Deep Learning based classification of vocal folds' vibration dynamics

Mona Kirstin Fehling[1,2,3], Maximilian Linxweiler[2], Bernhard Schick[2], and Jörg Lohscheller[1]

[1]Department of Computer Science, University of Applied Sciences Trier, Schneidershof, Trier, Germany
[2]Department of Otorhinolaryngology, Head and Neck Surgery, Saarland University Hospital, Homburg/Saar, Germany
[3]Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Mannheim / Medical Faculty Mannheim of Heidelberg University, Mannheim, Germany

Abstract. Vocal fold (VF) dynamics can be captured in real-time using high-speed videolaryngoscopy, laying the basis for quantitative assessment of the VFs vibration properties. A compact representation of the vibrational behavior as captured in these high-speed videos (HSV) is provided by the so-called Phonovibrogram (PVG). The PVG encodes the VFs vibrational behavior by characteristic spatial and temporal patterns in a three-dimensional representation. Based on these characteristic PVG patterns, this work realizes a fully automatic classification of different voice disorders. For this purpose, a Convolutional Neural Network (CNN) was trained and evaluated using a stratified 10-fold cross-validation strategy on PVGs from 220 subjects to solve two different classification tasks: (a) Classification of the vibrational behavior as *physiologic* or *pathologic* and (b) classification of the PVGs according to the subjects actual clinical diagnosis as *healthy*, *muscle tension dysphonia (MTD)*, *paresis*, or *polyp*. The trained CNN distinguished with an average classification accuracy of 0.82 ± 0.07 between *physiologic* and *pathologic* VF vibration (sensitivity: 0.81 ± 0.12, specificity: 0.82 ± 0.12) and achieved an average classification accuracy of 0.85 ± 0.07 across all classes (sensitivity: 0.71 ± 0.19, specificity: 0.91 ± 0.07) for classification according to the clinical diagnoses. Based on the PVG representation, the presented approach reliably differentiates between physiologic and pathologic VF vibration and is even eligible to distinguish types of voice disorders without user interaction. However, to further increase the method's performance, a larger amount of training data is required.

Keywords: vocal fold vibration, voice disorders, high-speed video, Phonovibrogram, classification, deep neural network

1 Introduction

Voice disorders emerge from disturbances in the vocal folds (VFs) vibrational behavior [1]. At any time, about 7.6% of the adult population in the USA is affected by any sort of voice disorder, with a lifetime prevalence of voice disorders being 30% [2, 3].
Quantitative assessment of the VFs' vibrational behavior can be done in real time using high-speed videolaryngoscopy [4, 5]. High-speed videolaryngoscopy captures the vibrating VFs at frame rates of 2,000 - 20,000 fps [6], allowing for detection of even slight variations in VF vibration. However, the huge amount of recorded data makes evaluation of the captured high-speed videos (HSVs) time-consuming and thus challenging during voice assessment in clinical routine [7]. Moreover, visual assessment and rating of the HSVs, demand an experienced clinician to make a proper diagnosis [8]. Therefore, various approaches were presented to assist the clinicians with voice assessment providing a compact and clinically meaningful representation of the relevant information on VF vibration as contained in the HSVs [9–16].
One of the so far most comprehensive approache is the so-called Phonovibrogram (PVG). The PVG provides a compact and clinically meaningful three-dimensional representation of the VFs vibrational behavior contained in an HSV sequence [11]. Figure 1(a) shows the construction process of the PVG; a detailed description of the PVG construction can be found in Lohscheller et al. [11]. To build the PVG, the glottal area is initially segmented in the subsequent HSV frames. Based on this segmentation result, the glottal symmetry axis is defined between the posterior (P) and the anterior (A) end of the glottis. Afterwards, the glottal symmetry axis is used to split the retrieved glottis contour into two halves representing the left and the right VF edge. That following, the distances between the contours and the glottal symmetry axis are computed for all positions along the glottal symmetry axis. Color-coding the computed distances results into one color stripe per frame. The PVG visualization is finally built by temporal concatenating all these color stripes. Spatio-temporal patterns in the PVG encode the VFs vibrational behavior and show characteristic patterns for physiologic as well as pathologic VF vibrations. Exemplarily, PVGs from a healthy subject, a subject with muscle tension dysphonia (MTD), a subject with unilateral VF paresis, and a subject diagnosed with a VF polyp are shown in Figure 1(b).
PVG-based classification of voice disorders was so far done by training Support Vector Machines (SVMs). Suitable features were retrieved from the PVG and partly supplemented with additional features from the acoustical

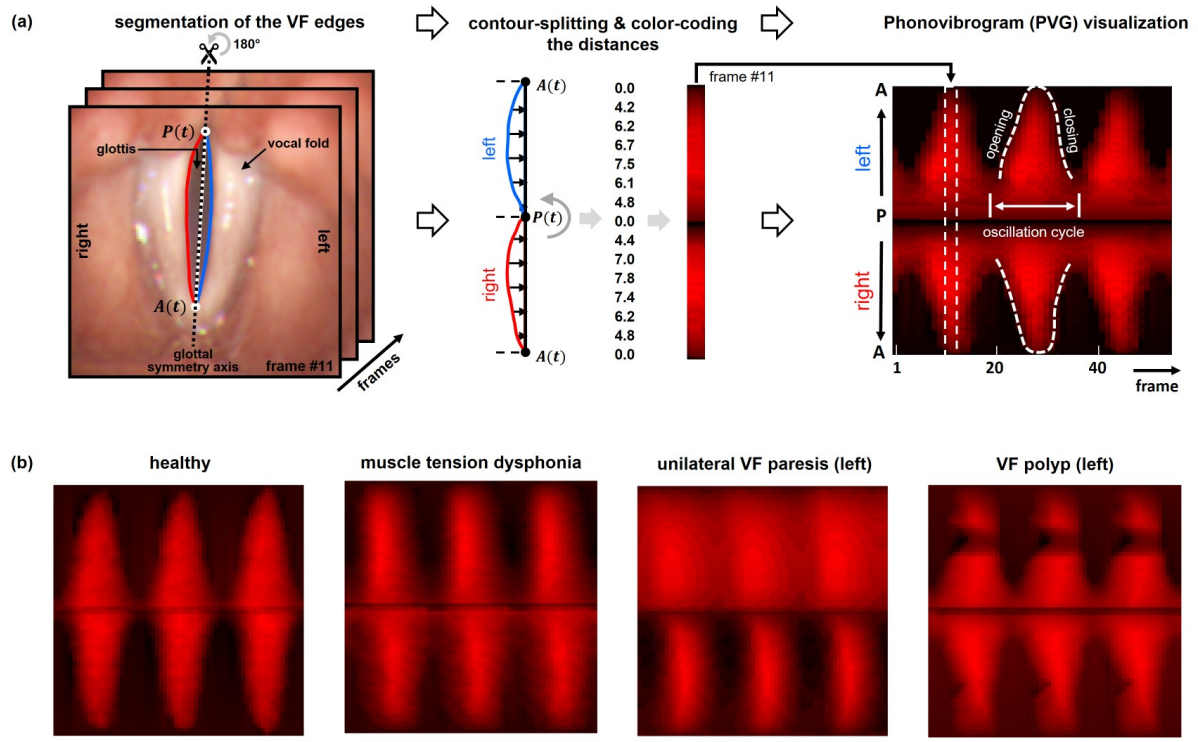


Fig. 1: Phonovibrogram (PVG) visualization of VF vibration. (a) Computation of the PVG. The VF edges are retrieved from each HSV frame based on the segmented glottis. Then, the contour of the left VF edge is rotated at 180°, followed by computation of the distances between the respective VF edge and the glottal symmetry axis and color-coding them. Concatenation of the resulting color stripes results in the PVG representation of the VFs spatiotemporal vibrational behavior. (b) Examples of PVGs computed from clinical HSVs: Healthy subject, subject with muscle tension dysphonia (MTD), subject with unilateral VF paresis (left VF), and subject with a polyp on the left VF.

voice signal.
By employing such an SVM-based classification approach, Voigt et al. achieved an average classification accuracy of 0.744 on a 2-class discrimination task by retrieving features solely from the PVG (healthy vs. muscle tension dysphonia patients) [17].
In another work, Voigt et al. achieved an average classification accuracy of 0.93 on a SVM-based classification of PVGs as either healthy or paralytic [18].
Unger et al., on the other hand, achieved an accuracy of 0.69 ± 0.02 on a four-class classification task (healthy vs. MTD vs. paresis vs. polyp) by training an SVM on a 12-dimensional feature set [8]. In that study, the PVG-based feature set was retrieved by using a combined wavelet- and PCA-based approach for dimensionality reduction on the initially 1000 frames long PVG sequences. The features retrieved from the PVG describe the glottal closure type, the phase information, the asymmetry, and the irregularity of VF vibration.
In this work, a fully automatic classification of voice disorders based on characteristic PVG patterns is presented, which for the first time is realized by using a deep Convolutional Neural Network (CNN).

2 Material and Method

Figure 2 provides an overview of the conducted study by showing the dataset used, the employed CNN architecture, and the performed training and evaluation procedure. Details on the clinical data and the trained CNN are provided in the following.

2.1 Data

Clinical data were collected from the *Department of Otorhinolaryngology, Head and Neck Surgery* at the *Saarland University Medical Center* (Homburg/Saar, Germany) within a previous study [8]. Ethical approval was obtained from the local ethics committee (*Ethik-Kommission bei der Ärztekammer des Saarlandes*, reference number:

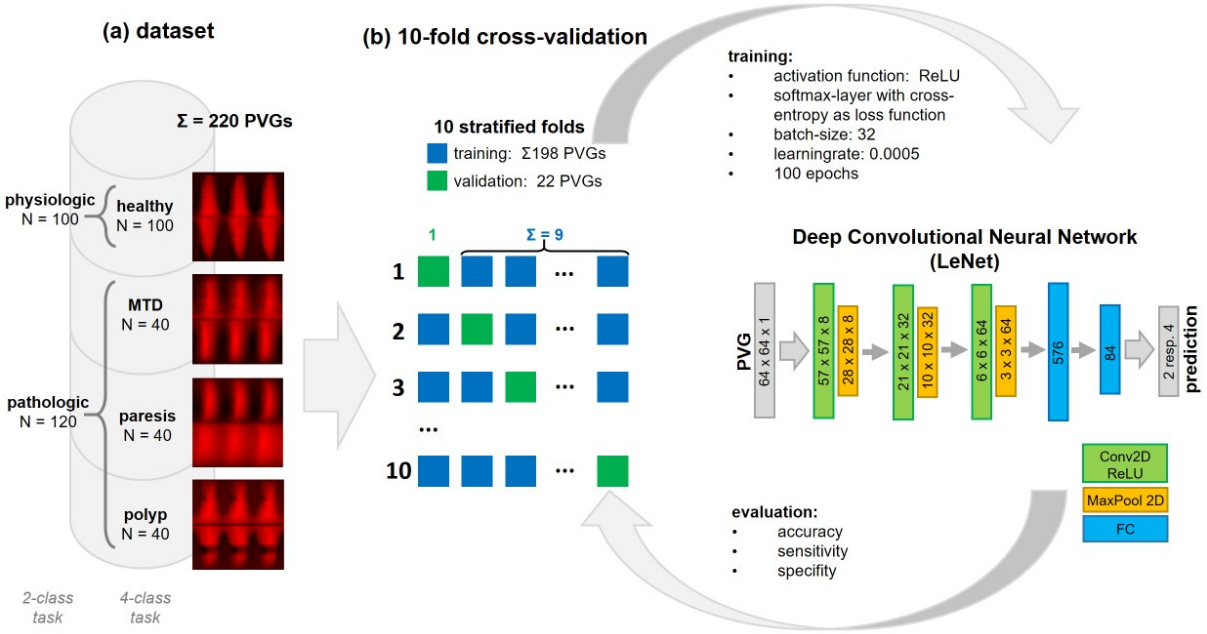


Fig. 2: Overview on the conducted study. (a) Composition of the dataset used, and (b) schematic representation of training and evaluation procedure. The PVGs were classified with a CNN (based on the LeNet) using a stratified 10-fold cross-validation strategy.

103/12), and the participants gave their consent prior to participation. Laryngeal HSVs were recorded with a 90° tip and in color using the rigid endoscopy system *HRES ENDOCAM 5562* from *Richard Wolf GmbH* (Knittlingen, Germany). The HSVs were captured with a spatial resolution of 256 × 256 px and at a frame rate of 4,000 fps. All participants were asked to perform a sustained phonation of the vowel /ae/ at comfortable pitch and loudness for at least 1 s during the examination.
In total, PVGs from clinical HSV recordings acquired from 220 individual subjects comprising four clinical groups *healthy*, *muscle tension dysphonia (MTD)*, unilateral VF *paresis*, and unilateral VF *polyp* were included in the dataset. The dataset comprises $N_{healthy} = 100$ recordings from healthy subjects, $N_{MTD} = 40$ recordings from subjects diagnosed with muscle tension dysphonia, $N_{paresis} = 40$ subjects with paresis, and $N_{polyp} = 40$ subjects with polyp.
To avoid any bias caused by the subjects' individual fundamental frequencies, three complete oscillation cycles were extracted from each PVG. The VF-deflections encoded in each PVG extract were normalized to the interval $[0; 1]$ and the PVGs were futher rescaled to $64px \times 64px$.

2.2 Classification Tasks

A deep CNN architecture is used to classify the PVGs. This study investigated two different classification scenarios: (a) First, a CNN was trained to classify the VFs vibrational behavior as represented by the PVG as either *physiologic* or *pathologic* (2-class task). For differentiation between physiologic and pathologic VF vibration, the patients diagnosed with MTD, paresis, and polyp were combined to the class *pathologic* ($N_{pat} = 120$), while the healthy subjects compose the class *physiologic* ($N_{phys} = 100$). (b) In a second step, a CNN was analogously trained to classify the PVGs according to their actual clinical diagnosis as *healthy*, *MTD*, *paresis*, or *polyp* (4-class task).

2.3 Architecture

This study used a CNN based on the LeNet-architecture [19], a comparatively small and elementary CNN architecture for computer vision tasks. Contrary to the original LeNet, we used as input single PVGs of size $64px \times 64px$, ReLU activations instead of tanh and sigmoid activations [20]. Moreover, dropout layers were used to avoid overfitting and improve classification performance despite the small dataset [21].
As depicted in Fig. 3, the CNN used here consists of seven layers in total: the input layer, three convolutional layers, two fully connected layers, and an output layer. These layers build two functional parts: a feature extractor

and a classifier.
The feature extractor is built from the convolutional layers, where each convolutional layer consists of convolutions with varying kernel sizes and dropout, ReLU as a nonlinear activation function, and 2D-max-pooling. The extracted features are then classified by the classifier that is composed of two fully connected layers which are activated with ReLU and a final linear softmax layer that returns the class probabilities for the respective number of classes as $out\_features$. The properties of the consecutive CNN layers are as follows (Fig. 3):

```
LeNetOpt64DropOut(
  (feature_extractor): Sequential(
    (0): Conv2d(1, 8, kernel_size=(8, 8), stride=(1, 1))
    (1): Dropout(p=0.2, inplace=False)
    (2): ReLU()
    (3): MaxPool2d(kernel_size=2, stride=2, padding=0,
         dilation=1, ceil_mode=False)
    (4): Conv2d(8, 32, kernel_size=(8, 8), stride=(1, 1))
    (5): Dropout(p=0.2, inplace=False)
    (6): ReLU()
    (7): MaxPool2d(kernel_size=2, stride=2, padding=0,
         dilation=1, ceil_mode=False)
    (8): Conv2d(32, 64, kernel_size=(5, 5), stride=(1, 1))
    (9): Dropout(p=0.2, inplace=False)
    (10): ReLU()
    (11): MaxPool2d(kernel_size=2, stride=2, padding=0,
          dilation=1, ceil_mode=False)
  )
  (classifier): Sequential(
    (0): Linear(in_features=576, out_features=84, bias=True)
    (1): ReLU()
    (2): Linear(in_features=84, out_features= {2 or 4}, bias=True)
  )
)
```

Fig. 3: Properties of the consecutive CNN layers.

2.4 Training & Evaluation.

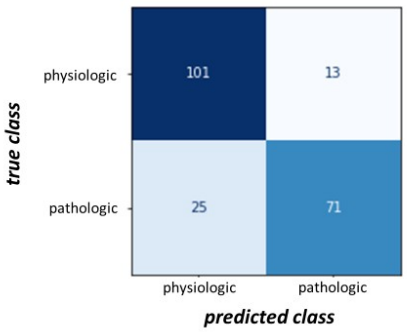
In order to find suitable hyperparameters, a train-validation-split pre-experiment with randomly but stratified selected 198 PVGs as training data and 22 PVGs as validation data was conducted on the 2-class as well as on the 4-class task.
The CNN was then trained and evaluated using a 10-fold cross-validation strategy (c.f. Fig. 2(b)). Each fold comprises a stratified training set of 198 PVGs and a stratified validation set of 22 PVGs. Based on the findings of the pre-experiment, for both classification tasks, the CNN was trained with a $batch\_size$ of 32 and over 100 $epochs$. As loss function, an according to the class frequencies weighted cross-entropy was used, with an Adam Optimizer and a $learning\_rate = 0.0005$. The dropout probability for the convolutions was set to 0.2.
The predictive power of the CNN was assessed after each fold and for the individual classes using accuracy, sensitivity, and specificity. The overall predictive power of the CNN was finally evaluated by considering all folds.

3 Results

This work realized a classification of Phonovibrograms using a CNN. We investigated two different classification tasks: (a) Differentiation between *physiologic* and *pathologic* VF vibration (2-class task) and (b) PVG Classification according to the actual diagnosis as *healthy*, *MTD*, *paresis*, or *polyp* (4-class task).

(a) Differentiation between *physiologic* and *pathologic* VF-vibration

The CNN trained to distinguish between physiologic and pathologic VF-vibration achieved over all 10 folds an average accuracy of $ACC^{avg}_{2-class} = 0.82 \pm 0.07\ (0.81)$ with an average sensitivity of $SEN^{avg}_{2-class} = 0.81 \pm 0.12\ (0.80)$ and an average specificity of $SPEC^{avg}_{2-class} = 0.82 \pm 0.12\ (0.81)$ [all values indicated as: mean $\pm$ standard deviation (median)]. The respective confusion matrix, as well as the detailed results, are depicted in Figure 4.



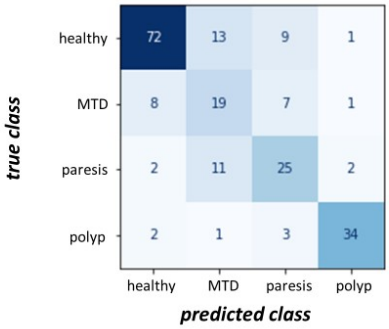
	sensitivity	specificity	accuracy	N
physiologic	0.89 ± 0.08 (0.91)	0.74 ± 0.16 (0.70)	0.82 ± 0.07 (0.81)	100
pathologic	0.74 ± 0.16 (0.70)	0.89 ± 0.08 (0.91)	0.82 ± 0.07 (0.81)	120
weighted average	**0.81 ± 0.12 (0.80)**	**0.82 ± 0.12 (0.81)**	**0.82 ± 0.07 (0.81)**	220

Fig. 4: Results on the differentiation between physiologic and pathologic VF vibration. Classification results were retrieved from all 10 folds and incorporated the entire dataset. The confusion matrix on the left visualizes the frequency of predicted classes for the two classes, *physiologic* and *pathologic*, individually. Detailed information on classification performance as measured by the parameters sensitivity, specificity, and accuracy is given inside the table. Moreover, overall performance is given by the weighted average of the respective measures, where class-individual measures were weighted with class frequencies to avoid any bias by unequal class sizes. Values are stated as: mean ± standard deviation (median), N states the class frequency.

Physiologic VF vibration patterns were classified correctly with a sensitivity of $SEN_{phys} = 0.89 \pm 0.08\ (0.91)$, while the specificity for these subjects was $SPEC_{phys} = 0.74 \pm 0.16\ (0.70)$. On the other hand, sensitivity for the pathologic VF-vibration patterns was reduced to $SEN_{pat} = 0.74 \pm 0.16\ (0.70)$ with a specificity of $SPEC_{pat} = 0.89 \pm 0.12\ (0.81)$.
In order to systematically investigate the misclassifications and to analyze whether these confusions might be related to individual pathologies, in a second step a CNN was trained on the classification of PVGs according to their actual clinical diagnosis.

(b) PVG Classification according to the actual diagnosis

When classifying the PVGs according to their actual diagnosis, the CNN achieved over all 10 folds an average accuracy of $ACC^{avg}_{4-class} = 0.85 \pm 0.07\ (0.84)$ with a sensitivity of $SEN^{avg}_{4-class} = 0.71 \pm 0.19\ (0.76)$ and an average specificity of $SPEC^{avg}_{4-class} = 0.91 \pm 0.07\ (0.92)$. Detailed results and the respective confusion matrix are shown in Figure 5.



	sensitivity	specificity	accuracy	N
healthy	0.76 ± 0.13 (0.74)	0.90 ± 0.08 (0.91)	0.83 ± 0.06 (0.81)	100
MTD	0.53 ± 0.30 (0.58)	0.86 ± 0.08 (0.86)	0.80 ± 0.09 (0.79)	40
paresis	0.62 ± 0.20 (0.75)	0.89 ± 0.10 (0.91)	0.84 ± 0.09 (0.86)	40
polyp	0.85 ± 0.23 (1.00)	0.98 ± 0.03 (1.00)	0.95 ± 0.04 (0.95)	40
weighted average	**0.71 ± 0.19 (0.76)**	**0.91 ± 0.07 (0.92)**	**0.85 ± 0.07 (0.84)**	220

Fig. 5: Results on PVG classification according to the actual clinical diagnosis. Classification results were retrieved from all 10 folds and incorporated the entire dataset. The confusion matrix on the left visualizes the frequency of predicted classes for each of the four classes *healthy*, *MTD*, *paresis*, and *polyp*. Detailed information on classification performance for the four classes as measured by the parameters sensitivity, specificity, and accuracy are given inside the table. Moreover, overall performance is given by the weighted average of the respective measures, where class-individual measures were weighted with class frequencies to avoid any bias by unequal class sizes. Values are stated as: mean ± standard deviation (median), N states the class frequency.

On this classification task, the class *polyp* outperformed all other classes and in all three measures ($ACC_{polyp} = 0.95 \pm 0.04$ (0.95), $SEN_{polyp} = 0.85 \pm 0.23$ (1.00), $SPEC_{polyp} = 0.98 \pm 0.03$ (0.92)). Organic lesions on the VFs, like polyps, lead to characteristic local alterations in the PVG pattern (c.f. Fig. 1(b)). Moreover, as pathologies with organic lesions are represented in this study only by the class *polyp*, high sensitivity and high specificity for this class indicate that these local PVG alterations caused by the polyp might help to detect this pathology.
On the other hand, PVGs for the remaining three classes *healthy*, *MTD*, and *paresis* mainly differ in their degrees of symmetry and regularity of the VFs vibrational behavior (c.f. Fig. 1(b)), making differentiation between these classes harder. This fact also becomes clear from the classification results, as confusions often occurred between the classes *healthy* and *MTD* as well as between the classes *MTD* and *paresis* (c.f. confusion matrix in Fig. 5). Nevertheless, when comparing these three classes, the class *healthy* outperformed both pathologic classes, in both sensitivity and specificity ($SEN_{healthy} = 0.76 \pm 0.13$ (0.74), $SPEC_{healthy} = 0.90 \pm 0.08$ (0.91)). This is likely due to the different sized classes where much more cases were available from healthy subjects than from pathological cases ($N_{healthy} = 100$ vs. $N_i = 40$ for all pathologic groups *i = MTD, paresis, polyp)*.

4 Discussion

This work presents a fully automatic classification of the VFs vibrational behavior from PVGs using a CNN. Based on the PVG representation, the presented approach has shown to differentiate between physiologic and pathologic vibrational behavior reliably and has also demonstrated reliable classification of the PVG according to the four considered clinical diagnoses.
On a 2-class discrimination task between physiologic and pathologic VF vibration from PVG, Voigt et al. achieved an average classification accuracy of 0.744 when the pathologic subjects were MTD patients [17] and average classification accuracy of 0.93 when the pathologic subjects were paralytic [18]. While the here presented CNN-based approach with an average accuracy of $ACC^{avg}_{2-class} = 0.82 \pm 0.07$ outperforms the SVM-based classification of healthy vs. MTD patients, it shows clearly reduced performance when compared to the SVM-based classification of healthy vs. paralytic patients. These differences in classification performance might be due to the fact that in the here presented work, the pathologic class consisted of subjects with three different diagnoses. And as these diagnoses present themselves totally differently in the PVG, making the classification task somewhat harder. Moreover, by only 220 PVGs, the dataset is relatively small compared to datasets usually used to train a CNN comprising thousands of samples. This is also limiting the achievable classification performance.
However, when considering the actual clinical diagnosis, on the 4-class classification task, the here presented Deep Learning based approach with an average accuracy of $ACC^{avg}_{4-class} = 0.85 \pm 0.07$ outperformed the previous approach by Unger et al., which achieved an average accuracy of $ACC^{avg}_{SVM} = 0.69 \pm 0.02$ [8]. These results are directly comparable as both works are based on the identical data set.
Confusions between the classes occured mainly between *healthy* and *MTD* subjects as well as between *MTD* and *pareses*. All these three classes share that there are no organic lesions located at the VFs altering the PVG representation. Instead, the pathologies (*MTD* and *paresis*) are mainly characterized by asynchronities and asymmetries in the VFs vibrational behavior. These asynchronities and asymmetries are presumably harder to detect as irregularities marked by characteristic alterations in the PVG patterns caused by organic lesions, i.e. by a polyp (c.f. 1 (b)). This is especially relevant for such a comparatively small dataset as the one used in this study.
Despite the small dataset, the CNN showed overall promising results. However, much more data are needed to extend this approach further and increase classification performance. Acquisition of additional clinical data and refinement of the here presented approach will be a focus of future works.

5 Conclusion

Due to its time-consuming nature, visual assessment of the VFs vibrational behavior as comprised in the HSVs is not feasible in clinical routine and moreover demands an experienced clinician.
In this work, we presented for the first time an approach that employs a deep Convolutional Neural Network to fully automatically classify different types of voice disorders. The presented approach is based on the VFs vibrational behavior as encoded in the PVG representation and has shown to reliably differentiate between physiologic and pathologic vibrational behavior. It is also eligible to classify different types of voice disorders without the need of any user interaction.
More training data will be required to increase the classification performance further. With an increased amount of training data, in the future it will be possible to realize more complex and thus even more powerful CNNs.

Acknowledgments

Computational resources were provided by the High Performance Compute Cluster 'Elwetritsch' at the University of Kaiserslautern, which is part of the 'Alliance of High Performance Computing Rheinland-Pfalz' (AHRP). We kindly acknowledge the support.

References

1. Titze, I.R.: Principles of Voice Production. National Center for Voice and Speech (2000)
2. Roy, N., Merrill, R.M., Gray, S.D., Smith, E.M.: Voice Disorders in the General Population: Prevalence, Risk Factors, and Occupational Impact. The Laryngoscope **115**(11) (nov 2005) 1988–1995
3. Bhattacharyya, N.: The prevalence of voice problems among adults in the United States. The Laryngoscope **124**(10) (may 2014) 2359–2362
4. Mehta, D.D., Hillman, R.E.: Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. Current Opinion in Otolaryngology & Head & Neck Surgery **16**(3) (June 2008) 211–215
5. Hertegård, S.: What have we learned about laryngeal physiology from high-speed digital videoendoscopy? Current Opinion in Otolaryngology & Head and Neck Surgery **13**(3) (#jun# 2005) 152–156
6. Deliyski, D.D., Hillman, R.E.: State of the Art Laryngeal Imaging: Research and Clinical Implications. Current Opinion in Otolaryngology & Head and Neck Surgery **18**(3) (#jun# 2010) 147–152
7. Lohscheller, J., Eysholdt, U.: Phonovibrogram Visualization of Entire Vocal Fold Dynamics. The Laryngoscope **118**(4) (#apr# 2008) 753–758
8. Unger, J., Schuster, M., Hecker, D.J., Schick, B., Lohscheller, J.: A multiscale product approach for an automatic classification of voice disorders from endoscopic high-speed videos. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (jul 2013)
9. Timcke, R., von Leden, H., Moore, P.: Laryngeal Vibrations: Measurements of the Glottic Wave: Part I. The Normal Vibratory Cycle. A.M.A. Archives of Otolaryngology **68**(1) (1958) 1–19
10. Švec, J.G., Schutte, H.K.: Videokymography: High-Speed Line Scanning of Vocal Fold Vibration. Journal of Voice **10**(2) (1996) 201–205
11. Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M.: Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibrations Into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics. IEEE Transactions on Medical Imaging **27**(3) (#mar# 2008) 300–309
12. Deliyski, D.D., Petrushev, P.P., Bonilha, H.S., Gerlach, T.T., Martin-Harris, B., Hillman, R.E.: Clinical Implementation of Laryngeal High-Speed Videoendoscopy: Challenges and Evolution. Folia Phoniatrica et Logopaedica **60**(1) (#nov# 2007) 33–44
13. Karakozoglou, S.Z., Henrich, N., d'Alessandro, C., Stylianou, Y.: Automatic glottal segmentation using local-based active contours and application to glottovibrography. Speech Communication **54**(5) (jun 2012) 641–654
14. Unger, J., Meyer, T., Herbst, C.T., Fitch, W.T.S., Döllinger, M., Lohscheller, J.: Phonovibrographic wavegrams: Visualizing vocal fold kinematics. The Journal of the Acoustical Society of America **133**(2) (2013) 1055
15. Chen, G., Kreiman, J., Alwan, A.: The glottaltopogram: A method of analyzing high-speed images of the vocal folds. Computer Speech & Language **28**(5) (2014) 1156–1169
16. Herbst, C.T., Unger, J., Herzel, H., Švec, J.G., Lohscheller, J.: Phasegram Analysis of Vocal Fold Vibration Documented With Laryngeal High-speed Video Endoscopy. Journal of Voice **30**(6) (nov 2016) 771.e1–771.e15
17. Voigt, D., Döllinger, M., Braunschweig, T., Yang, A., Eysholdt, U., Lohscheller, J.: Classification of functional voice disorders based on phonovibrograms. Artificial Intelligence in Medicine **49**(1) (may 2010) 51–59
18. Voigt, D., Döllinger, M., Yang, A., Eysholdt, U., Lohscheller, J.: Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. Computer Methods and Programs in Biomedicine **99**(3) (sep 2010) 275–288
19. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation **1**(4) (1989) 541–551
20. Shuang, Y., Gang, W.: Research on Meter Image Recognition Based on Improved LeNet-5. In: Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering, ACM (nov 2020)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. The Journal of Machine Learning Research **15**(1) (2014) 1929–1958