Categorizing Normal and Pathological Voices: Automated and Perceptual Categorization

*Virgilijus Uloza, †,‡Antanas Verikas, †Marija Bacauskiene, †Adas Gelzinis, *Ruta Pribuisiene, *Marius Kaseta, and §Viktoras Saferis, *|§Kaunas, Lithuania, and †Halmstad, Sweden

Summary: Objectives. The aims of the present study were to evaluate the accuracy of an elaborated automated voice categorization system that classified voice signal samples into healthy and pathological classes and to compare it with classification accuracy that was attained by human experts.

Material and Methods. We investigated the effectiveness of 10 different feature sets in the classification of voice recordings of the sustained phonation of the vowel sound /a/ into the healthy and two pathological voice classes, and proposed a new approach to building a sequential committee of support vector machines (SVMs) for the classification. By applying “genetic search” (a search technique used to find solutions to optimization problems), we determined the optimal values of hyper-parameters of the committee and the feature sets that provided the best performance. Four experienced clinical voice specialists who evaluated the same voice recordings served as experts. The “gold standard” for classification was clinically and histologically proven diagnosis.

Results. A considerable improvement in the classification accuracy was obtained from the committee when compared with the single feature type-based classifiers. In the experimental investigations that were performed using 444 voice recordings coming from 148 subjects, three recordings from each subject, we obtained the correct classification rate (CCR) of over 92% when classifying into the healthy-pathological voice classes, and over 90% when classifying into three classes (healthy voice and two nodular or diffuse lesion voice classes). The CCR obtained from human experts was about 74% and 60%, respectively.

Conclusion. When operating under the same experimental conditions, the automated voice discrimination technique based on sequential committee of SVM was considerably more effective than the human experts.

Key Words: Voice pathology categorization–Feature selection–Genetic search–Support vector machine.

INTRODUCTION

Diagnostics of laryngeal diseases is based on a variety of instrumental endoscopic methods (video laryngoscopy and kymography; indirect and direct microlaryngoscopy) and evaluation of phonatory function. Perceptual and objective evaluation is fundamental to assess voice quality, the severity of voice deterioration, and results of therapeutic and/or phonosurgical treatment.4–3 Because of its noninvasive character and the possibility to provide quantitative data, the evaluation of vocal function using acoustic measurements of the voice signal is already commonly used in research and clinical practice.4–4 In contrast, despite the convenience and utility of these computer-based measurement devices, several practical limitations in the application of acoustic voice parameters (a rather poor relationship between acoustic measures and auditory-perceptual evaluation, difficulty to assessing aperiodic signals, and differences in signal acquisition and detection algorithms) still remain.7

Automated acoustic analysis of voice is increasingly used for the screening of laryngeal disorders.8–12 Time, frequency, and cepstral domains are usually used to extract features that characterize a voice signal. Analysis of the literature related to automated categorization of voice that aims to detect laryngeal pathologies showed that the categorization is usually based on one, two, or three types of features. There are no reports about attempts to extract a larger variety of features that characterize a voice signal. Various classifiers were used to make a decision about a voice signal represented by a feature vector. Gaussian mixture models,13,14 the linear discriminant,9 k-NN8 learning vector quantization (LVQ),10 hidden Markov models,15 a multi-layer perceptron,16 and radial basis function networks are the most popular classifiers that have been applied. In most studies, the two-class classification problem is solved, namely, a voice signal is assigned into a healthy or a pathological class. The correct classification rate (CCR) obtained in different studies, when solving the two-class classification problem, varies significantly: 80.0%,11 85.8%,17 89.1%,12 91.3%,16 and 96%.10 Because of a large variety of data sets used in the different studies, comparison of the results obtained in these studies is rather problematic.

Therefore, the present study concentrated on the investigation of the utility of a large variety of feature types in categorizing the voice signal into a healthy and two pathological voice classes using a committee of support vector machines (SVMs).18 Consequently, the aims of the present study were to evaluate the classification accuracy of the elaborated automated voice categorization system that categorizes voice signal samples into a healthy voice and two pathological voice classes, and to compare it with the classification accuracy attained by human experts.
MATERIAL AND METHODS

Voice recordings
The voice recordings came from 148 subjects, three recordings from each subject. In total, there were 79 patients representing the pathological voice classes (237 recordings) and 69 persons representing normal (healthy) voice class (207 recordings). The study group was homogenous with respect to the individuals’ age and gender (Table 1).

Thus, the mixed gender database of voice recordings used in this study contained 444 digital voice recordings of sustained phonation of the vowel sound /a/ (as in the English word “large”). Three separate voice samples obtained from each subject were recorded at the Department of Otolaryngology, Kaunas University of Medicine, Kaunas, Lithuania, in a sound-proof booth on a digitized Sony Mini Disc Recorder MDS-101 (Tokyo, Japan) through a D60S Dynamic Vocal Acoustics, Vienna, Austria) microphone placed at a 10.0-cm distance from the mouth, keeping at about 90° microphone-to-mouth angle. The typical original utterance duration before processing was about 2–5 seconds. The voice recordings were made in the “wav” file format at the rate of 44,100 samples per second. Sixteen bits were allocated for one sample. The average length of each recording was 2.4 seconds. The very first parts of the phonation sample (0.25 seconds) were cut off, and the subsequent 2.0 seconds were used for the measurements, thus minimizing the variability resulting from sampling errors. The remaining parts of the sustained vowel /a/ were discarded. This was done to ensure that the rather unstable beginning and the end of sampling had no effect on the final result.

Voice categorization
The control group, representing a “healthy voice” was composed of 69 randomly selected healthy volunteer individuals who considered their voice as normal. They had no complaints concerning their voice and no history of chronic laryngeal diseases or other long-lasting voice disorders. They also had never seen an otolaryngologist for voice problems. The voices of this group of individuals were also evaluated as healthy voice by clinical voice specialists. No pathological alterations in the larynx of the subjects of the control group were found during video laryngostroboscopy. Acoustic voice signal parameters of the control group subjects that were obtained using Dr. Speech software Voice Assessment, Version 3.0 (Tiger Electronics, Seattle, WA) were within the normal range.3

The pathological voice group represented a rather common, clinically discriminative group of laryngeal diseases, that is, mass lesions of vocal folds. Mass lesions of vocal folds included in the study consisted of nodules, polypi, cysts, papillomata, chronic hyperplastic laryngitis with keratosis, and carcinoma.19–23 Seventy-nine pathological voice group patients were recruited from the consecutive patients who were diagnosed with the laryngeal diseases mentioned above. All the patients underwent a clinical evaluation that included perceptual and acoustic voice assessments. The clinical diagnosis was based on typical clinical signs revealed during video laryngostroboscopy and direct microlaryngoscopy. All the patients underwent endolaryngeal microsurgical interventions; therefore, the final diagnosis, which later served as “gold standard,” was proven by the results of histological examination of the removed mass lesions of vocal folds. The tissue samples were obtained from all the subjects with pathological voices.

For this study, two groups of mass lesions of vocal folds were distinguished, that is, nodular lesions (localized thickenings): nodules, polypi, cysts with low malignancy potential and diffuse lesions: papillomata, chronic hyperplastic laryngitis with keratosis that presented a relatively high malignancy potential, and carcinoma.19

It was determined that among 237 recordings of the pathological voice, 147 were from the diffuse, and 90 were from the nodular classes.

Voice categorization by experts
Four experienced clinical voice specialists who worked in the field of voice pathology for at least one-third of their professional activity time and for at least 15 years served as experts/judges. They performed perceptual blind evaluation and classification of the same digitized recordings of the sustained vowel /a/ into a “healthy” and two “pathological” classes without using any additional information about the subject’s age, gender, diagnosis, etc.

Before the evaluation, the experts were instructed about the goal of the study and agreed with the experimental categorization of voice samples into three classes, that is, “healthy” and “pathological—nodular or diffuse.” An independent “training set” consisting of 30 voice recordings (10 examples represented each categorization class), which were not included in the test set was presented to the experts before the experimental evaluation of voice samples to ensure that the judges were rating to

<table>
<thead>
<tr>
<th>TABLE 1. Demographic Data of Study Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
</tr>
<tr>
<td>---------------------------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Control, N = 69</td>
</tr>
<tr>
<td>Diffuse lesions, N = 49</td>
</tr>
<tr>
<td>Nodular lesions, N = 30</td>
</tr>
<tr>
<td>P</td>
</tr>
</tbody>
</table>
Automated voice categorization

The usefulness of a large variety of feature types for categorization of a voice signal into a healthy and two pathological classes has been explored. A committee of classifiers was used to make the categorization. To find the optimal feature subsets of the various types, a genetic search procedure (genetic search or genetic algorithm (GA) is an approach to stochastic optimization) was applied.

FEATURE SETS AND FEATURE SELECTION

A large variety of classification problems can be characterized by multiple feature sets. Voice characterization is also the case. In this study, we used 10 different feature sets (the number of available features is given in the parentheses)\(^3\)^:

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 (HNR spectral) (11);
8. Harmonics-to-noise ratio in cepstral domain (HNR cepstral) (11);
9. Linear prediction (LP) coefficients (16); and
10. Linear prediction cosine transform (LPCT) coefficients (16).

It is well known that not all the features are useful for classification. Some of them can even reduce the accuracy of classification. Despite a large variety of techniques that are available for selecting variables for a single classifier, publications on feature selection for classification or regression committees are not numerous.\(^25,26\) It has been demonstrated that even simple random sampling in the feature space may be an effective technique for increasing the accuracy of classification committees.\(^27\) In Refs. 28 and 29, GAs have been used for ensemble feature selection, probably for the first time, by exploring all possible feature subsets. However, only one ensemble at a time was considered in these efforts. Kim et al\(^30\) proposed meta-evolutionary ensembles considering multiple ensembles simultaneously. When dealing with multiple feature set-based classification or prediction tasks, it is desirable to exploit as much information as possible. The analysis of the results obtained by different authors regarding variable selection for ensembles suggests that a genetic search in which a chromosome encodes an ensemble is the most promising approach. However, pure genetic search-based approaches are computationally prohibitive for large sets of variables, which is almost always the case with multiple feature sets. In this work a two-staged ensemble generation procedure was developed, to mitigate the computational burden problem.

PROCEDURE

The basic classifier used in this work was a binary classifier, the so-called SVM. Thus, to obtain the \(Q\)-class classification, \(Q(Q-1)/2\) binary classifiers \textit{one-against-one} were designed. When \(Q\) is large, the \textit{one-against-all} scheme can be applied. Given a database consisting of \(L\) (10 in our case) feature sets characterizing \(Q\) classes, the procedure to generate an ensemble for data classification into \(Q\) classes is summarized in the following steps:

1. Design a SVM using features of the \(j\)th type to separate data coming from the \(i\)th pair of classes. The optimal feature subset and the optimal number of features were determined by applying the genetic search procedure.
2. Generate \(K\) additional sets of features of the size determined in the previous step by simple random selection from the feature set of the \(j\)th type. Design \(K\) SVM classifiers to separate the \(i\)th pair of classes using the generated feature sets.
3. Present the training data to all the \(K+1\) classifiers (one classifier designed in Step 1 and \(K\) classifiers designed in Step 2) and calculate the outputs. These outputs were used as features in the second stage (Step 5).
4. Repeat Steps 1–3 for all the feature types, \(j = 1, \ldots, L\).
5. Use the outputs of the \((K+1)\ast L\) classifiers as input features to a new SVM designed to separate the \(i\)th pair of classes. The optimal feature subset and the optimal number of features are determined by applying the genetic search procedure.
6. Repeat Steps 1–5 for all the \(i = 1, \ldots, Q(Q-1)/2\) pairs of classes.
7. The committee decision was derived by aggregating decisions obtained from the \(Q(Q-1)/2\) binary classifiers.

The rationale behind using \(K\) random feature subsets (Step 2 and Step 3) was to increase the diversity of information conveyed from the first stage to the second stage.

GENETIC SEARCH

Genetic search or GA is an approach to solve optimization problems.\(^31,32\) The problem being solved is encoded in the so-called chromosome. Each chromosome consists of a string of random numbers, usually binary numbers, representing genes. The optimization problem considered in this article was maximization of the classification accuracy through selection of salient features and parameters of the classifier. A chromosome contains all the information that is needed to build an
SVM classifier. For example, when the \(i\)th gene takes the value of 1 or 0 means that the \(i\)th feature is used or not used by the classifier.

Once the encoding is defined, GA generates a random population of the probable solutions in the form of chromosomes. To encourage evolution of a better solution, the so-called genetic operations of crossover, mutation, and selection were then performed. In crossover, pairs of chromosomes (parents) are combined to create new chromosomes called offsprings. The crossover points were randomly selected, and the corresponding parts of two chromosomes selected for the crossover operation were exchanged at the selected points. The crossover operation for the two selected chromosomes is executed with the probability of crossover \(p_c\). The mutation operation usually reverses the value of one randomly selected gene—from 0 to 1 or vice versa. Each gene is selected for mutation with the probability \(p_m\).

A fitness function was used to evaluate the chromosomes. The fitness function used to evaluate the chromosomes in this work was derived from the CCR of the validation data set. The selection process of a new population was governed by the fitness values. A chromosome exhibiting a higher fitness value had a higher chance to be included in the new population. The selection probability of the \(i\)th chromosome \(p_i\) was given by

\[
p_i = \frac{r_i}{\sum_{j=1}^{M} r_j}
\]

where \(r_i\) is the CCR obtained from the classifier encoded in the \(i\)th chromosome and \(M\) is the population size.

In the reproduction process, a 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 \(p_i\) (parameter chosen by the user) or if the fitness value of the offspring is larger than that of the chromosome with the smallest fitness value. It is expected that by applying the genetic operations iteratively, a generation of good chromosomes leading to a good solution will be created.

The parameter values used in the genetic search have been found experimentally. The following values worked well in the tests: \(p_c = 0.95\), \(p_m = 0.01\), and \(p_r = 0.05\).

**RESULTS**

If we have enough data, the best approach to model assessment is to randomly divide the data set into a training set, a validation set, and a test set. The training set is used to train the models, the validation set is used to select a model, and the test set is used to assess the prediction error of the chosen model. Because data sets are often small, such position is not possible. Therefore, the leave-one-out technique is often used in the case of relatively small data sets. The leave-one-out technique gives approximately unbiased estimate of the true prediction error, but can have a rather high variance, because the training sets used in different runs are very similar.

Because we had a relatively small data set, the leave-one-out approach was used in the tests. When applying the technique, the classifier is trained using all available data points except one, which is left aside and used for testing after the training session. The experiment is repeated as many times as there are data points in the data set, each time leaving aside a different data point. The prediction error is then obtained by averaging the test results obtained in the different experiments. Regardless of the computation burden, this technique is widely used when the available data sets are small. In our tests, the data of one subject (three voice recordings) were left aside.

In the first set of experiments, a single classifier—SVM—was used for each type of features. The genetic search procedure was used to determine the optimal parameter values of the classifier and the optimal feature set.

Table 2 presents the results obtained from the tests using a single classifier for each type of features to discriminate between the two (healthy-pathological) classes. In addition to the CCR, Table 2 also presents the initial \(N\) and the selected \(N_s\), number of features. As it can be seen from Table 2, the \(HNR\) cepstral, \(Mel\) coefficients, and \(perturbation\) features provided the best performance. The CCR obtained for the three-class pairs: \(nodular\)-\(diffuse\), \(nodular\)-\(healthy\) and \(diffuse\)-\(healthy\), and the three-class (\(nodular\), \(diffuse\), and \(healthy\)) classification is shown in Table 3.

**STATISTICS**

Statistical analysis was performed using PASW Statistics 18 (SPSS Inc., Chicago, IL). The significance level of 0.05 was chosen for testing statistical hypotheses. For the evaluation of internal compatibility of experts, Cohen’s Kappa coefficient was calculated: 0.21–0.40—fair, 0.41–0.60—moderate, and 0.61–0.80—substantial agreements. The probability test for paired samples was used to reveal differences between the experts and the SVM classifier. The magnitude of the difference was considered statistically significant when type I error \(\alpha = 0.05\) and type II error \(\beta < 0.2\). The Bonferroni correction method was used for the control of the alpha error in multiple comparisons of the CCRs.

**Table 2.**

<table>
<thead>
<tr>
<th>N#</th>
<th>Type of Features</th>
<th>CCR %</th>
<th>N</th>
<th>(N_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perturbation</td>
<td>86.2</td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Frequency</td>
<td>84.2</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Mel frequency</td>
<td>84.4</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Cepstrum</td>
<td>83.1</td>
<td>100</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>Mel coefficients</td>
<td>87.3</td>
<td>35</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>Autocorrelation</td>
<td>81.8</td>
<td>80</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>HNR cepstral</td>
<td>82.4</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>HNR cepstral</td>
<td>87.8</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>LP coefficients</td>
<td>79.3</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>LPCT coefficients</td>
<td>80.7</td>
<td>17</td>
<td>6</td>
</tr>
</tbody>
</table>
In the next set of experiments, a committee was built according to the proposed design procedure. Three versions of committees with $K$ equal to 0, 1, and 2 were explored. Because we had 10 different feature types, the number of the available input features $N$ for the committee was 10, 20, and 30, depending on the $K$ value used. Note that the committee features are given by the outputs of the first stage classifiers. The results of the tests are summarized in Table 4.

As seen in Table 4, a considerable improvement in classification accuracy was obtained when SVM committees were used compared with the results obtained from single classifiers. The results indicate that the randomly selected feature sets contributed to the increase in CCR. For example, the committee formed using $K = 1$ selected 8 features from the 20 available. Among those eight, four features (HNR spectral, LP coefficients, Mel frequency, and Frequency) were obtained using the original, and four were obtained using the randomly generated feature sets (Step 2 of the procedure).

When classifying into “healthy-pathological” voice classes, the sensitivity and specificity of the committee created using $K = 3$ were 94.1% and 91.3%, respectively. There were 14 misses from 237 positives (pathological class) and 18 false alarms for 207 records from the healthy class. Figure 1 presents the true positive rate (sensitivity) as a function of the false positive rate (specificity)—receiver operating characteristic (ROC). As it can be seen from the ROC curve, detection of all the pathological cases would generate a false alarm rate close to 45%.

When they were classifying the same data set into the “healthy-pathological” voice classes, four human experts showed the following CCRs: 77.7%, 79.1%, 79.7%, and 73.2%, respectively (mean, 77.4%; standard deviation [SD], 2.94). The sensitivity and specificity obtained from the best expert classification were 82.7% and 76.3%, respectively; including 41 misses and 49 false alarms. Table 5 presents the confusion matrix obtained for the best human expert in the two-class classification problem. As can be seen from Table 5, the categorization errors are approximately evenly distributed between misses and false alarms.

Next, the CCR obtained from the best three experts was compared with the CCR obtained from the SVM committees (Table 6). As shown in Table 6, statistically significant differences among the CCR obtained from the experts and SVM committees were revealed, where the automated classification system shows on average by $13.2 \pm 2.3\%$ higher CCR, compared with that attained by the experts.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$N$</th>
<th>CCR %</th>
<th>$N_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>91.0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>92.0</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>92.6</td>
<td>13</td>
</tr>
</tbody>
</table>

In the next set of experiments, a committee was built according to the proposed design procedure. Three versions of committees, with $K$ equal to 0, 1, and 2 were explored. Because we had 10 different feature types, the number of the available input features $N$ for the committee was 10, 20, and 30, depending on the $K$ value used. Note that the committee features are given by the outputs of the first stage classifiers. The results of the tests are summarized in Table 4.

As seen in Table 4, a considerable improvement in classification accuracy was obtained when SVM committees were used compared with the results obtained from single classifiers. The results indicate that the randomly selected feature sets contributed to the increase in CCR. For example, the committee formed using $K = 1$ selected 8 features from the 20 available. Among those eight, four features (HNR spectral, LP coefficients, Mel frequency, and Frequency) were obtained using the original, and four were obtained using the randomly generated feature sets (Step 2 of the procedure).

When classifying into “healthy-pathological” voice classes, the sensitivity and specificity of the committee created using $K = 3$ were 94.1% and 91.3%, respectively. There were 14 misses from 237 positives (pathological class) and 18 false alarms for 207 records from the healthy class. Figure 1 presents the true positive rate (sensitivity) as a function of the false positive rate (specificity)—receiver operating characteristic (ROC). As it can be seen from the ROC curve, detection of all the pathological cases would generate a false alarm rate close to 45%.

When they were classifying the same data set into the “healthy-pathological” voice classes, four human experts showed the following CCRs: 77.7%, 79.1%, 79.7%, and 73.2%, respectively (mean, 77.4%; standard deviation [SD], 2.94). The sensitivity and specificity obtained from the best expert classification were 82.7% and 76.3%, respectively; including 41 misses and 49 false alarms. Table 5 presents the confusion matrix obtained for the best human expert in the two-class classification problem. As can be seen from Table 5, the categorization errors are approximately evenly distributed between misses and false alarms.

Next, the CCR obtained from the best three experts was compared with the CCR obtained from the SVM committees (Table 6). As shown in Table 6, statistically significant differences among the CCR obtained from the experts and SVM committees were revealed, where the automated classification system shows on average by $13.2 \pm 2.3\%$ higher CCR, compared with that attained by the experts.

When classifying into three voice classes ("healthy" and "nodular-diffuse" lesion voice), the CCRs obtained from the
four experts were as follows: 56.8%, 60.4%, 61.7%, and 62.4%, respectively. Table 7 presents the confusion matrix obtained for the best human expert in the three-class classification problem. There were 49 misses (29.3%), 41 false alarms (24.6%), and 77 incorrect disease categorizations (46.1%) among the errors made by the best expert.

The experts were clearly outperformed by a committee of SVMs. The highest CCR was obtained from the committee with \( K = 3 \) and was equal to 90.8%. Table 8 presents the confusion matrix obtained from the committee, when classifying into the three voice classes mentioned above. False alarms constituted 14.6% of all the errors, and misses constituted 41.5% of all the errors. The remaining errors were the result of an incorrectly determined laryngeal lesion type.

Figure 2 presents the ROCs created using the committee with \( K = 3 \). The three ROCs were created by separating one class at a time (positive) from the union of the other two (negative). As can be seen from Figure 2, the healthy class exhibits the highest separability from the other two.

Subsequently, the highest three CCRs obtained from the experts were compared with the CCRs obtained from three SVM committees created using \( K = 0, 1, \) and 2 (Table 9).

As can be seen in Table 9, when classifying into three voice classes, statistically significant differences between CCR obtained from the experts and the SVM committees were found, with the automated classification system showing on average even by 28.2 ± 3.8% higher CCR, when compared with the CCR attained by the experts.

There was no statistically significant difference (\( P<0.1 \)) between the CCR obtained from the automated classification system when classifying into two and/or three voice classes. Thus, the SVM classifier demonstrated a stable CCR result. However, the CCR obtained from the experts was reduced significantly (on the average, by 17.2 ± 2.6%), when compared with the classification accuracy into three against two voice classes.

Generally, the internal compatibility of the CCR obtained from the experts was statistically significant (\( P<0.1 \)) and reached a moderate-to-substantial classification agreement, when classifying into two voice classes (Table 10). However, the internal compatibility of the CCR obtained from the experts decreased to fair-to-moderate classification agreement when classifying into three voice classes (Table 11).

Thus, when only a sustained vowel /a/ was used as an information source, the automatic classification system was by far more precise than the human experts.

DISCUSSION

At present, the need for noninvasive methods of diagnosis has increased because of the demands of the population requiring fast, simple, and convenient clinical investigations. These methods have become possible because of the development of a technology that provides the necessary means for the acquisition and processing of biological signals. Noninvasive, objective, and automated methods of voice assessment and categorization have the potential to screen and monitor individuals with increased risk for laryngeal diseases, including laryngeal carcinoma. Therefore, qualified medical care could be provided as early and effectively as possible.
Comprehensive analysis of the voice signal including both time- and spectral-/cepstral-based measures is necessary to precisely evaluate the highly variable signals observed both between and within pathological voice types. In contrast, these measures are understandable and rather easily communicated between clinicians and researchers in clinical acoustics. However, only the combination of these measures may provide an easily applicable method for voice categorization that can be applied to the clinical diversity of voice pathologies.35

These theoretical/practical considerations were kept in mind when we were developing an automated system for voice classification based on a large variety of feature types. More specifically, the categorization of a voice signal into one healthy and two pathological (nodular or diffuse lesion) classes using a committee of SVMs was addressed.

Sustained phonation of vowel /a/ was chosen for the analysis at this initial stage of research because the steady-state phonations allow the reduction of the variances in sustained vowels and provide reliable detection and computation of acoustic features.6,12 Recordings of pathological voice representing a rather common, clinically discriminative category of laryngeal diseases—that is, mass lesions of vocal folds—were analyzed.

The experimental investigations performed in this study demonstrated that the techniques developed allowed for a significant increase in the CCR compared with using the best feature set of a single type. The CCR of over 92% was obtained when classifying into two—that is healthy-pathological—voice classes, and over 90% when classifying into three—that is, healthy and nodular-diffuse lesion voice classes. There were 43.8% of misses and 56.2% of false alarms among the errors appearing in the two-class case. The sensitivity and specificity obtained for the two-class case was 94.1% and 91.3%, respectively. In the three-class classification case, false alarms constituted 14.6% of all the errors, and misses constituted 41.5% of all the errors. The remaining errors were because of an incorrectly determined type of lesion.

The human experts were significantly less accurate than the machine. In the two-class case, the best expert showed 79.7% CCR, 82.7% sensitivity, and 76.3% specificity, respectively. As in the machine case, the errors were approximately equally distributed between misses and false alarms: 45.6% of misses and 54.4% of false alarms. In the three-class case, the best human expert showed the CCR of only 62.4%. Among the errors made by the expert, 24.6% were false alarms, 29.3% were misses, and 46.1% were incorrect disease categorizations. The three-class classification task appeared to be much more difficult for the human experts when compared with the machine. This is expected, because it has been demonstrated that human decisions are more robust when fewer categories are considered.36

There was no statistically significant difference (P<0.01) between the CCR obtained from the automated classification system classifying into two or three voice classes, thus the SVM classifier demonstrated a stable CCR irrespective of different classification conditions.

However, the CCR obtained from the human experts was about 74% when classifying into two classes, and 60% when classifying into three voice classes, respectively. Thus, the SVM-based

| TABLE 9. Comparison of the Correct Classification Rate (CCR) Obtained From the Experts and the SVM Committees When Classifying Into Three Voice Classes (Healthy and Nodular-Diffuse Lesion Voice) |
|-----------------|-----------------|-----------------|
| **Classifier**   | **CCR (%)**     | **Expert 2**    |
| Experts (1-2-3)  | 60.4            | 61.7            |
| SVM committees (K = 0-1-2) | 88.2 | 90.2 | 90.8 |
| Difference %     | 27.8            | 28.5            |
| $p_\alpha$       | 0.01            | 0.01            |
| $\beta^*$        | 0.01            | 0.01            |

* $\rho = 0.05$, when calculating $\beta$.  

TABLE 10. Compatibility of Experts’ Correct Classification Rate Performing Classification Into Three Voice Classes (Healthy and Nodular-Diffuse Lesion Voice) (Cohen’s Kappa Coefficient) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classifier</strong></td>
<td><strong>Expert 2</strong></td>
<td><strong>Expert 3</strong></td>
</tr>
<tr>
<td>Expert 1</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.60</td>
<td>0.33</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.42</td>
<td>0.33</td>
</tr>
</tbody>
</table>

* $\rho = 0.05$, when calculating $\beta$.  

TABLE 11. Compatibility of Experts’ Correct Classification Rate Performing Classification Into Three Voice Classes (Healthy and Nodular-Diffuse Lesion Voice) (Cohen’s Kappa Coefficient)
The voice classification system demonstrated, on the average, from 13.2 ± 2.3% to 28.2 ± 3.8% statistically significantly higher CCR when compared with the human voice experts.

The results of the present study are rather encouraging as they allow for obtaining a high CCR corresponding to the best two-class automated voice classification results presented in literature.\(^\text{10–12,16,17}\) Moreover, a rather high CCR obtained when classifying into three voice classes—that is, healthy and two pathological—presents a new step toward the improvement of an automated voice categorization system that has the capability for future clinical utility. As the group of diffuse lesions of vocal folds presents an increased malignization probability, a high CCR obtained by automatically discriminating this class of voice disorders allows the anticipation that future development of an automated voice categorization system will serve as a screening role in clinical applications.

To the best of our knowledge, there is a lack of data in literature on automated voice categorization that solves the problem of more than two-class (healthy-abnormal) voice classification or that uses automated voice categorization system for a specified group or single voice disorders. Therefore, the results of our study are to some extent only compatible with the 78% diagnostic accuracy of an automated, remote system for correctly identifying vocal fold paralysis as single voice disorder presented by Wormald et al.\(^\text{12}\)

Further development of the automated voice categorization system, training and “self-learning” based on exploiting data from a larger number of patients and wider range of laryngeal pathologies should allow even more precise recognition and categorization of different voice disorders, potentially including the screening for laryngeal carcinoma.

Although the system is not seen as a substitute for clinical examination, it has a potential role in screening for laryngeal diseases and for subsequent referral for clinical examination and visualization of the larynx (video laryngostroboscopy, indirect/direct microlaryngoscopy) in persons who work in high-risk industries and/or in patients whose access to otorhinolaryngological services is more restricted because they live in remote or rural areas.

The system was developed using the MATLAB (MathWorks, Natick, MA) programming language and, for example, can be easily converted into C++ and implemented in an ordinary PC for use in clinical practice.

CONCLUSIONS

We have developed a new approach to the design of a sequential committee of SVMs for multiple feature sets and genetic search-based automated discrimination of pathological voices. The proposed approach mitigates the computation burden characteristic of genetic search procedures that explore high-dimensional spaces. A considerable improvement in CCR was obtained from the committee when compared with the single feature type-based classifiers. In our series, when operating in the same experimental conditions and using only a sustained vowel /a/ as an information source, the automated voice discrimination technique was considerably better than the human experts.

Acknowledgments

An Independent Ethics Committee at Kaunas University of Medicine has approved the study.

REFERENCES