International Journal of Speech Technology (2019) 22:341–350
https://doi.org/10.1007/s10772-019-09592-y

The automatic assessment of the severity of dysphonia

Miklós Gábriel Tulics[1] · **Klára Vicsi**[1]

Received: 14 June 2018 / Accepted: 14 January 2019 / Published online: 11 March 2019

Abstract
Perceptual evaluation of the patient's voice is the most commonly used method in everyday clinical practice. We propose an automatic approach for the prediction of severity of some types of organic and functional dysphonia. By means of an unsupervised learning method, we have demonstrated that acoustic parameters measured on different phonetic classes are suitable for modelling the four grade assessments of the specialists (RBH subjective scale from 0 to 3). In this study, the overall hoarseness H was examined. Four specialists were asked to determine the severity of dysphonia. A k-means cluster analysis was performed for the decision of each specialist separately; the average accuracy of the four-grade classification was 0.46. The four-grade classification has been surprisingly close to the subjective judgements. Moreover, automatic estimation of severity of dysphonia was also determined. Linear regression and RBF kernel regression models were compared. The average rating of the four specialists were used as target in the experiments. Low RMSE and high correlation measures were obtained between the automatically predicted severity and perceptual assessments. The best RMS value of H was 0.45 for the model with RBF kernel, however, a simpler linear model provided the highest correlation value of 0.85, using only eight acoustic parameters.

Keywords Speech analysis · Pathological speech production · Interrater reliability · Regression analysis · Cluster analysis · Diagnostics

1 Introduction

Dysphonia refers to the dysfunction in the ability to produce voice. Perceptually, dysphonia can be characterized by hoarse, breathy, harsh or rough vocal qualities, but some kind of phonation remains (Hirschberg et al. 2013). Any disorder occurring in phonation affects private life as well as professional position and livelihood. Consequently, there is a need for new, cheap and effective methods that help the work of professionals in recognizing dysphonic voices and follow the development of speech therapy in an easy way. Acoustic analysis-based automatic detection of dysphonia and its severity is exactly such a research area, as it gives the possibility of non-invasive and objective quantification of pathological information, using only the speech of the patient.

In the diagnosis and management of dysphonic speech, a voice clinician typically assesses the voice quality of a patient personally. The assessment is subjective by nature. The target severity of a voice is usually defined as one clinician's assessment or as the median or average severity rating determined by a group of experienced raters assessing the voice (Chien et al. 2017; Laaridh et al. 2017). If multiple raters are recruited for the objective assessment of severity of dysphonia, the assessment is done by listening to the previously recorded voice samples. The assessment can vary among raters; thus, analysis of rating consistency is advisable. In the work of Law et al. (2012), it was found that higher intra-rater reliability was achieved with continuous speech than with sustained vowel samples. In most voice clinics, acoustic measures are derived from sustained vowel samples; however, continuous speech has several advantages over analysis of sustained vowels. It contains a variation of fundamental frequency, pauses and phonation onsets, and there is the opportunity to examine different variations of speech sounds.

✉ Miklós Gábriel Tulics
tulics@tmit.bme.hu

Klára Vicsi
vicsi@tmit.bme.hu

[1] Budapest University of Technology and Economics, Budapest, Hungary

Researchers have been focused on the development of those acoustic features that can efficiently represent the pathological condition of the speech production system. Selecting the right acoustic parameters and machine learning technics is essential for the recognition of several types of pathological voice disorders like Parkinson's disease (Benmalek et al. 2018), depression (Kiss and Vicsi 2017), dysarthria (Nidhyananthan and Shenbagalakshmi 2016), etc.

The most widely used acoustic parameters regarding dysphonia include: jitter, shimmer and Harmonics-to-Noise Ratio (HNR). Zhang and his colleagues in Zhang and Jiang (2008) found that jitter and shimmer statistically differentiate between normal and pathological sustained vowels but did not show such a significant difference between normal and pathological continuous speech. SNR, correlation dimension, and second-order entropy seemed to be capable of distinguishing normal groups from patient groups for sustained vowels and for continuous speech as well. In Wang et al. (2016) a total of 65 dimensionality measures including traditional acoustic methods, MFCC, Glottal-to-Noise Excitation methods and nonlinear dynamical analysis were measured on sustained vowels and used to compose a matrix of features. The multiclass classification results were moderately correlated with GRBAS ratings of severity, with the best accuracy around 77.55 and 80.58%, respectively.

In this study we focus on the automatic assessment of the severity of voice in cases of organic and functional dysphonia. Organic disorders include vocal cord nodules, polyps, recurrent paresis, gastroesophageal reflux disease (GERD), cyst, etc. We propose an automatic method for estimating dysphonia severity levels using read texts uttered by Hungarian patients and by a healthy control population. Our previous research has confirmed that acoustic parameters like jitter, shimmer, HNR and the first component (c1) of the mel-frequency cepstral coefficients (referred to as 'mfcc01') are useful in the automatic classification of healthy and pathological voices using continuous speech (Vicsi et al. 2011; Kazinczi et al. 2015; Grygiel et al. 2012). Moreover, in Tulics and Vicsi (2017) we demonstrated that these parameters correlate with the severity of dysphonia, as well as Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based frequency band ratios acoustic parameters measured on different phonetic classes (for example nasals, vowels, fricatives, etc.). In this research jitter, shimmer, HNR mfcc01 and frequency band ratios were used as input features. Speech defect severity was determined by 4 specialists: one of them treated the patient and directly listened to and evaluated the quality of the patient's speech during the consultations, while the other three specialists, not knowing the patient, only listened to the previously recorded sound files and determined the severity of dysphonia. The RBH scale, a four-grade subjective assessment scale from 0 to 3, where R stands for roughness, B for breathiness, H for overall hoarseness (Schönweiler et al. 2000), gave the severity of dysphonia. A four-class classification by an unsupervised learning method (k-means clustering) was used to examine whether acoustic parameters selected in our earlier research were suitable for modelling the four grade assessments of the specialists. The RBH's subjective nature was examined, as well as the consistency of the four specialists' ratings. The system proposed can be useful for clinical practice, as it is designed to provide clinical decision support.

Section 2 briefly describes the speech material used in the experiments, the measured acoustic parameters and the evaluation methodology. Our results are shown in Sect. 3, followed by the discussion and the future direction in Sect. 4.

2 Methods and materials

The system proposed in this study comprises several steps: the speech recordings of the patients are arranged into speech databases (Pathological and Healthy Adults Speech Database). The recordings are normalized and segmented on phoneme level. After selecting the phonemes to be analyzed, acoustic parameters are extracted and arranged into a feature vector. The feature vector is given to a classifier to perform the binary classification (healthy or unhealthy), or to a regression module, performing the estimation of the severity of dysphonia, in possession of prior knowledge. Prior knowledge is gained by the procession of a carefully built speech database and optimal classification and regression models. In case of a new speech sample the class (healthy/pathological) or the severity of dysphonia is unknown. The preprocessing of the speech record is the same and after the acoustic parameters are measured on phoneme level a testing feature vector is constructed that enters a comparative unit, thus the classification or regression is performed. This process is summarized in Fig. 1. This study focuses on the automatic assessment of voice severity, while analyzing the subjective nature of the specialists' ratings, too.

2.1 Pathological and healthy adults speech database

Sound samples from patients were collected during patient consultations in a consulting room at the Department of Head and Neck Surgery of the National Institute of Oncology. Several types of diseases occurred during the survey: functional dysphonia, recurrent paresis, tumors at various points of the vocal tract, gastroesophageal reflux disease, chronic inflammation of the larynx, bulbar paresis, amyotrophic lateral sclerosis, leucoplakia, spasmodic dysphonia, etc. Recordings from healthy people were collected as well. These recordings were used as comparison, and the

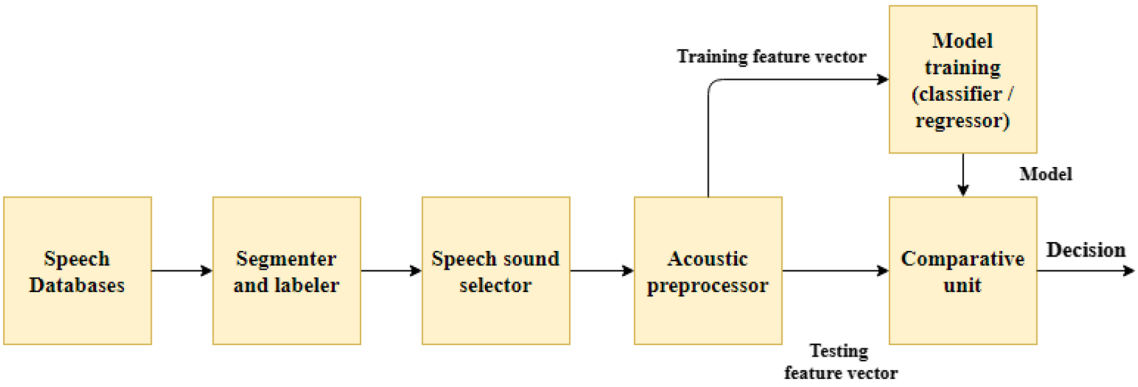

Fig. 1 The framework of this study

recordings were collected from people who had attended for unrelated check-ups.

2.2 Recording environment and text material

The recordings were made using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX outer USB sound card, with good quality A/D converter and low noise level (audio coding: PCM, sampling rate: 16 kHz, quantization: 16-bit). The recordings were made in a quiet office environment (medical office). Each patient had to read out aloud one of Aesop's Fables, "The North Wind and the Sun". This folktale is frequently used in phoniatrics as an illustration of spoken language. It has been translated into several languages, Hungarian included. The text is eight sentences long. The database was annotated and segmented on phoneme level with the help of an automatic phoneme segmentator which was developed in the Laboratory of Speech Acoustics (Kiss et al. 2013).

In the present study two datasets were used, the Initial database and the Selected database.

2.2.1 Initial database

The database containing a total of 263 speech recordings, 127 recordings from healthy subjects (62 male and 65 female) and 136 recordings from patients suffering from functional or organic dysphonia (66 male and 70 female), thus each recording is from a separate subject. The specialist who treated the patient determined the diagnosis. The specialist directly listened to and evaluated the quality of the patient's speech during the consultations. This database was used for the two-class classification experiment.

2.2.2 Selected database

The Selected database contains a total of 148 recordings, and it was used for the unsupervised cluster and regression analysis. The database contains all the 136 pathological recordings from the Initial database. Furthermore, 12 healthy recordings were selected from the Initial database, because the number of samples for each hoarseness severity category (from H0 to H3) must be balanced for the unsupervised cluster and regression analysis. Table 2 summarizes the diagnoses and their occurrences in the patient group. Four specialists examined these recordings. One of the four specialists set up the diagnosis and evaluated the quality of the patient's speech during the consultations; the other three specialists did not know the patient and only listened to the previously recorded sound files and determined the severity of dysphonia. Every rater is experienced in working with patients with voice disorders. Table 1 summarizes the diagnoses and their occurrences in the patient group.

Table 1 Diagnoses for the patient group

Diagnosis	Frequency
Benignus	2
Closure insufficiency	4
Dysarthria	2
Functional dysphonia	57
GERD	6
Healthy	12
Laryngeal paralysis	1
Laryngitis	5
Partial laryngeal surgery	1
Recurrent paresis	35
Spasmodic dysphonia	1
Tongue resection	2
Tractional stenosis	1
Tumor	12
Vocal cord alteration	1
Vocal node	4
Vocal tremor	2
Total	148

Table 2 Two-class classification results

Kernel type	Feature selection?	Hyperparameter optimalization?	Number of features	C	Gamma	Accuracy (%)
Linear	No	No	33	1	–	87
Linear	No	Yes	33	2	–	87
Linear	Yes	Yes	24	24	–	88
RBF	No	No	33	1	0.03	81
RBF	No	Yes	33	8	0.25	88
RBF	Yes	Yes	18	16	0.5	**89**

Bold values represent the best results (highest accuracies or lowest RMS values)

2.3 RBH scale

The RBH scale gives the severity of dysphonia, where R stands for roughness, B for breathiness and H for overall hoarseness. The degree of the category H cannot be less than the highest rate of the other two categories. For example, if B = 3 and R = 2, H is 3, and cannot be 2 or 1. A healthy voice's code is R0B0H0; the maximum H and respectively RBH value is 3, so a voice's code with severe dysphonia is R3B3H3. Ptok and his colleagues demonstrated that the application of the RBH scale is suitable for clinical purposes (Ptok et al. 2006). In this study the overall hoarseness H was examined.

2.4 Acoustic parameters

In Tulics and Vicsi (2017) we have done a detailed correlation analysis experiment in the shaping an extended parameter set. The following parameter set has been selected: jitter(ddp), shimmer(ddp), Harmonics-to-Noise Ratio (HNR) and mfcc01 means and standard deviations were measured on the vowel [E] (SAMPA), being the most frequent vowel read out in the folk tale. Moreover, Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based frequency band ratios were measured on the voiced parts of speech, and the measured parameter were grouped into different phonetic classes. While the quality of the continuous speech is determined not only by the quality of the vowels but also by the distortion of speech sounds of other voiced phonetic classes, like nasals, voices fricatives, etc. Therefore, these acoustic parameters which were selected by the detailed correlation analysis were used in this study.

SPI is the average ratio of energy of the speech signal in the low frequency band (70–1600 Hz) to the high frequency band (1600–4500 Hz). If the ratio is large it means the energy is concentrated in the low frequencies, indicating a softer voice (Roussel and Lobdell 2006). The parameter was calculated based on mel-frequency bands. The first band starts at 100 mel (64,95 Hz) and each band is 100 mel wide. Thus, SPI can be represented by the energy ratio of the band with the index from 1 to 13 to the bands with the index from 14 to 22.

EMD decomposes a multicomponent signal into elementary signal components called intrinsic mode functions (IMFs) (Huang et al. 1998). Each of these IMFs contributes both in amplitude and frequency towards generating the speech signal. The IMFs are arranged in a matrix in sorted order according to frequency. The first few IMFs are the high frequency components of the signal, the latter IMFs represent the lower frequency components. We calculate the entropy (E) for each IMF. The frequency band ratios of entropy were calculated the following way:

$$IMF_{entropy} = \frac{\sum_{d=1}^{2} E_d}{\sum_{d=2}^{D} E_d} \tag{1}$$

Hd is the value of Shannon entropy for each d = 1, 2,… D of the log-transformed IMFs. D is the total number of extracted IMFs. Shannon entropy for a discrete signal is defined as

$$E(p_i) = -K \sum_{i=1}^{n} p_i log p_i \tag{2}$$

where K is a positive constant. To extract the parameter, the toolkit presented in Tsanas (2013) was used.

The means and standard deviations of Soft Phonation Index (SPI) and IMFentropy were also calculated on the vowel [E], moreover SPI and IMFentropy were measured on the whole voiced parts of the speech samples and were grouped according to the following phonetic classes:

- on nasal sounds marked with [m], [n] and [J]
- on high vowels marked with [E], [e:], [i], [2] and [y]
- on low vowels marked with [O], [A:], [o] and [u]
- voiced spirants marked with [v], [z] and [Z]
- voiced plosives and affricates marked with [b], [d], [g], [dz], [dZ] and [d']

Moreover, SPI was calculated on the whole sample as well; no standard deviation was calculated here. Thus, a total

of 33 acoustic parameters were measured per patient voice sample, as starting parameter set in this research.

2.5 Decision methods

A two-class classification was performed on the Initial database using leave-one-out cross validation, with SVM (support vector machine) classifier. SVM is a supervised machine learning algorithm which is used mainly for binary classification tasks. It uses the kernel trick to transform data and based on these transformations it finds an optimal boundary between the possible outputs. The classifier was used successfully in our previous work achieving high accuracy separating the healthy and pathological voices (Kazinczi et al. 2015). The goal of the two-class classification was to find out whether the chosen acoustic parameters are rich enough in information to differentiate between healthy and pathological voices, while reducing the dimensionality of the input vector.

In order to reduce dimensionality of the input vector the forward feature selection (FFS) algorithm was used. Forward feature selection is an iterative algorithm, choosing the best feature that improves the performance in regard to a cost or objective function in each step and adding it to the already selected features. Here, the features were selected using maximum accuracy as an objective function.

It is also an important question whether the acoustic parameters selected by the correlation analysis are suitable for modelling the four grade assessments of the specialists (RBH subjective scale). For this reason, an unsupervised learning method, the k-means algorithm was used on the Selected database. The k-means is one of the simplest algorithms that uses unsupervised learning method to solve known clustering issues. This method is a fast and simple approach to the problem: it is easy to implement, and it is easy to interpret the clustering results.

The consistency of the four specialists' ratings was also examined with Cronbach's Alpha and the Intra Class Correlation Coefficient (ICC). Both methods are widely used to estimate the reliability of a composite score.

Our main aim is the automatic estimation of the severity of dysphonia. Linear regression and support vector regression (SVR) with radial basis function (RBF) kernel was used for model building. By its nature, linear regression only looks at linear relationships between dependent and independent variables; linear regression also assumes that there is a straight-line relationship between the input variables and the target. SVR with RBF kernel has good generalization and strong tolerance to input noise.

3 Results

3.1 Two-class classification results

Classification experiments were made using several combinations. Liner and RBF kernels were also tried out. The default value of C of support vector machine is 1, while Gamma is 1/number of features. In order to choose the optimal hyperparameters for the SVM classifier grid search was used. Leave-one-out cross validation was used in all cases. Results are summarized in Table 2.

As Table 2 suggests, the highest accuracy of 89% was reached by using RBF kernel. The features selection algorithm reduced the input dimensionality to 18 acoustic parameters, while achieving higher accuracy than the default setting. The acoustic parameters selected by the FFS algorithm are the following: $\text{jitter}_{\text{mean}}$, $\text{jitter}_{\text{std}}$, $\text{shimmer}_{\text{mean}}$, $\text{shimmer}_{\text{std}}$, hnr_{mean}, hnr_{std}, $\text{mfcc01}_{\text{mean}}$, $\text{mfcc01}_{\text{std}}$, $\text{SPI} \rightarrow \text{E}_{\text{std}}$, $\text{SPI} \rightarrow \text{Nasal}_{\text{mean}}$, $\text{SPI} \rightarrow \text{Nasal}_{\text{std}}$, $\text{SPI} \rightarrow \text{LowVowels}_{\text{std}}$, $\text{SPI} \rightarrow \text{VoicedSpirants}_{\text{mean}}$, $\text{SPI} \rightarrow \text{VoicedSpirants}_{\text{std}}$, $\text{IMF} \rightarrow \text{E}_{\text{std}}$, $\text{IMF} \rightarrow \text{Nasal}_{\text{mean}}$, $\text{IMF} \rightarrow \text{VoicedPlosives}_{\text{mean}}$, $\text{IMF} \rightarrow \text{VoicedPlosives}_{\text{std}}$. These parameters are referred to as '18 parameter set' in further experiments.

3.2 Unsupervised cluster analysis

It is an interesting question whether the chosen acoustic parameters (the 18-parameter set) can model the individual assessments. Cluster analysis is used to classify cases into relative groups called clusters, in this case: individual assessments of severity of dysphonia. In cluster analysis, there is no prior information about the cluster membership for any of the data. If the acoustic parameter set and the unsupervised learning method are fixed, it is possible to compare four cluster models for each case labelled by a specialist's judgement. In order to examine the subjective nature of RBH k-means cluster analysis was done.

The confusion matrices for each specialist are shown separately in Tables 3, 4, 5 and 6. The accuracies for the

Table 3 Confusion matrix in case of Specialist 1

	Predicted label				Sum
	0	1	2	3	
Specialist 1 (true label of H)					
0	**12**	13	9	2	36
1	1	**33**	25	4	63
2	2	5	**10**	6	23
3	1	3	5	**17**	26
Sum	16	54	49	29	148

Bold values represent diagonal values in the confusion matrices

Table 4 Confusion matrix in case of Specialist 2

	Predicted label				Sum
	0	1	2	3	
Specialist 2 (true label of H)					
0	**11**	6	5	0	22
1	3	**24**	26	2	55
2	2	16	**23**	15	56
3	0	3	0	**12**	15
Sum	16	49	54	29	148

Bold values represent diagonal values in the confusion matrices

Table 5 Confusion matrix in case of Specialist 3

	Predicted label				Sum
	0	1	2	3	
Specialist 3 (true label of H)					
0	**11**	2	3	0	16
1	2	**20**	15	1	38
2	2	25	**16**	9	52
3	1	7	15	**19**	42
Sum	16	54	49	29	148

Bold values represent diagonal values in the confusion matrices

Table 6 Confusion matrix in case of Specialist 4

	Predicted label				Sum
	0	1	2	3	
Specialist 4 (true label of H)					
0	**12**	7	6	0	25
1	2	**24**	18	6	50
2	2	18	**17**	6	43
3	0	5	8	**17**	30
Sum	16	54	49	29	148

Bold values represent diagonal values in the confusion matrices

decision in case of each specialist in order is: 0.49, 0.44, 0.45, 0.47. The average classification accuracy is 0.46.

From this experiment we can conclude that the acoustic parameter set is suitable for modelling the individual assessments of dysphonia severity. Looking at the individual confusion matrices, the following observations can be made. If the clustering process was not able to accurately determine the specialist's assessment, it was classified into the adjacent cluster. The separation between healthy (H0) and unhealthy (H1, H2 and H3) is satisfactory. While Specialist 1 rated the voices less severe, Specialist 4 rated more voices H3. The results also show that the H1 and H2 classes are the least distinct from the clustering point, in the cases of all four specialists.

Table 7 Average deviations from the average of H for individual specialists

	Sum	Specialist 1	Specialist 2	Specialist 3	Specialist 4
Average deviation from average of H	0.4	0.5	0.3	0.4	0.3

Table 8 Item reliability statistics if one rater is removed

	If item dropped	
	Item-rest correlation	Cronbach's α
Specialist 1	0.714	0.881
Specialist 2	0.777	0.857
Specialist 3	0.782	0.852
Specialist 4	0.787	0.850

This demonstrates that in this case, a continuous scale prediction process, such as regression, would better approximate the subjective assessments than clustering with disaggregated sets. In this way, we can get a more accurate system, with fewer errors.

3.3 Reliability analysis

Perceptual evaluation of voice is the most commonly used tool in everyday clinical practice when assessing voice disorders. Perceptual evaluation plays a crucial role in both clinical outcome measures and our own task. In this section we analyze if the rater reliability is satisfying enough; moreover, if any difference can be observed between the clinician who was present during patient consultations and the three specialists who determined the severity of dysphonia while listening to the recordings.

According to Table 7, on average, the specialists gave a difference of 0.4 to the average. Specialists 2 and 4 gave the most average ratings, so their assessments will change the average H value the least; while Specialist 1's assessments differ the most from the average. It should be noted, however, that only Specialist 1 was present during the sound recordings. Having a different acoustic experience could have significantly influenced his assessment.

For measuring internal consistency ("reliability") Cronbach's Alpha was used. It has been proposed that Cronbach's alpha values of 0.80 or above indicate a high level of internal consistency. In our case, Cronbach's alpha was found to be 0.891, which indicates a high level of reliability. The removal of any expert assessment would not result in a higher alpha value. It is to be noted, though, that removal

of Specialist 1 would ruin the internal consistency the least. The Item Reliability Statistics are shown in Table 8.

The Intra class correlation coefficient (ICC) was also calculated. An ICC of 0.75 or above indicates good reliability. In our case, a high degree of reliability was found between the severity judgements. The average measure ICC was 0.890 with a 95% confidence interval from 0.857 to 0.917 ($F(137;411) = 9.172$; $p < 0.001$).

3.4 Regression analysis

Regression has a significant advantage compared with cluster analysis, since the prediction follows a function almost continuously. This property can significantly improve the quality of the model. Due to the small sample size, leave-one-out cross validation was used. The performance of the regression methods is evaluated by the root mean square error (RMSE) value, the linear relationship between the target and the predicted H scores is described by Pearson correlation. To find the optimal hyperparameters grid search was used.

In this analysis linear regression and support vector regression (SVR) with radial basis function (RBF) kernel were used. To reach the best performance the 18-parameter set and the result of the FFS algorithm was used, for linear regression and SVR with RBF kernel separately.

Table 9 summarizes the results. The mean of the four specialists' was used as target. The FFS algorithm reduced the original 33-dimension input to only eight parameters using linear kernel. The following parameters were selected: $mfcc01_{mean}$, $shimmer_{mean}$, hnr_{std}, $SPI \rightarrow HighVowels_{mean}$, $SPI \rightarrow LowVowels_{std}$, $SPI \rightarrow VoicedPlosives_{std}$, $IMF \rightarrow Nasal_{mean}$, $IMF \rightarrow LowVowels_{std}$. This configuration gave the highest 0.853 correlation. When RBF kernel was used, the FSS algorithm selected 14 parameters, these were the following: $mfcc01_{mean}$, $shimmer_{mean}$, hnr_{mean}, hnr_{std}, $SPI \rightarrow E_{mean}$, $SPI \rightarrow E_{std}$, $SPI \rightarrow Nasal_{std}$, $SPI \rightarrow HighVowels_{mean}$, $SPI \rightarrow LowVowels_{mean}$, $SPI \rightarrow LowVowels_{std}$, $SPI \rightarrow VoicedPlosives_{mean}$, $IMF \rightarrow Nasal_{mean}$, $IMF \rightarrow VoicedPlosives_{mean}$, $IMF \rightarrow VoicedPlosives_{std}$. The lowest RMSE value of 0.454 was obtained here. Furthermore, the FFS models gave only slightly better results than the models with the 18-parameter set, which demonstrates the generalizing ability of the acoustic parameters employed.

Figure 2 depicts the automatically predicted severity of the dysphonia compared to the reference perceptual assessment of speaker severity. The figure shows the linear regression models created by the result of the FFS algorithm.

The figure illustrates once again the capacity of the proposed approach in predicting the severity of dysphonia regardless of the speaker's pathology or severity degree. It can be observed that the model gives good prediction of severity of H1.

Figure 3 shows the distribution of the predicted H values from the linear regression model and the mean of the four specialist's ratings. The means are the same, but the predicted values have a lower standard deviation.

Table 9 Regression analysis results—the mean of the four specialist's ratings as target

Acoustic parameter set	Type of regression	Correlation	RMSE	Hyperparameters
18-Parameter set	Linear	0.837	0.502	C = 1
Result of FFS, 8-parameter set	Linear	**0.853**	**0.462**	C = 1
18-Parameter set	RBF kernel	0.808	0.506	C = 2, gamma = 0.125
Result of FFS, 14-parameter set	RBF kernel	**0.849**	**0.454**	C = 4, gamma = 0.25

Bold values represent the best results (highest accuracies or lowest RMS values)

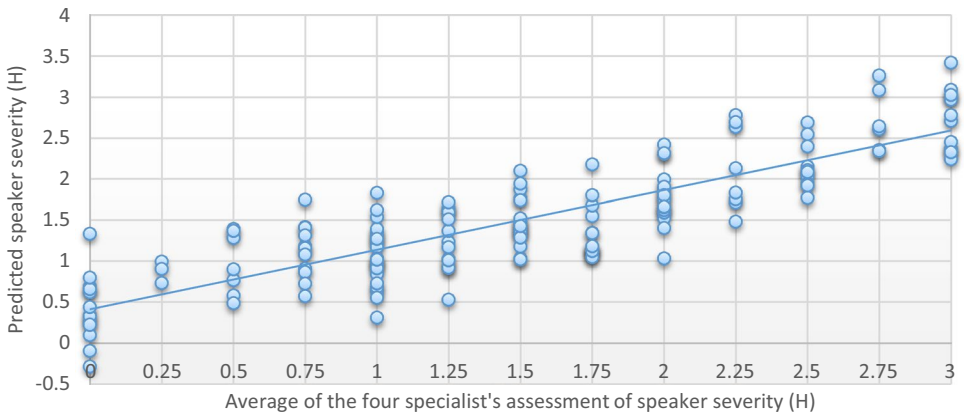

Fig. 2 Automatically predicted dysphonia severity degree according to perceptual assessment of H, using linear regression with 8 parameters

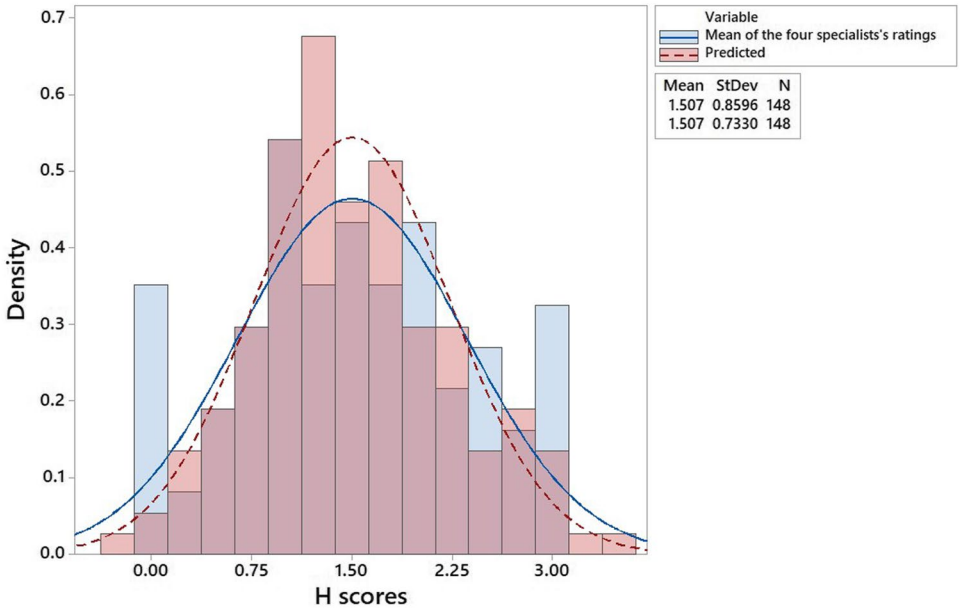

Fig. 3 Histogram of the mean of the four specialist's ratings and the predicted H scores

4 Conclusion and discussion

This paper investigates an automatic approach for the prediction of severity of dysphonia based on acoustic parameters measured on different phonetic classes. In these phonetic classes the type of excitation is different, thus several phonation problems may appear differently in the phonetic classes. This research was based on continuous speech, since it is more applicable to practical work than sustained vowels (Kim et al. 2015).

Four specialists evaluated all the speech samples perceptually: the specialist who heard the patient live and set up the diagnosis (Specialist 1), and three specialists who determined the severity of dysphonia while listening to the recordings (Specialists 2, 3 and 4). The specialists were asked to rate all the speakers by the RBH scale; in this study the overall hoarseness H was examined.

Generally speaking, in statistical model building methods, where only a limited number of samples are available, it is very important to choose the optimal set of the characteristic parameters, thus, to minimalize the noise entering the system. This is the case for modeling the assessment of pathological speech, where the database collection is very difficult. For this reason, the optimization of the set of the input parameters is necessary for the construction of models for the automatic assessment of the severity of dysphonia. When creating a model overfitting may also be problem, which happens when a model learns the detail and noise in the training data. This can have a negative impact on performance when the model is tested with new data. Overfitting happens more likely with nonparametric and nonlinear models as they have more flexibility when learning a target function. One should try to generalize the model as much as possible.

Based on our previous works (Kazinczi et al. 2015; Tulics and Vicsi 2017), a total of 33 acoustic parameters were selected and measured per patient voice sample, including jitter, shimmer, HNR and the first component (c1) of the mel-frequency cepstral coefficients measured on vowel [E], and Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based frequency band ratios on different phonetic classes. We can ask the question whether these acoustic parameters are really suitable for modelling the assessments of the specialists. To answer this question three different experiments were made.

4.1 Two-class classification and parameter selection

The goal of the two-class classification was to find out whether the chosen acoustic parameters are suitable for the automatic separation of healthy and pathological classes. With the help of FFS feature selection algorithm, the dimensionality of the input vector was reduced to 18 parameters reaching accuracy of 89% between the two classes.

4.2 Clustering

With k-means unsupervised learning method, four cluster models were compared, each case labelled by a specialist's assessments using the four grade RBH subjective scale. The 18-parameter set, selected in the two-class classification

experiment was used to model the four grade individual assessments of dysphonia severity. The accuracies for each specialist's model in order is: 0.49, 0.44, 0.45, 0.47.

RBH perceptual evaluation of experts was the bases for our classification and regression models. Thus, it was important to analyze if the rater reliability of the 4 experts is consistent enough. For measuring internal consistency, ("reliability") Cronbach's Alpha and Intra Class Correlation Coefficient (ICC) method was used. Through the analysis, we have found that Specialist 1 rated more voices less severe than the other three specialists. One explanation of these phenomena can be that Specialist 1 was personally involved in the patient diagnosis and therapy; moreover, the specialist had a different acoustic experience than the other jury members. Despite the interesting differences among the decision of the specialists, a high degree of reliability (Cronbach's alpha = 0.891, ICC = 0.890) was measured between their severity judgements when measuring internal consistency. Despite its significance, we have not encountered such consistency testing so far in the scientific literature.

4.3 Regression

The four-class clustering results show that the H1 and H2 classes are the least distinct from the clustering point, for all four specialists. Our hypothesis was that in the case of a continuous scale prediction process, such as regression, the subjective assessments would be better approximated than in the case of clustering with disaggregated sets. In this way we can get a more accurate system with fewer errors. A method for dysphonia severity assessment has been presented, which is a regressor that uses acoustic parameters measured on different phonetic classes. The best RMS value of H was 0.45 for the model using RBF kernel, where the feature selection algorithm selected 14 parameters. A simpler linear model has provided the highest correlation value of 0.85, using only eight acoustic parameters. The result of the FFS shows the importance of those acoustic parameters, which were measured on different phonetic classes. The distribution of the predicted values of the linear model is very similar to the H values' mean given by the specialists, resulting in the same means, but the predicted values have a lower standard deviation. Using a linear model reduces the probability of overfitting, as a linear model using only a few parameters has less flexibility when learning a target function. It is also important to mention that the regression model using the 18-parameter set provides similar correlation and RMSE values as the linear regression model using only 8 acoustic parameters. This demonstrates the generalizing ability of the used acoustic parameters.

The end system proposed in this study can help young physicians or general practitioners filter out patients with dysphonia more efficiently and automatically determine the severity of dysphonia. Future work includes involving more specialists in the severity assessments, and the evaluation of our system on a larger dataset. Based on a larger dataset, the classification of the different types of dysphonia would also be possible. We also believe that the results are generalizable to other languages.

Acknowledgements Open access funding provided by Budapest University of Technology and Economics (BME). We would like to thank Krisztina Mészáros from the Department of Head and Neck Surgery of the National Institute of Oncology for her continued cooperation in helping us collect and evaluate the patient data, which is the basis of our research. We would also like to thank Tamás Hacki, György Smehák and Nóra Damásdi for the evaluation of the sound recordings.

References

Benmalek, E., Elmhamdi, J., & Jilbab, A. (2018). Multiclass classification of Parkinson's disease using cepstral analysis. *International Journal of Speech Technology, 21*(1), 39–49.

Chien, Y. R., Borskỳ, M., & Guðnason, J. 2017. Objective severity assessment from disordered voice using estimated glottal airflow. In *Proceedings of the Interspeech* 2017 (pp. 304–308).

Grygiel, J., Strumołło, P., & Niebudek-Bogusz, E. (2012). Application of mel cepstral representation of voice recordings for diagnosing vocal disorders. *Delta, 12*, 2.

Hirschberg, J., Hacki, T., & Mészáros, K. 2013. Foniátria és társtudományok: A hangképzés, a beszéd és a nyelv, a hallás és a nyelés élettana, kórtana, diag-nosztikája és terápiája (I. kötet).

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., & Liu, H. H. 1998. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences, The Royal Society* (pp. 903–995).

Kazinczi, F., Mészáros, K., & Vicsi, K. 2015. Automatic detection of voice disorders. In: *Proceedings of the International Conference on Statistical, Language and Speech Processing* (pp. 143–152). Springer.

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language, 29*, 132–144.

Kiss, G., Sztaho, D., & Vicsi, K. 2013. Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features. In: *Proceedings of the 2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 579–582). IEEE.

Kiss, G., & Vicsi, K. (2017). Mono-and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology, 20*(4), 919–935.

Laaridh, I., Kheder, W. B., Fredouille, C., & Meunier, C. 2017. Automatic prediction of speech evaluation metrics for dysarthric speech. In: *Proceedings of the Interspeech 2017* (pp. 1834–1838).

Law, T., Kim, J. H., Lee, K. Y., Tang, E. C., Lam, J. H., van Hasselt, A. C., & Tong, M. C. (2012). Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *Journal of Voice, 26*, 666–613.

Nidhyananthan, S. S., & Shenbagalakshmi, V. (2016). Assessment of dysarthric speech using Elman back propagation network (recurrent network) for speech recognition. *International Journal of Speech Technology, 19*(3), 577–583.

Ptok, M., Schwemmle, C., Iven, C., Jessen, M., & Nawka, T. (2006). On the auditory evaluation of voice quality. *HNO, 54*, 793–802.

Roussel, N. C., Lobdell, M. (2006). The clinical utility of the soft phonation index. Clinical Linguistics & Phonetics, *20*, 181–186.

Schönweiler, R., Hess, M., Wübbelt, P., & Ptok, M. (2006). Novel approach to acoustical voice analysis using artificial neural networks. *JARO-Journal of the Association for Research in Otolaryngology, 1*, 270–282.

Tsanas, A. (2013). Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. *Models and Analysis of Vocal Emissions for Biomedical Applications, 2*, 37–40.

Tulics, M. G., & Vicsi, K. 2017. Phonetic-class based correlation analysis for severity of dysphonia. In: *Proceedings of the 2017 8th IEEE Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 21–26). IEEE.

Vicsi, K., Imre, V., & Mészáros, K. 2011. Voice disorder detection on the basis of continuous speech. In: *Proceedings of the 5th European Conference of the International Federation for Medical and Biological Engineering* (pp. 86–89). Springer.

Wang, Z., Yu, P., Yan, N., Wang, L., & Ng, M. L. (2016). Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the grbas scale. *Journal of Signal Processing Systems, 82*, 241–251.

Zhang, Y., & Jiang, J. J. (2008). Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice, 22*, 1–9.