https://doi.org/10.2352/EI.2023.35.9.IPAS-300

Automatic Estimation of Mucosal Waves Lateral Peak Sharpness – Modern Approach

Aleš Zita [a]***, Šimon Greško*** [a]***, Adam Novozámský*** [a]***, Michal Šorel*** [a]***, Barbara Zitová*** [a]***, Jan G. Švec*** [b]***, Jitka Vydrová*** [c]
[a] ***The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague, Czech Republic***
[b] ***Palacký University, Faculty of Sciences, Department of Experimental Physics, Voice Research Lab, Olomouc, Czech Republic***
[c] ***Voice Centre Prague, Medical Healthcom, Ltd., Prague, Czech Republic***

Abstract

Videokymographic (VKG) images of the human larynx are often used for automatic vibratory feature extraction for diagnostic purposes. One of the most challenging parameters to evaluate is the presence of mucosal waves and their lateral peaks' sharpness. Although these features can be clinically helpful and give an insight into the health and pliability of vocal fold mucosa, the identification and visual estimation of the sharpness can be challenging for human examiners and even more so for an automatic process. This work aims to create and validate a new method that can automatically quantify the lateral peak sharpness from the VKG images using a convolutional neural network.

Introduction

Videokymography is one of the fast-growing fields of vocal cords vibration visualization techniques. The method uses a line scanner camera to visualize a vibratory pattern of the larynx and its neighboring tissue (Figure 1). Vertically stacked scanned lines create a spatial-temporal videokymographic image (see Figure 2). Physicians use this visualization to evaluate the vibration characteristics of the vocal folds for diagnostic purposes, often with the help of an automatic software tool capable of extracting the essential characteristics and features [1]. The line scanner camera operates with a frequency of 7200 fps and typically produces 25 VKG images every second [2, 3]. Due to a large amount of data, VKG images are suitable for computer processing.

A phase difference between the movement of the lower and upper vocal cord edges causes the mucosal wave on the medial surface of the vocal cords. In the VKG images, a significant phase difference appears as a double contour during the glottis closure phase and as sharp lateral oscillation peaks (refer to Figure 3). If the phase difference between the upper and lower vocal cord margins is relatively small, it will show up as a restricted glottal wave and rounded lateral peaks in the VKG images. The occurrence and the shape of mucosal waves on the vibrating vocal cords are crucial indicators of larynx health conditions.

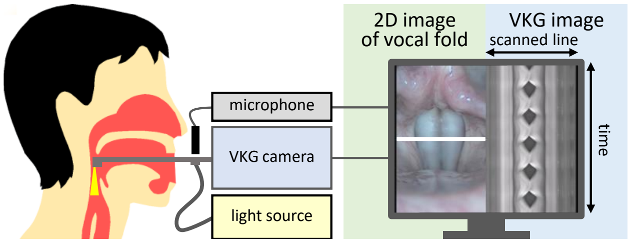

Figure 1. *Videokymography examination of the patient. The whole videokymographic frame comprises a 2D space image of the vocal fold (left side) and the temporal image of the scanned middle line highlighted in white in the 2D image (right side).*

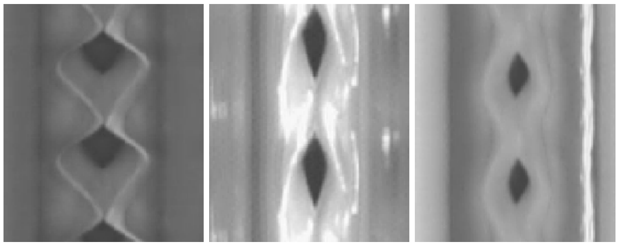

Figure 2. *Videokymographic images, where the vertical axis represents the temporal and the horizontal axis denotes the spatial domain. Here we see three different types of sharpness from left to right - sharp, rather rounded, and rounded.*

The performance of deep learning systems increased significantly in the last few years. In some areas, the machine learning approach exceeds the actual human experts. The main goal of this study is to verify the usability of deep learning for mucosal wave presence detection and the sharpness evaluation of waves' lateral peaks, one of the most complicated tasks in videokymographic image analysis.

In lateral peak sharpness assessment, manual ratings of the same images can vary between the examining experts; even single human professional evaluations may differ when repeated. Several influences cause these inconsistencies in rating, such as different levels of experience, length of practice, or such a trifle as the order of individual images. The combination of all these can bias the final assessment.

Related Work

Several researchers focused on the mucosal wave properties automatic estimation from the Videokymographic images [4, 5, 6], but due to its complexity, only a few have addressed the wave lateral peak sharpness.

Jiang et al. in 2000 [7] used an indirect method of peak sharpness estimation by quantifying the vertical phase difference using a sinusoidal model approximation. Although the method is correct in theory, the more complex shapes of the glottal contour are hard to process or interpret.

Yamauchi et al. [8] choose a different approach. They defined a *Lateral Peak Index* as an angle formed by two lines be-

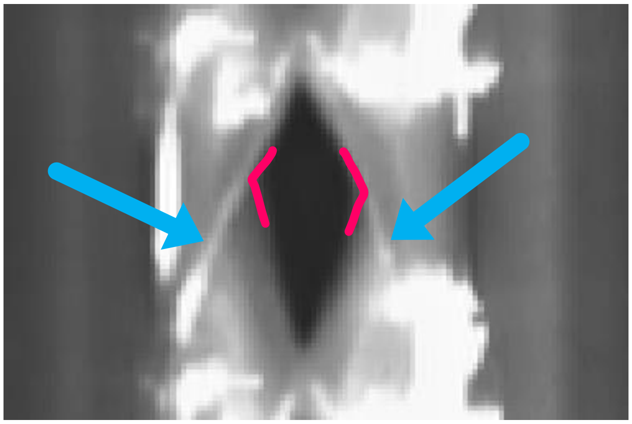

Figure 3. Mucosal wave as viewed on the VKG image. The blue arrows denotes the mucosal wave and the red marking shows the lateral peak

tween the start of the open phase and the lateral peak; and between the lateral peak and the end of the open phase. They quantified the sharpness of the lateral peak using the defined index. The drawbacks of this approach are that the index is sensitive to unrelated factors and discounts the changes of curvature of the vocal fold waveform that influence peak sharpness.

In [9], the researchers proposed four quotients that are good indicators of lateral peak sharpness. A set of proposed quotients was automatically calculated from the glottal contour line using the VKG Analyzer tool [1]. Four of the derived quotients had the best correspondence with the visual ratings of human experts, namely, PQ_{95}, PQ_{80}, OTQ_{95}, and OTQ_{80}. They are the variants of the *Plateau Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds $R\%$ of vibration amplitude within the open phase (denoted as PQ_R) and the *Open Time Percentage Quotients*, defined as the proportion of time during which the vocal fold displacement exceeds a chosen percentage (R) of the vibration amplitude within a period (denoted as OTQ_R).

All the publications mentioned above attempted to estimate the sharpness of lateral peaks from VKG images using conventional image processing and mathematical approaches. To our knowledge, no other teams pursued this topic using a machine learning system.

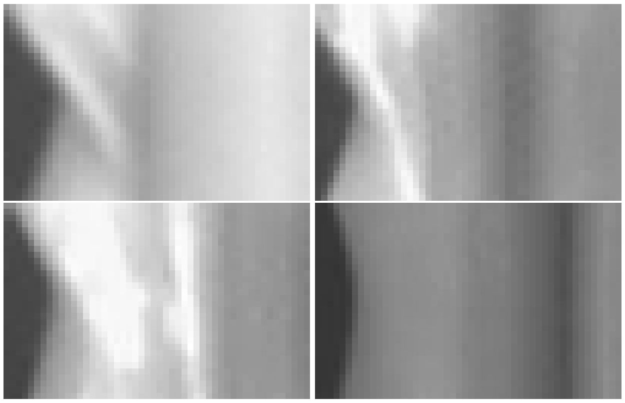

Figure 4. Cropped parts of videokymograms for sharpness classification; from left to right - sharp, rather sharp, rather rounded, and rounded.

Methodology

The proposed automatic mucosal wave lateral sharpness estimation method is defined as a deep learning classification task using a convolutional neural network (CNN). The technique uses fine-tuning of a pre-trained CNN architecture, trained on a set of data from a given application domain, in our case, videokymograms. The final system works with two same neural networks with different trained weights. The first one classifies the lateral peaks of a videokymogram into one of four classes of sharpness [*sharp, rather sharp, rather rounded, rounded*]; see the examples in Figure 4. The second one estimates the mucosal wave length into one of four ranges [*0-25%, 25-50%, 50-75%, 75-100%*].

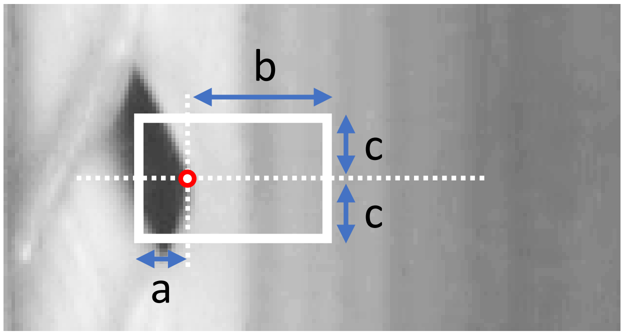

Figure 5. Cropping of the neighborhood of the lateral peak.

After trying and testing several different approaches, the best results were achieved using the following pipeline: We took advantage of the fact that videokymograms typically contain several visually almost identical vocal cord opening and closing cycles. Using a VKG Analyzer tool [1], we segment the vocal folds and detect significant points (lateral peaks, opening, closing). That gives us the beginning and end of each cycle. Within each cycle, we cropped the local neighborhood following these coordinates $[x_l - a : x_l + b, y_l - c + 1 : y_l + c]$; where $[x_l, y_l]$ represents the lateral point of a particular cycle. The graphic representation is in Figure 5.

A neural network subsequently classifies these cropped parts one after the other. In this way, we detect the relevant features separately for each cycle. The resulting overall value for the whole videokymogram is the most frequent (mode) for both the right and left sides. Additionally, the agreement of the results on individual cycles within a videokymogram determines the reliability of the estimation.

Parameters a, b, and c need to be set according to the size of one cycle in the image (in pixels). Gender, length of the throat, or the frequency of the vocal cords, is one of the main factors which cause these differences. We can normalize the whole image or adapt the size of the cropping. In our case, we selected images with similar sizes of cycles from different examinations of patients. Therefore we could set up the values globally as a=10, b=40, and c=16. No other normalization of size was needed.

We used *MobileNetV2*[10] with the *Adam*[11] optimizer algorithm as the backbone of our algorithm for its advantage of few parameters and few operations, which leads to easy implementation, fast inference, and not demanding hardware. The same neural network architecture and dataset were used to automatically estimate the lateral peak's sharpness and mucosal waves' presence. We trained the system on the left and right halves of the

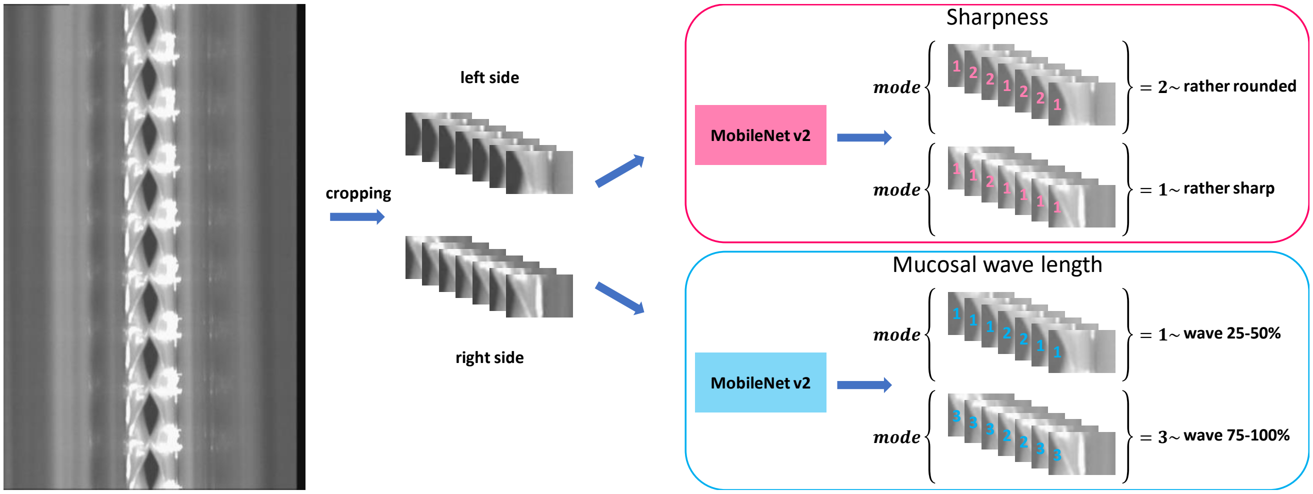

Figure 6. *Schematic representation of our method. The first step is a specific pre-processing operation (image enhancement and normalization). Then we cut individual complete cycles from the left and right sides of the kymogram. The cropped data is sent to CNN for sharpness and mucosal wave length classification in the next step. The final decision is to find the most common class on both sides.*

videokymogram together, with the left side first rotated around the vocal tract axis to the same position as the right side. Figure 6 shows the whole pipeline.

Dataset

All data used in this study are from examinations performed on patients of different ages, gender, and health conditions at the *Voice and Hearing Centre, Medical Healthcom, Ltd, Prague*. All VKG images come from the second generation VKG camera (Kymocam, CYMO, b.v. Groningen, the Netherlands) with a combination of different connected laryngoscopes, objective adapters, or light sources. We used two vocal cord specialists from the same department to evaluate the images.

The robustness of the CNN-based approach strongly depends on the quality of the training set used. Therefore, we have focused on data collection and subsequent annotation using the proposed web annotation tool we created for this task, see Figure 7. Using this tool, we could randomly display individual images to experts and store their sharpness ratings. For control, we showed each image several times in random order. In the resulting database, we only included images on which the experts agreed.

After this evaluation, our database consists of 319 expert-rated VKG images from clinical records. The images were processed and analyzed by VKG Analyzer software [1], and particular cycles were extracted and saved for subsequent sharpness analysis. In this way, a database of 3695 cropped parts with a size of 32x50 pixels was created, which was then divided into training

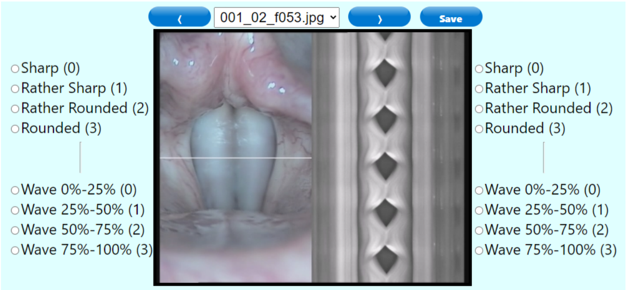

Figure 7. *A screenshot from the proposed tool used for the training dataset manual annotations. Annotators evaluated the sharpness of left and right lateral peaks and the level of mucosal wave length (percentage denotes the distance to the neighboring tissue, see Figure 3 for illustration of 100% wave).*

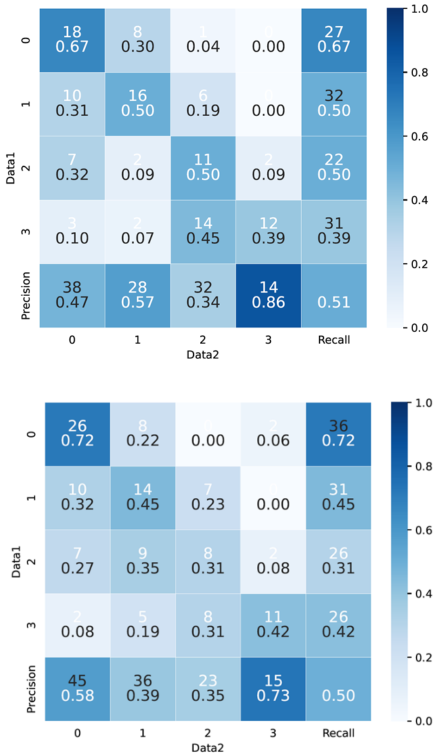

Figure 8. *The confusion matrices between professional physicians and the convolutional network in **wave presence detection**. The tables show the evaluated features' precision, recall, and accuracy. Professional physicians' ratings are on the vertical axes and the proposed method results on the horizontal axis. The upper table belongs to the left side of the vocal cord, the lower to the right.*

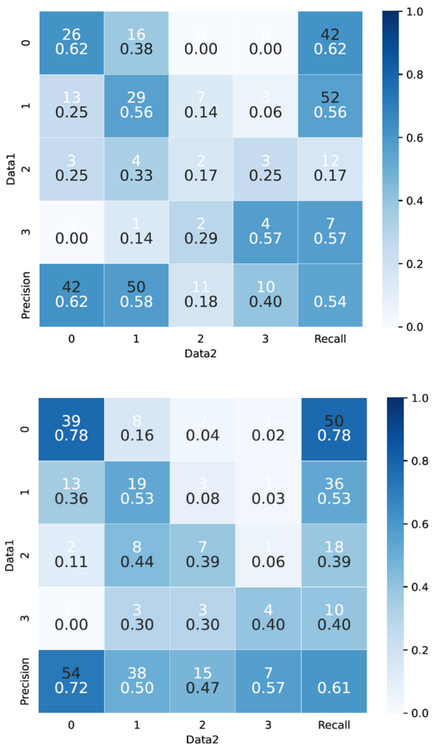

Figure 9. *The confusion matrices between professional physicians and the convolutional network in **lateral peak sharpness**. The tables show the evaluated features' precision, recall, and accuracy. Professional physicians' ratings are on the vertical axes and the proposed method results on the horizontal axis. The upper table belongs to the left side of the vocal cord, the lower to the right.*

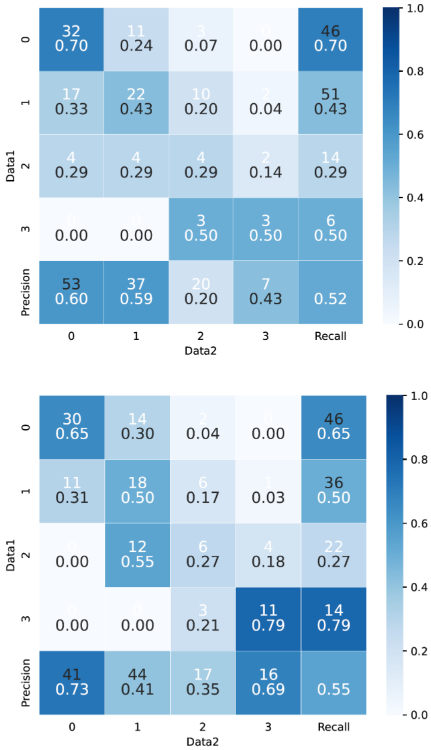

Figure 10. *The performance comparison of the junior evaluators' ratings vs. the professional physician. The table shows the professional (ground truth) on the vertical axis and the examining rater on the horizontal axis. The bottom right corner denotes the overall accuracy. The upper table belongs to the left side of the vocal cord, the lower to the right.*

and test parts in a ratio of 3:2. Examples of individual sharpness classes can be seen in Figure 4. Due to the vocal cords' symmetry, we could evaluate the left and right vocal cords simultaneously. The results of our method on the validation set in the form of confusion matrices can be seen in the following section of this document.

Results

The network performance results are presented in the form of confusion matrices. Figures 8 and 9 show both evaluated features' values for the left and right sides. The vertical axis corresponds to the ground truth value determined by the expert (values 0 to 3), while the horizontal axis corresponds to the values determined by the machine learning algorithm. The integer values in the contingency table correspond to the number of cases (combinations of expert and algorithm values). Numbers below the combination values are the same values normalized to the sum of one, row-wise. The rightmost column shows the recall values, and the bottom row shows the precision values. The bottom right corner then displays the overall percentage of exact match (accuracy). Then we wanted to compare the ratings of two junior evaluators with our professional physicians. The results of this comparison are shown in Figure 10; the interpretation is the same as in Figure 8.

Finally, the proposed machine learning algorithm achieves an accuracy of **0.54** and **0.61** for right and left lateral peak sharpness, which exceeds the match of success of junior evaluators (0.52 and 0.55). Similarly, compared to junior evaluators' values, it improves the rating for wave length, reaching an accuracy of **0.51** and **0.50** for the left and right sides. The system's performance will improve in the future through our continuous fine-tuning of the network with newly acquired data. We consider this to be the study's main result, as it will allow us to automate and objectify the estimation of an important parameter for evaluating the condition of the vocal cords.

Conclusion

We have developed a CNN-based tool for automatically estimating lateral peak sharpness and the mucosal wave length in videokymograms. The performance of this tool was evaluated on a small dataset, and the results indicate that the proposed method can accurately assess the sharpness of lateral peaks, and the level of accuracy is higher or comparable to that of non-specialists.

In addition, when trained on a more extensive and diverse dataset, we expect significant improvements which will be approaching accuracy achieving professional physicians. In that case, we could handle a more comprehensive range of videokymograms and accurately assess the sharpness of lateral peaks in a broader range of vocal conditions. Overall, this tool shows promise as a reliable and efficient method for evaluating the health and function of the vocal cords.

Acknowledgments

This work was partially supported by the Technology Agency of the Czech Republic, grant No. TH04010422, by the Czech Science Foundation, grant no. GA21-03921S, and by the *Praemium Academiae* awarded by the Czech Academy of Sciences.

References

[1] Aleš Zita, Adam Novozámský, Barbara Zitová, Michal Šorel, Christian T Herbst, Jitka Vydrová, and Jan G Švec, "Videokymogram analyzer tool: Human–computer comparison," *Biomedical Signal Processing and Control*, vol. 78, pp. 103878, 2022.

[2] Jan G. Švec and Harm K. Schutte, "Videokymography: High-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, pp. 201–205, 1996.

[3] J. G. Švec and F. Šram, "Videokymographic examination of voice," in *Handbook of Voice Assessments*, E. P. M. Ma and E. M. L. Yiu, Eds., pp. 129–146. San Diego, CA: Plural Publishing, 3rd edition, 2011.

[4] Adam Novozámský, Jiři Sedlář, Aleš Zita, Filip Šroubek, Jan Flussef, Jan G. Švec, Jitka Vydrová, and Barbara Zitová, "Image analysis of videokymographic data," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 78–82.

[5] Jack J Jiang, Yu Zhang, Michael P Kelly, Erik T Bieging, and Matthew R Hoffman, "An automatic method to quantify mucosal waves via videokymography," 2008.

[6] S. Pravin Kumar and Jan G. Švec, "Kinematic model for simulating mucosal wave phenomena on vocal folds," *Biomedical Signal Processing and Control*, vol. 49, pp. 328–337, 3 2019.

[7] Jack J. Jiang, Ching I.B. Chang, Joseph R. Raviv, Sameer Gupta, Franklin M. Banzali, and David G. Hanson, "Quantitative study of mucosal wave via videokymography in canine larynges," *Laryngoscope*, vol. 110, pp. 1567–1573, 2000.

[8] Akihito Yamauchi, Hisayuki Yokonishi, Hiroshi Imagawa, Ken-Ichi Sakakibara, Takaharu Nito, Niro Tayama, and Tatsuya Yamasoba, "Quantitative analysis of digital videokymography: A preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers," 2015.

[9] S.P. Kumar, K.V. Phadke, J. Vydrová, A. Novozámský, A. Zita, B. Zitová, and J.G. Švec, "Visual and automatic evaluation of vocal fold mucosal waves through sharpness of lateral peaks in high-speed videokymographic images," *Journal of Voice*, vol. 34, 2020.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[11] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.

Author Biography

Aleš Zita received his M.Sc. degree in informatics from the Faculty of Mathematics and Physics, Charles University, Prague, in 2013. He is pursuing his Ph.D. with the Institute of Information Theory and Automation Cooperating Institute of Charles University, Prague. His research interests include medical imaging, image segmentation, machine learning, and image forensics.