounded by a normal tissue of the vocal fold. Respectively, diffuse lesions visually appear as irregular, rough, multiple thickenings without distinct margins, often surrounded by an inflamed tissue. It is worth stressing that according to the task of the study, the categorization into the nodular and diffuse classes was based on visual appearance of vocal fold mass lesions, evaluated under direct micro-laryngoscopy. However, the final diagnosis was confirmed by histological examination of laryngeal specimens removed during endolaryngeal microsurgical intervention.

Laryngeal images have been recorded at the Department of Otolaryngology, Kaunas University of Medicine, Lithuania. The images were acquired during routine direct micro-laryngoscopy employing the *Moller-Wedel Universa* 300 surgical microscope. The 3-CCD *Elmo* colour video camera of 768 × 576 pixels was used to record the images. To lessen the influence of variation of the image capturing conditions on image appearance, we apply the multi-scale *retinex* theory-based colour image enhancement [21,22]. Details on how the enhancement has been applied can be found in [18].

Voice recordings of the sustained phonation of the vowel sound /a/ (as in the English word "large") are the voice signals utilized. There are three voice recordings from each subject. The average length of each recording is 2.4 s. The recordings are made in the "*wav*" file format at 44,100 samples/s rate. There are 16 bits allocated for one sample. During preprocessing, the beginning and the end of each recording was eliminated. The D60S Dynamics Vocal microphone has been used to make the recordings.

There are fourteen questions in the questionnaire utilized in this study.

3. Feature sets

3.1. Features extracted from colour images

Fig. 1 presents characteristic examples from the three decision classes considered, namely, *nodular*, *diffuse*, and *healthy*. However, it is worth noting that due to the large variety of appearance of vocal fold mass lesions, the classification task can sometimes be difficult even for a trained physician [23,24].

Eight types of features are used to characterize colour, texture, and geometry of biological structures seen in colour images of vocal folds [19,18,25]. The list of feature types used is given below. The features are given by the first kernel principal components extracted from each type of measurements. The number of components (features) used is such that 99.5% of variance is accounted for by the components utilized. In the parentheses, the number of features used is provided.

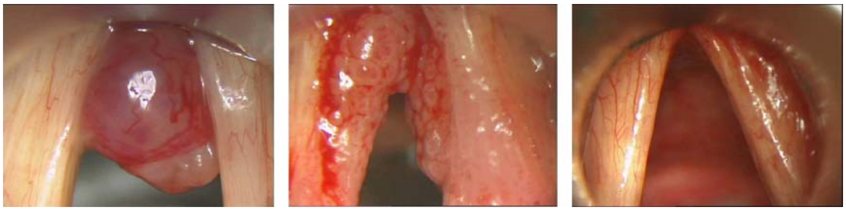

Fig. 1. Images from the *nodular* (left), *diffuse* (middle), and *healthy* (right) classes.

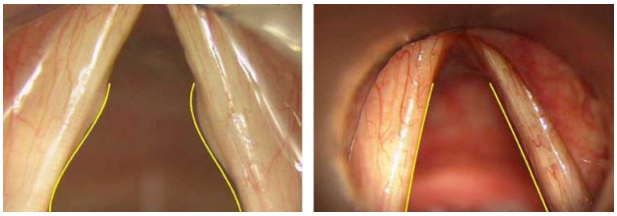

Fig. 2. Laryngeal images coming from the *nodular* (left) and *healthy* (right) classes along with two third order curves used to calculate the geometrical features.

(i) The probability distribution of colour represented by the 3D colour histogram, (80) [18].

(ii) Features calculated from the co-occurrence matrices. A polynomial $p(x)$ of degree n

$$p(x) = p_0 + p_1x + \cdots + p_{n-1}x^{n-1} + p_nx^n \quad (1)$$

has been fitted to the values of each of the 14 Haralick's coefficients [26] calculated from the co-occurrence matrices evaluated for several distance parameter values. Kernel principal components are then extracted from the parameters of the polynomials and used as features, (33) [27].

(iii) The distribution of responses obtained from the multi-channel Gabor filtering. An image is filtered by a bank of Gabor filters of frequency f and orientation θ. Having the filtered image, a 24-bin histogram of the image is calculated. Thus, having N_f frequencies and N_θ orientations, $N_f \times N_\theta$ of such histograms are obtained from one image. Leaving the first bin of the histograms aside, the other bins are concatenated into one long vector. Features are then given by the first kernel principal components of the vector, (182) [18].

(iv) Fourier spectrum based features characterizing the distribution of image frequencies in frequency rings. The frequency plane is divided into several rings R_i of different average frequency. The Chi-square χ_i and the entropy M_i of the Fourier power are then computed in each of the rings and used to extract the image frequency content based features (F_2), (17) [19].

(v) Fourier spectrum based features characterizing the distribution of image frequencies in frequency wedges. To compute the feature vector, the upper part of the frequency plane is divided into M equidistant wedges W_i and the average power is computed in each of the wedges. The average power values are used to extract features (F_1), (75) [19].

(vi) The distribution of the image intensity gradient direction. We use a histogram to represent the distribution of the gradient angle. The histogram vector is then projected onto the space spanned by the first eigenvectors of the kernel covariance matrix. The vector of the kernel principal components is utilized as a feature vector of this type, (65) [19].

(vii) The run-length matrices based features. Seven features, *short-run emphasis*, *long-run emphasis*, *grey-level non-uniformity*, *run-length non-uniformity*, *run percentage*, *low grey level run emphasis*, and *high grey level run emphasis* [28] have been extracted based on the run-length matrices. Since red colour dominates in the vocal fold images, the $a^*(x, y)$ (*red-green*) image component ($L^*a^*b^*$ colour space) has been employed for extracting the run-length matrices based features (27) [25].

(viii) Geometrical features characterizing the shape of the edges of vocal folds. Three polynomial curves given by Eq. (1) – one of the first, one of the second, and one of the third order – were fitted to the lower part of edges of vocal folds. Thus, in total, we have 18 parameters p_i characterizing the six curves. Fig. 2 presents two examples of laryngeal images coming from the healthy and nodular classes along with the third order polynomial curves found.

To extract geometrical features, two more geometrical measurements are made. A vocal fold image is first segmented into a set of homogenous regions. Two lines, ascending in the left-hand part and descending in the right-hand part of the image are then drawn in such a way as to maximize the number of segmentation boundary points intersecting the lines. Fig. 3 presents two examples of the segmentation boundaries found and the two lines drawn according to the determined directions. The first geometrical measurement is then given by the sum of the squared number of the boundary points intersecting the two lines. The second geometrical measurement is obtained in the same way, except that colour edge points are utilized instead of the segmentation boundary points. These two geometrical measurements together with the 18 parameters mentioned above are then subjected to the kernel principal component analysis and used as features, (40) [19,25].

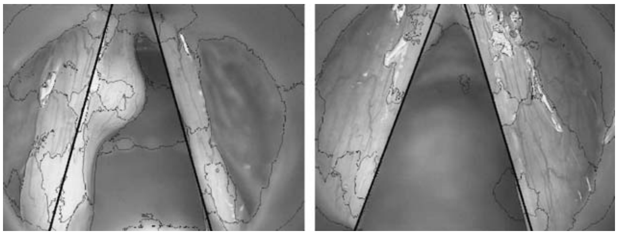

Fig. 3. Vocal cord images coming from the *nodular* (left) and the *healthy* (right) classes along with two lines used to calculate the geometrical feature.

3.2. Features extracted from a voice signal

In this study, we used 12 different feature sets, presented in the list below. In the parentheses, the number of features of each type is provided. A comprehensive description of the first eleven feature sets can be found in [16].

1. Pitch and amplitude perturbation measures, (24).
2. Frequency features, (100).
3. Mel-frequency features, (35).
4. Cepstral energy features, (100).
5. Mel-frequency cepstral coefficients, (35).
6. Autocorrelation features, (80).
7. Harmonics to noise ratio in spectral domain, (11).
8. Harmonics to noise ratio in cepstral domain, (11).
9. Linear prediction coefficients. It is known that different number of coefficients p is required to model male and female voices [29,30]. According to [29], $p = 33$ for female and $p = 44$ for male voices. Since we are dealing with a mixed gender database, we modeled the voice signals two times using $p = 33$ and $p = 44$. Thus, we have 77 features in total, (77).
10. Linear prediction cosine transform coefficients, (77).
11. Feature set used in the commercial "Dr.Speech" software, (23).
12. Signal shape. Several periods of a voice signal are averaged and represented by the signal amplitude at 128 equally spaced points. Features are given by the amplitude values. Fig. 4 presents an example of the voice signal, (128).

3.3. Query features

The query data are represented by the following features (components x_i of the data vector $\mathbf{x}$):

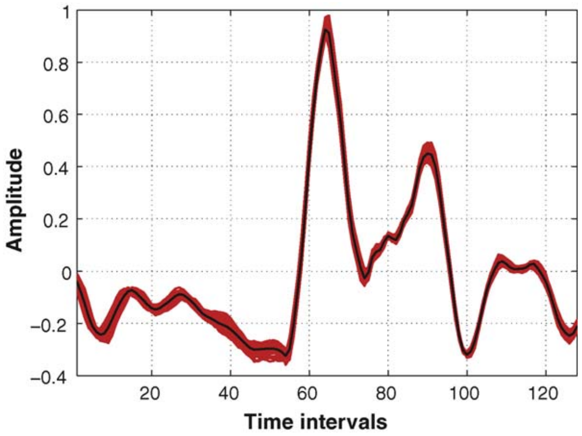

Fig. 4. Several superimposed periods and the averaged (bold line) voice signal.

1. Patient's age;
2. Duration of voice disorder (months);
3. Patient's education (five grades);
4. Average duration of intensive speech use (h/day);
5. Number of days of intensive speech use (days/week);
6. Smoking (yes/no);
7. Smoked cigarets/day;
8. Smoking duration (years);
9. Subjective voice function assessment by the patient on Visual Analogue Scale. The patients were asked to rate hoarseness on this scale, ranging from 0 ("no hoarseness") to 100 ("severe hoarseness");
10. Maximum phonation time (s);
11. Functional domain of the voice handicap index (VHI) (F);
12. Emotional domain of VHI (E);
13. Physical domain of VHI (P);
14. Voice Handicap Index (VHI) (the maximum value is 120), assessed from answers to questions from a specially designed questionnaire and was used in this study to evaluate person's level of handicap resulting from a voice disorder [31].

Thus, in total, the data are represented by 21 feature sets, 8 feature sets are extracted from an image, 12 from a voice signal and 1 from a questionnaire.

4. The classifier

4.1. The basic classifier

A support vector machine is used as the basic classifier in this work. Depending on the definition of the optimization problem, several forms of SVM can be distinguished, for example, 1-norm or 2-norm SVM. Since there are examples demonstrating that the 1-norm SVM outperforms the 2-norm, especially if there are redundant noise features [32], the 1-norm SVM is used in this work. Assuming that $\Phi(\mathbf{x})$ is the non-linear mapping of the data point $\mathbf{x}$ into the new space, the 1-norm soft margin SVM can be constructed by solving the following minimization problem [33]:

$$\min_{\mathbf{w},b,\gamma,\xi} -\gamma + C\sum_{i=1}^{N}\xi_i \tag{2}$$

subject to

$$y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b) \geq \gamma - \xi_i, \xi_i \geq 0, \|\mathbf{w}\|^2 = 1, \quad i = 1, \ldots, N \tag{3}$$

where $\mathbf{w}$ is the weight vector, $y_i = \pm 1$ is the desired output (± 1), N is the number of training data points, $\langle\ \rangle$ stands for the inner product, γ is the margin, ξ_i are the slack variables, b is the threshold, and C is the regularization constant controlling the trade-off between the margin and the slack variables. The discriminant function for a new data point $\mathbf{x}$ is given by

$$f(\mathbf{x}) = \mathcal{H}\left[\sum_{i=1}^{N}\alpha_i^* y_j k(\mathbf{x}, \mathbf{x}_i) + b^*\right], \tag{4}$$

where $k(\mathbf{x}, \mathbf{x}_i)$ stands for the kernel and the Heaviside function $\mathcal{H}[y(\mathbf{x})] = -1$, if $y(\mathbf{x}) \leq 0$ and $\mathcal{H}[y(\mathbf{x})] = 1$ otherwise. In this work, the Gaussian kernel, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma\}$, has been used. The optimal values α_i^*, b^* of the parameters α_i and b are found during training. The salient variables x_i, the regularization constant C and the Gaussian width σ have been found by the genetic search.

Since an SVM is a binary classifier while the task is to distinguish between three classes, the one-against-one scheme is used to make the categorization. To make a decision, outputs of the binary SVM are converted into probabilities and the class corresponding to the highest probability is chosen.

4.2. The committee

A separate SVM is used for each set of features. Decisions obtained from the separate SVMs are then combined into a committee decision. Five aggregation alternatives, discussed in Section 6.2, have been studied. SVM committees have been successfully used in several studies, for instance concerning facial expression recognition [34,35]. One can also consider the AdaBoost-based approach to committee design. However, the AdaBoost tends to over-fitting for higher noise levels [36].

5. Feature selection and classifier design

The issue of selecting an optimal subset of relevant features plays also an important role in successful design of a pattern recognition system. Some of features, that can be measured in many pattern recognition applications, may be redundant or even irrelevant. Usually better performance may be achieved by discarding such features. Moreover, as the number of features used grows, the number of training samples required grows exponentially. Therefore, in many practical applications we need to reduce the dimensionality of the data.

Feature selection in general is a difficult problem. In a general case, only an exhaustive search can guarantee an optimal solution. A large variety of feature selection techniques that result in a sub-optimal feature set have been proposed [37,38], ranging from the sequential forward and backward selection to sequential forward floating selection [39], genetic [40], tabu [41], or branch and bound algorithm-based search [42]. Genetic search-based selection is one of the most promising approaches to feature selection. However, genetic search procedures are rather time consuming. Therefore, when selecting features for individual classifiers (SVMs), we reduced the number of features based on the feature saliency-based ranking first and then applied genetic search to the remaining set of features. Such feature reduction was applied to the feature sets containing more than 32 features only. This value was found experimentally. When designing a classifier not only a feature set, but also the hyper-parameter values of the classifier are to be selected. Since we use an SVM with the Gaussian kernel as a basic classifier, values of two hyper-parameters, namely the kernel width σ and the regularization constant C are to be selected. Genetic search has been used to find values of these parameters. We combine selection of features and values of the hyper parameters into one learning process based on genetic search.

5.1. Feature saliency

The feature saliency measure used in this work is based on two factors, namely, the fuzzy derivative of the classifier output with respect to the feature and the similarity between the feature and the feature set [43]. Having a feature set F, the following feature saliency score Γ_i is assigned to the i th feature [43]:

$$\Gamma_i = \frac{\beta\Upsilon_i}{\max_{j=1,\ldots,N}\Upsilon_j} + (1-\beta)\bar{\lambda}_{i,F} \tag{5}$$

where Υ_i is the classifier output sensitivity-based feature saliency measure, $\bar{\lambda}_{i,F}$ stands for the average similarity between the i th feature and the set F, N is the number of features and β is a parameter.

The classifier output sensitivity based saliency measure for the i th feature is given by [43]

$$\Upsilon_i = \frac{1}{PQ}\sum_{p=1}^{P}\sum_{j=1}^{Q}|\bar{y}'_j(\tilde{D}_{ip})| \tag{6}$$

where P is the number of data points, $\tilde{D}_{ip}$ is the p th fuzzy location for the i th feature, Q is the number of classes (outputs) and $\bar{y}'_j(\tilde{D}_{ip})$ is the defuzzified value of the derivative $y'_j(\tilde{D}_{ip})$ of the j th output with respect to the input feature x_i at the fuzzy location $\tilde{D}_{ip}$. The derivative $y'_j(\tilde{D}_{ip})$ is a fuzzy set. The average similarity $\bar{\lambda}_{i,F}$ is calculated as follows [43]. Let for two features i and j $\lambda(i,j)$ be given by [44]

$$\lambda(i,j) = \frac{1}{2}[\mathrm{var}(i) + \mathrm{var}(j) - \sqrt{[\mathrm{var}(i)+\mathrm{var}(j)]^2 - 4\mathrm{var}(i)\mathrm{var}(j)[1-\rho(i,j)^2]}] \tag{7}$$

where var(i) stands for the feature i variance and $\rho(i,j)$ is the coefficient of correlation between the features i and j.

In the multi-class case, we calculate the average value:

$$\bar{\lambda}(i,j) = \frac{1}{Q}\sum_{k=1}^{Q}\lambda_k(i,j) \tag{8}$$

where Q is the number of classes and $\lambda_k(i,j)$ stands for the measure value calculated using data coming from the k th class. The measure is normalized:

$$\bar{\lambda}_n(i,j) = \frac{\bar{\lambda}(i,j)}{\max_{i,j\in F}\bar{\lambda}(i,j)} \tag{9}$$

with F being a feature set. The similarity between the feature i and the set F is then given by

$$\bar{\lambda}_{i,F} = \min_{j\in F}\bar{\lambda}_n(i,j) \tag{10}$$

5.2. Genetic search

The most important issues to consider when solving a problem by genetic search are encoding of the problem into a *chromosome* and evaluation, where the genetic representation of the problem and the *fitness function* for evaluating the suggested solution are defined [45]. Once the encoding and evaluation are defined, GA randomly generates the population of the probable solutions in the form of chromosomes. Members of the population are evaluated using the fitness function and, based on the evaluation results, a portion of the members are selected for subjection to the genetic operations. The higher the fitness function value, the higher is the selection probability. *Crossover* and *mutation* are the genetic operations applied. In crossover, pairs of parents are combined to create new chromosomes called *offsprings*. In mutation, random changes on the genes are introduced. Thus, information representation in a chromosome, generation of initial population, evaluation of population members, selection, crossover, mutation, and reproduction (survival) are the issues to consider when designing a genetic search algorithm.

In our case, a chromosome contains all the information needed to build an SVM classifier. We divide the chromosome into three parts. One part encodes the regularization constant C, one the kernel width parameter σ, and the third one encodes the inclusion/noninclusion of features. To generate the *initial population*, the features are masked randomly and values of the parameters C and σ are chosen randomly from the interval $[C_0 - \Delta C, C_0 + \Delta C]$ and $[\sigma_0 - \Delta\sigma, \sigma_0 + \Delta\sigma]$, respectively, where C_0 and σ_0 are the very approximate parameter values obtained from the experiment. The fitness function used to evaluate chromosomes is given by the classification accuracy of the validation set data.

The *selection process* of a new population is governed by the fitness values. The selection probability of the i th chromosome p_i is given by

$$p_i = \frac{r_i}{\sum_{j=1}^{M} r_j} \tag{11}$$

where r_i is the correct classification rate obtained from the classifier encoded in the i th chromosome and M is the population size.

The *crossover operation* for two selected chromosomes is executed with the probability of crossover p_c. If a generated random number from the interval [0,1] is smaller than the crossover probability p_c, the crossover operation is executed. Crossover is performed separately in each part of a chromosome. The crossover point is randomly chosen in the "feature mask" part and two parameter parts. The corresponding parts of two chromosomes selected for the crossover operation are exchanged at the chosen points.

The *mutation operation* adopted is such that each gene is selected for mutation with the probability p_m. The mutation operation is executed independently in each chromosome part. If the gene selected for mutation is in the feature part of the chromosome, the value of the bit representing the feature in the feature mask (0 or 1) is reversed. To execute mutation in the parameter part of the chromosome, the value of the offspring parameter determined by the selected gene is mutated by $\pm\Delta\gamma$, where γ stands for C or σ, as the case may be. The mutation sign is determined by the fitness values of the two chromosomes, namely the sign resulting into a higher fitness value is chosen. The way of determining the mutation amplitude $\Delta\gamma$ is somewhat similar to that used in [46] and is given by

$$\Delta\gamma = w\beta(\max(|\gamma - \gamma_{p1}|, |\gamma - \gamma_{p2}|)) \tag{12}$$

where γ is the actual parameter value of the offspring, $p1$ and $p2$ stand for parents, $\beta \in [0,1]$ is a random number, and w is the weight decaying with the iteration number:

$$w = k\alpha^t \tag{13}$$

where t is the iteration number, $\alpha = 0.95$ and k is a constant. The constant k defines the initial mutation amplitude. The value of $k = 0.4$ worked well in our tests.

In the *reproduction process*, the newly generated offspring replaces the chromosome with the smallest fitness value in the current population, if a generated random number from the interval [0,1] is smaller than the reproduction probability p_r or if the fitness value of the offspring is larger than that of the chromosome with the smallest fitness value.

Table 1
The statistics of the data used in the experiments.

Data type	Patients	Nodular	Diffuse	Healthy	Records
Voice	316	151	64	101	948
Image	270	157	78	35	1349
Questionnaire	260	157	78	25	260

Table 2
The test set data classification accuracy obtained from the separate classifiers.

Feature type	*N* # features	Classification accuracy (%)
Voice		
Perturbation	13	65.5
Frequency	17	66.0
Mel-frequency	3	61.7
Cepstrum	45	69.0
Mel-coefficients	9	67.0
Autocorrelation	5	61.6
HNR-spectral	4	60.0
HNR-cepstral	4	60.0
LP-coefficients	14	67.9
LPCT-coefficients	22	72.1
DrSpeech	11	66.4
Signal shape	21	69.9
Image		
Colour	13	69.6
Co-occurrence	15	75.0
Gabor	19	65.6
Frequency, F_2	9	63.3
Frequency, F_1	21	69.8
Gradient	14	71.7
Run-length	10	68.9
Geometrical	13	76.4
Questionnaire		
Questionnaire	8	85.0

6. Experimental investigations

6.1. Experimental setup

Data of all the three types were available from 240 patients. Amongst those, 151 patients belong to the nodular class, 64 to the diffuse class and 25 to the healthy class. The data were randomly split into the learning set S_l containing data of 200 patients and the test set S_t with data of 40 patients. In all the tests involving estimation of the classification accuracy, we run an experiment 25 times with different random split of the data set into the learning and tests subsets. The results presented here are average values calculated from such 25 runs. In addition to this set of 240 patients, there were available data from patients characterized by only one or two data modalities, out of the three. These data have been used for training only. Several voice and image recordings were available from one patient. Table 1 presents statistics of the data used in the experiments. The data used were normalized to zero mean and variance one.

The genetic search lasted for 80 generations with the following parameters: the population size was set to 75, the number of offsprings produced for creating the next population was equal to 40, and the probability of including a variable into the initial chromosome was set to 0.3. The values of crossover, mutation and reproduction probabilities were found experimentally. The following values worked well in the tests: $p_c = 0.95$, $p_m = 0.02$, and $p_r = 0.05$. The appropriate β value in Eq. (5) has been found to be $\beta = 0.6$.

6.2. Results

In the first experiment, a separate SVM was used for each type of features. The genetic search was applied for finding both values of the hyper-parameters C and σ, and relevant features. Features in the sets containing more than 32 features were first ranked using the saliency score Γ_i and then the first 32 features were subjected to the genetic search. Table 2 summarizes results of the tests. In Table 2, shown is the number of features providing the best performance along with the test data set classification accuracy. Surprisingly enough, the questionnaire features exhibited the highest classification accuracy. It seems that features extracted from images of vocal folds are more informative than those characterizing voice signals.

In the next experiment, multiple feature sets have been used to solve the classification task. The a posteriori probabilities obtained from the separate SVMs were aggregated in different ways. Five aggregation alternatives presented below have been considered.

1. The class a posteriori probabilities obtained using a separate feature set for each SVM (previous experiment) are averaged exploiting all the separate classifiers.
2. The average class a posteriori probabilities are calculated using not all, but selected separate classifiers. A technique of sequential forward member inclusion was applied. The first member selected for classification was the one providing the highest classification accuracy. The classifier included into the committee in the j step of the designing procedure was that providing the highest classification accuracy of the committee. The process continued until all the members were included into the committee. Thus, a series of committees with the increasing number of members was created. The final committee chosen was that providing the highest performance.
3. The a posteriori probabilities obtained from the separate SVMs are considered as new variables. Outputs of the separate SVMs are aggregated into a committee output via the k-NN classification rule applied to these variables. The Euclidean distance measure is used to find the nearest neighbors. All three outputs of all the separate SVMs are used to calculate the distance.
4. The same as the third alternative, except that the sequential forward variable selection is used to determine the variables used to calculate the distance.

Table 3
The average test set data classification accuracy (%) obtained from the five aggregation alternatives, along with the average number of new variables (given in the parentheses) used in the classification.

Features	Voice	Image	V + I	All
Aggregation				
1	69.7 ± 2.6 (36.0)	71.3 ± 2.7 (24.0)	69.7 ± 2.9 (60.0)	71.6 ± 2.9 (63.0)
2	80.4 ± 1.9 (9.6)	81.3 ± 2.1 (9.0)	86.0 ± 2.0 (13.5)	91.7 ± 1.7 (10.5)
3	74.8 ± 2.2 (36.0)	78.6 ± 2.8 (24.0)	83.6 ± 2.5 (60.0)	87.6 ± 2.9 (63.0)
4	86.4 ± 1.9 (9.1)	87.0 ± 2.5 (6.3)	91.1 ± 1.9 (6.4)	95.7 ± 1.2 (5.9)
5	87.7 ± 1.8 (14.5)	89.5 ± 2.4 (6.9)	94.5 ± 1.7 (22.6)	98.5 ± 0.8 (22.6)

Table 4
Confusion matrix for the ‘All features’’ case and the 4th aggregation alternative providing the classification accuracy of 95.7 ± 1.2%.

Prediction	True			
	Nodular	Diffuse	Healthy	Total
Nodular	23.40	1.06	0.09	24.55
Diffuse	0.58	9.77	0	10.35
Healthy	0	0	5.09	5.09
Total	23.98	10.84	5.18	40.00

Table 5
Confusion matrix for the “All features” case and the 5th aggregation alternative providing the classification accuracy of 98.5 ± 0.8%.

Prediction	True			
	Nodular	Diffuse	Healthy	Total
Nodular	23.77	0.29	0.08	24.15
Diffuse	0.21	10.53	0	10.74
Healthy	0	0.009	5.10	5.11
Total	23.98	10.84	5.18	40.00

5. Aggregation by an SVM classifier. The new variables are used as features for the classifier. Both the hyper-parameters and features are selected via the genetic search. Features are selected from all the new variables.

Table 3 presents the average test set data classification accuracy obtained from the five aggregation alternatives, where the 95% confidence interval is also provided. In the parentheses, the average number of new variables used in the classification is presented.

As can be seen from Table 3, the classification accuracy greatly depends on the type of aggregation applied. Simple averaging (alternative 1) even deteriorates the classification accuracy if compared to the best single classifier (see Table 2). It is well known, that averaging is a robust technique to aggregate decisions and very often improves the classification accuracy [47,48]. In our case, however, some single classifiers provide a very low classification accuracy. Therefore, the simple aggregation does not improve classification accuracy over the best single classifier. The classification accuracy is improved considerably when aggregating the most appropriate selected classifiers; the alternative 2 in Table 3. The same pattern of behavior is observed when using the *k*-NN classifier; the 3rd and 4th alternatives. The highest classification accuracy is achieved when using the genetic search based aggregation exploiting the space of all the new variables (the class a posteriori probabilities). However, the rather simple aggregation by the *k*-NN classifier is also very effective.

As can be seen from Tables 2 and 3, the classification accuracies obtained using the three modalities (voice, image, and questionnaire data) are rather similar; 87.7%, 89.5% and 85.0%, respectively. However, when exploiting all the three modalities, a considerable improvement in the accuracy is obtained.

To further explore the classification results, we present two confusion matrices for two aggregation alternatives providing the highest classification accuracy (alternative 4 and 5 for the all features case). The matrices were computed for the test set data and are given in Tables 4 and 5. We have 40 test data points in one experiment. However, we run the experiment 25 times using different random split of the available data into the learning and test sets. Thus, the numbers presented in Tables 4 and 5 are the average numbers calculated from these 25 trials. The confusion matrices reveal that, in spite of the unbalanced data sets, the error rates obtained for the three classes are rather similar. For example, when using the 5th aggregation alternative, correctly classified are 99.1%, 97.1%, and 98.5% of the test data coming from the nodular, diffuse, and the healthy classes, respectively.

The developed algorithms were implemented using MATLAB and C++ programming languages. The C++ based implementation is used in the operating phase (after training). Feature extraction is the most time-consuming part of the analysis in both image and voice processing. In the operating phase, processing of one image and one voice record take approximately 4 and 1 s, respectively, when using 2.0 GHz Laptop. By optimizing code of several feature extraction subroutines, the processing time can be reduced to about 1 s. The time required to process query data is negligible if compared to the image or voice processing. Currently, the analysis is performed off-line, since different hardware is used to record image and voice data.

7. Conclusions

A collection of soft computing techniques was developed for categorizing laryngeal disorders based on information extracted from three modalities: an image of patient’s vocal folds, a voice signal and questionnaire data. The data represented by multiple feature sets were categorized into the *healthy*, *nodular* and *diffuse* classes. The effectiveness of single SVM classifiers as well as committees of classifiers was studied.

It was found that features extracted from images of vocal folds are more informative than those characterizing voice signals. However, surprisingly enough, the questionnaire features exhibited the highest classification accuracy. Five alternatives to aggregate information available from multiple feature sets were considered. The classification accuracy obtained greatly depended on the type of aggregation applied. Due to classifiers of rather low accuracy, simple averaging of all available classifiers even deteriorated the classification accuracy if compared to the best single classifier. The classification accuracy improved considerably when aggregating the most appropriate selected classifiers. The highest classification accuracy was achieved when using the SVM and genetic search based aggregation exploiting the space the class a posteriori probabilities. The SVM based aggregation provided a higher classification accuracy than the *k*-NN approach.

When exploiting multiple feature sets, a rather similar classification accuracy was obtained from the three modalities. Combining information from all the three modalities a considerable improvement in classification accuracy was obtained. When testing the developed tools on the set of data collected from 240 patients, the classification accuracy of over 98.0% was achieved. Bearing in mind the ambiguity of the classes, the obtained classification accuracy is rather encouraging.

Acknowledgement

We acknowledge the support from the agency for international science and technology development programmes in Lithuania (COST Action 2103).

References

[1] Mafee MF, Valvassori GE, Becker M. Imaging of the head and neck. Thieme; 2005.

[2] Uloza V, Saferis V, Uloziene I. Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery. Journal of Voice 2005;19(1):138–45.

[3] Rumboldt Z, Gordon L, Ackermann RBS. Imaging in head and neck cancer. Current Treatment Options in Oncology 2006;7(1):23–34.

[4] Ruffing S, Struffert T, Reith AGW. Imaging diagnostics of the pharynx and larynx. Radiologe 2005;45(9):828–36.

[5] Hasso AN, Tang T. Magnetic resonance imaging of the pharynx and larynx. Topics in Magnetic Resonance Imaging 1994;6(4):224–40.

[6] Hoorweg JJ, Kruijt RH, Heijboer RJ, Eijkemans MJ, Kerrebijn JD. Reliability of interpretation of CT examination of the larynx in patients with glottic laryngeal carcinoma. Archives of Otolaryngology-Head & Neck Surgery 2006;135(1):129–34.

[7] Rubin JS, Lee S, McGuinness J, Hore I, Hill D, Berger L. The potential role of ultrasound in differentiating solid and cystic swellings of the true vocal fold. Journal of Voice 2004;18(2):231–5.

[8] Schade G, Kothe C, Leuwer R. Sonography of the larynx—an alternative to laryngoscopy? HNO 2003;51(7):585–90.

[9] Boyanov B, Hadjitodorov S. Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. IEEE Engineering in Medicine and Biology Magazine 1997;16:74–82.

[10] Hadjitodorov S, Mitev P. A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. Medical Engineering & Physics 2002;24:419–29.

[11] Moran RJ, Reilly RB, de Chazal P, Lacy PD. Telephony-based voice pathology assessment using automated speech analysis. IEEE Transaction on Biomedical Engineering 2006;53(3):468–77.

[12] Umapathy K, Krishnan S, Parsa V, Jamieson DG. Discrimination of pathological voices using a time-frequency approach. IEEE Transaction on Biomedical Engineering 2005;52(3):421–30.

[13] Hadjitodorov S, Boyanov B, Teston B. Laryngeal pathology detection by means of class-specific neural maps. IEEE Transaction on Information Technology in Biomedicine 2000;4(1):68–73.

[14] Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transaction on Biomedical Engineering 2004;51(2):380–4.

[15] de Oliveira Rosa M, Pereira JC, Grellet M. Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Transaction on Biomedical Engineering 2000;47(1):96–104.

[16] Gelzinis A, Verikas A, Bacauskiene M. Automated speech analysis applied to laryngeal disease categorization. Computer Methods and Programs in Biomedicine 2008;91(1):36–47.

[17] Ilgner JFR, Palm C, Schutz AG, Spitzer K, Westhofen M, Lehmann TM. Colour texture analysis for quantitative laryngoscopy. Acta Oto-Laryngologica 2003;123(6):730–4.

[18] Verikas A, Gelzinis A, Bacauskiene M, Uloza V. Towards a computer-aided diagnosis system for vocal cord diseases. Artificial Intelligence in Medicine 2006;36(1):71–84.

[19] Verikas A, Gelzinis A, Valincius D, Bacauskiene M, Uloza V. Multiple feature sets based categorization of laryngeal images. Computer Methods and Programs in Biomedicine 2007;85(3):257–66.

[20] Verikas A, Gelzinis A, Bacauskiene M, Uloza V, Kaseta M. Using the patient's questionnaire data to screen laryngeal disorders. Computers in Biology and Medicine 2009;39(2):148–55.

[21] Jobson DJ, Rahaman Z, Woodell GA. Properties and performance of a center/surround retinex. IEEE Transaction on Image Processing 1997;6(3):451–62.

[22] Rahman Z, Jobson DJ, Woodell GA. Retinex processing for automatic image enhancement. Journal of Electronic Imaging 2004;13(1):100–10.

[23] Dikkers FG, Nikkels PG. Benign lesions of the vocal folds: histopathology and phonotrauma. Annals of Otology Rhinology and Laryngology 1995;104(9/1):698–703.

[24] Poels PJP, de Jong FICRS, Schutte HK. Consistency of the preoperative and intraoperative diagnosis of benign vocal fold lesions. Journal of Voice 2003;17(3):425–33.

[25] Verikas A, Gelzinis A, Bacauskiene M, Uloza V. Intelligent vocal cord image analysis for categorizing laryngeal diseases. In: Ali M, Esposito F, editors. Lecture notes in artificial intelligence, vol. 3533. Berlin/Heidelberg: Springer; 2005. p. 69–78.

[26] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transaction on System Man and Cybernetics 1973;3(6):610–21.

[27] Gelzinis A, Verikas A, Bacauskiene M. Increasing the discrimination power of the co-occurrence matrix-based features. Pattern Recognition 2007;40(9):2367–72.

[28] Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing 1975;4:172–9.

[29] Markel JD, Gray AH. Linear prediction of speech. Berlin: Springer; 1976.

[30] Manfredi C, Peretti G. A new insight into post-surgical objective voice quality evaluation. IEEE Transaction on Biomedical Engineering 2006;53:442–51.

[31] Jacobson B, Jonhson A, Grywalski C, Silbergleit A. The Voice Handicap Index (VHI): development and validation. American Journal of Speech-Language Pathology 1997;6(1):66–9.

[32] Zhu J, Hastie SRT, Tibshirani R. 1-Norm support vector machines. In: Thrun S, Saul LK, Scholkopf B, editors. Advances in neural information processing systems, vol. 16. Cambridge, MA, USA: MIT Press; 2004. p. 49–56.

[33] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge, UK: Cambridge University Press; 2004.

[34] Hernandez B, Olague G, Hammoud RI, Trujillo L, Romero E. Visual learning of texture descriptors for facial expression recognition in thermal imagery. Computer Vision and Image Understanding 2007;106(2–3):258–69.

[35] Olague G, Hammoud R, Trujillo L, Hernandez B, Romero E. Facial expression recognition in nonvisual imagery. In: Hammoud RI, editor. Augmented vision perception in infrared. London: Springer; 2009. p. 213–39.

[36] Ratsch G, Onoda T, Muller KR. Soft margins for adaboost. Machine Learning 2001;42(3):287–320.

[37] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. Pattern Recognition 2000;33(1):25–41.

[38] Verikas A, Bacauskiene M. Feature selection with neural networks. Pattern Recognition Letters 2002;23(11):1323–35.

[39] Pudil P, Novovicova J, Somol P. Feature selection toolbox software package. Pattern Recognition Letters 2002;23:487–92.

[40] Yu S, Backer SG, Scheunders P. Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery. Pattern Recognition Letters 2002;23(1–3):183–90.

[41] Zhang H, Sun G. Feature selection using tabu search method. Pattern Recognition 2002;35:701–11.

[42] Chen XW. An improved branch and bound algorithm for feature selection. Pattern Recognition Letters 2003;24(12):1925–33.

[43] Verikas A, Bacauskiene M, Valincius D, Gelzinis A. Predictor output sensitivity and feature similarity-based feature selection. Fuzzy Sets & Systems 2008;159:422–34.

[44] Mitra P, Murthy CA, Pal SK. Unsupervised feature selection using feature similarity. IEEE Transaction on Pattern Analysis Machine Intelligence 2002;24(3):301–12.

[45] Konak A, Coit DW, Smith AE. Multi-objective optimization using genetic algorithms: a tutorial. Reliability Engineering and System Safety 2006;91(9):992–1007.

[46] Leung KF, Leung FHF, Lam HK, Ling SH. Application of a modified neural fuzzy network and an improved genetic algorithm to speech recognition. Neural Computing & Applications 2007;16(4/5):419–31.

[47] Taniguchi M, Tresp V. Averaging regularized estimators. Neural Computation 1997;9:1163–78.

[48] Verikas A, Lipnickas A, Malmqvist K, Bacauskiene M, Gelzinis A. Soft combination of neural classifiers: a comparative study. Pattern Recognition Letters 1999;20:429–44.