Automatic Evaluation of Speech Therapy Exercises Based on Image Data*

Zuzana Bílková[1], Adam Novozámský[1], Adam Domínec[1], Šimon Greško[1],
Barbara Zitová[1], and Markéta Paroubková[1]

The Czech Academy of Sciences, Institute of Information Theory and Automation,
Pod Vodárenskou věží 4, Praha 8, 182 08, Czech Republic
bilkova@utia.cas.cz

Abstract. The presented paper proposes a new method for unique automatic evaluation of speech therapy exercises, one part of the future software system for speech therapy support. The method is based on the detection of the lips and tongue movements, which will help to evaluate the quality of the exercise implementation. Four different types of exercises are introduced and the corresponding features, capturing the quality of the movements, are shown. The method was tested using manually annotated data and the proposed features were evaluated and analyzed. At the second part, the tongue detection is proposed based on the convolutional neural network approach and preliminary results were shown.

Keywords: Lip movement detection · Tongue segmentation · Speech therapy.

1 Introduction

Automatic evaluation of speech therapy exercises is one of the focuses in our research to create a software system to support speech therapy for children and adults with inborn and acquired motor speech disorders. The system aims to improve articulation and a tongue motion based on individual treatment using exercises recommended by a therapist. The key component of the system is the evaluation of a tongue and lips motion based on image data from an ordinary web camera.

Existing applications only passively show exercises in the form of images, videos or text descriptions for the voice formulation but none of the available applications offers the possibility of an automatic assessment with commonly available cameras. The use of digital imaging techniques is new and enables the degree of interactivity where the treatment itself becomes more effective and a patient's return to normal life is accelerated. In addition, the methodology of language motion detection is also applicable in other areas, such as controlling various devices in quadriplegics, where existing solutions include special sensors that are stuck to the tongue.

* Supported by TACR, grant No. TJ01000181, and the grant SVV–2017–260452.

1.1 Related work

To automatically evaluate the speech therapy exercises, tongue, its tip and lips motion must be detected. In the preliminary step a mouth must be located in the face image.

In the literature, there are several methods for face detection and detection of key facial features. Effective object detection proposed by Viola et al. [8] is a machine learning algorithm using Haar feature-based cascade classifiers. Widely used object detection method introduced in [2] by Dalal et al. is based on histogram of oriented gradients and support vector machines.

There are several methods for tongue segmentation. These methods are often used as the initial step for tongue diagnosis used in Chinese medicine. Zhang et al. [9] use tongue color, texture and and geometry features to detect diabetes. HSI color space was used for tongue segmentation in [3, 11]. Another common segmentation algorithm is the method of active contours, which is used in [6, 10]. All of these methods are slow and they are not robust. They also suppose the tongue to be stucked out in only one position which is not sufficient to evaluate the complex tongue motion.

We have not found any paper concerning detection of the tip of the tongue which is a very complicated problem due to the self-similarity of the tongue tissue. Another part of the solution is the detection of lips, which can be often found in the voice activity detection systems, for example in [5], but there are no papers about the lips detection for speech therapy. A systematic review of studies on computer-based speech therapy systems or virtual speech therapists is presented in [1].

1.2 Speech therapy software

The proposed software system will offer an adjustable set of exercises recommended by an expert and its evaluation and it will also motivate the patients by incorporating an augmented reality, such as an augmented picture of a ladybug sitting on the nose tip when the tongue should be stucked out. The augmented reality is employed into the solution to increase understanding of the proper exercise movements. Moreover, the system will offer session archiving allowing the therapist to evaluate treatment progress and adjust its schedule.

The main part of the software evaluates the quality of the exercise implementation. Using the therapist expertise, we have distinguished three groups of exercises based on the information necessary for their evaluation. They work with tracking of the patient's lips, with the segmentation of the tongue and of the tongue tip detection, respectively. In the next Section we will focus on processing of images of mouth to obtain lips movement patterns and in the Section 3 the tongue segmentation and localization of its tip is discussed.

2 Evaluation of the lips exercises

The evaluation of exercises when the lips have to follow specified motion patterns is based on the lips detection and their tracking. In the preliminary step a patient's mouth is located in the face image. The proposed solution for detection of face parts uses Dlib library [4], based on an ensemble of regression trees that are trained to estimate the facial landmark positions directly from the pixel intensities. The Dlib library automatically detects 68 points in the face image in real-time allowing an easy detection of the mouth and the lips. The features are described in Table 1.

For the further processing, we focus on the lips and mouth location only, thus the image of a face is cropped according the position of selected points, as it is shown in Figure 1. The cropped images are normalized with respect to their size and to the orientation between eyes. In order to evaluate the lips exercises we have proposed six features, derived from the positions of the detected lips points. They are illustrated in Figure 1 by blue drawings and described in Table 1. They capture main movement patterns of the lips. In our tests, these features proved to be sufficient to evaluate the quality of the patient exercises, see Section 2.1.

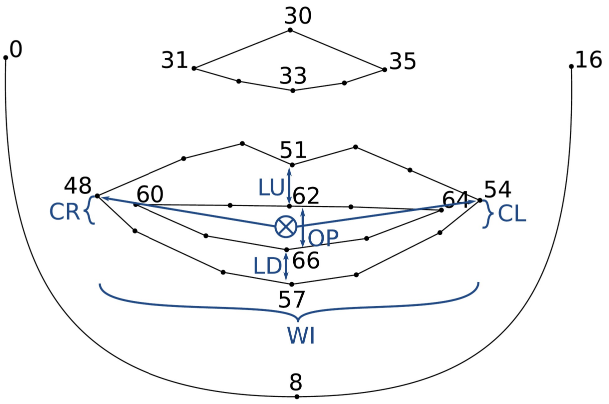


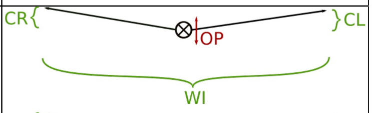
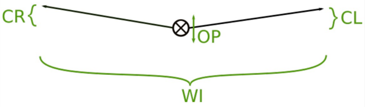
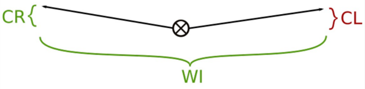
Fig. 1. Selection of key detected facial points and features for analysis of lips exercises.

Table 1. Description of main features.

Notation	Name	Description
CR	Corner - right	Vertical shift of the right corner with respect to the center.
CL	Corner - left	Vertical shift of the left corner with respect to the center.
LD	Lip - down	Height of the lower lip.
LU	Lip - up	Height of the upper lip.
OP	Open	Lips distance when mouth is open.
WI	Wide	Horizontal distance of left and right corner.

For this paper, we selected four representative lip exercises to show how the features are used in the evaluation process. The exercises are *closed smile, open smile* and *crooked smile*, when only one mouth corner, left or right, is lifted. Table 2 shows visualization of individual exercises with a formula for the corresponding feature calculation. If the exercise is executed correctly the values of the individual features are maximized. In the case of the closed smile both corners have to be lifted, therefore we control the minimum height of both corners, min(CL, CR), value -OP is maximized when the mouth is closed and finally the mouth widens when we smile, which captures the WI feature. The only difference in the open smile exercise is with the feature OP which controls the lips distance and here it should be maximized. In the case of the crooked smile, one of the corners is supposed to lift up and the other has to stay low and the mouth also widens, as it is shown in the last row of the table.

Table 2. Features used for evaluation of selected lips exercises.

Name	Features Calculation	Visualization
Closed smile	min(CL, CR), -OP, WI	
Open smile	min(CL, CR), OP, WI	
Crooked smile - right	CR, -CL, WI	

The individual exercises have to be practised several times. The number of repetition is set by the therapist. Figure 2 shows the automated counting of smiles performed. The successful attempts are appraised (the green smile) and counted, while the erroneous performance is highlighted by the yellow icon and the count of smiles stays intact.

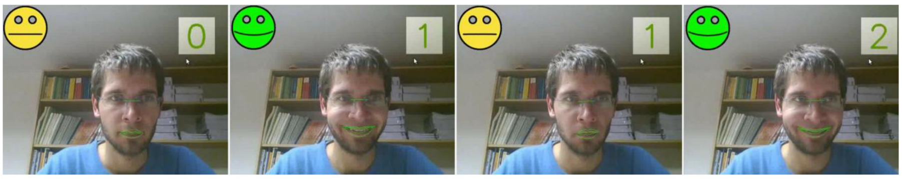


Fig. 2. Automated counting of smiles.

2.1 Analysis of the feature performance

Testing of the method for the exercise evaluation based on the feature computation was realized using manually annotated data. The positions of lips points were detected manually and the proposed features were evaluated and plotted to see if they reflect the expected motion patterns. Figure 4 shows the values of the four features from Table 2 (in different colors) for the four chosen exercises. Each exercise is three times repeated. On the x axis values for acquired video frames are shown. We can see that the chosen features correctly capture the desired characteristics of the exercises, namely the high values for width and height of both corners and no significant change in the feature describing opening for the closed smile and similar expected behavior for the other exercises.

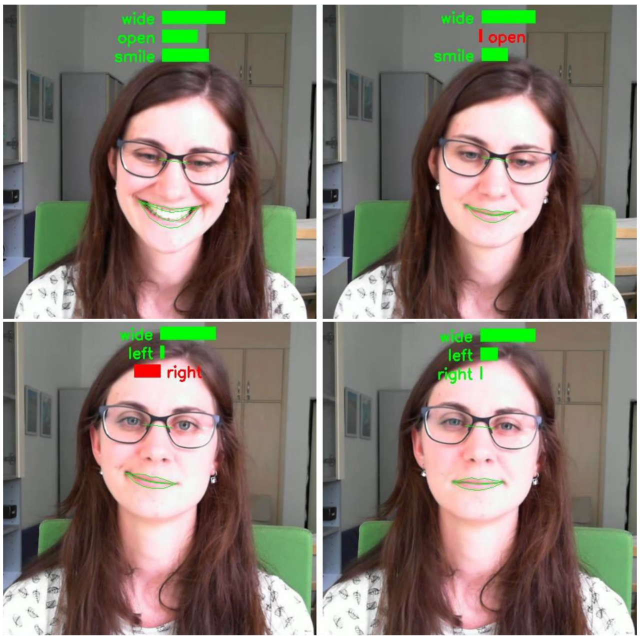


Fig. 3. Feature bars representing the value of features of selected lips exercises.

Our tests demonstrate the validity of our solution for the evaluation of the speech therapy exercises.

3 Segmentation of the tongue

The group of exercises based on the tongue and the tongue tip movements utilizes a segmentation of the tongue body and of its tip. For the speech therapy exercises we need to detect the tongue and its tip in real-time and in all their positions.

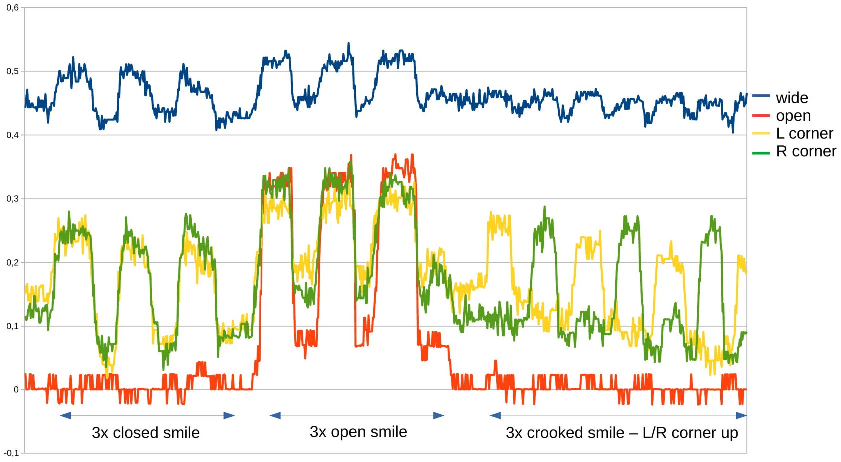


Fig. 4. Value of individual features in a video of performance of selected exercises.

Our solution is based on the convolutional neural network U-net [7] which proves to be sufficiently fast and robust. We use a single neural network to output both results - segmentation of the tongue and its tip - to achieve higher speed. The output of the network has thus two frames - one is a mask for the tongue segmentation and the other is a mask for its tip. The network is trained on data we partially recorded and partially downloaded from the Internet with different quality to ensure robustness of our method with respect to the data quality. The data were manually annotated. We are still enlarging our database to provide bigger and more versatile training dataset.

The results of our current network based on U-net for segmentation of tongue and its tip is shown in Figure 5. Green line and dot represent the ground truth - manually segmented tongue contour and the tip. The red lines are the prediction of the tongue contour. The red dots are the center of the mass of a mask predicted by the network, as shown in the middle column in the last row. We can see that the results correspond well to the ground truth even in the image where the whole tongue is in the mouth and the tip is hard to detect. However, to achieve required final robustness of the tongue detection method more testing and dataset collection creation is still needed.

4 Conclusion

The presented paper demonstrates the method for an automatic evaluation of speech therapy exercises. This is a part of the future software system for speech therapy support. The method is based on the detection of the lips and tongue movements and further evaluation of the quality of realized motion patterns, following the therapist recommendation. The efficiency of the method was demon-

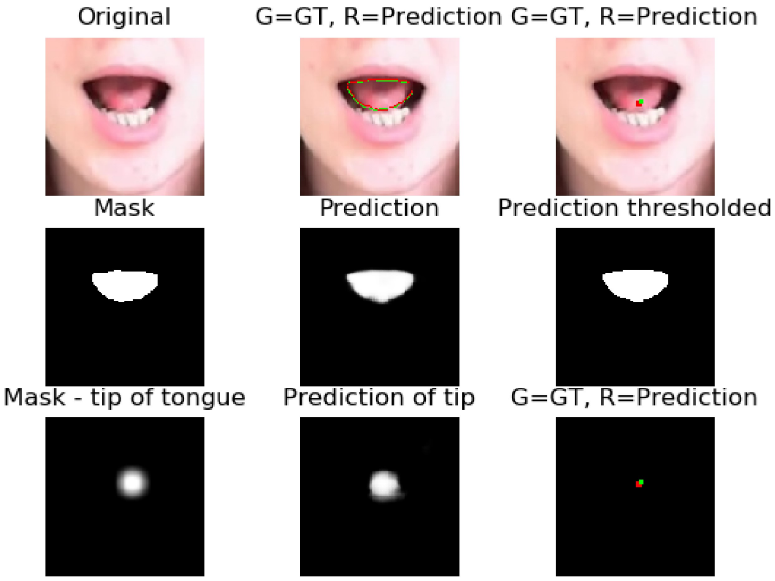


Fig. 5. Output of convolutional neural network Unet for segmentation of tongue and its tip.

strated on the four different types of exercises, which were manually annotated and the proposed features evaluated and analyzed. Their discriminability has to be shown to be sufficient to be able to distinguish between correct and erroneous realization of an exercise. The tongue detection was proposed using the neural network approach and preliminary results were shown.

References

1. Chen, Y.P.P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., et al.: Systematic review of virtual speech therapists for speech disorders. Computer Speech & Language **37**, 98–128 (2016)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: international Conference on computer vision & Pattern Recognition (CVPR'05). vol. 1, pp. 886–893. IEEE Computer Society (2005)
3. Du, J.q., Lu, Y.s., Zhu, M.f., Zhang, K., Ding, C.h.: A novel algorithm of color tongue image segmentation based on hsi. In: 2008 International Conference on BioMedical Engineering and Informatics. vol. 1, pp. 733–737. IEEE (2008)
4. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014)
5. Lopes, C.B., Gonçalves, A.L., Scharcanski, J., Jung, C.R.: Color-based lips extraction applied to voice activity detection. In: 2011 18th IEEE International Conference on Image Processing. pp. 1057–1060. IEEE (2011)

6. Pang, B., Zhang, D., Wang, K.: The bi-elliptical deformable contour and its application to automated tongue segmentation in chinese medicine. IEEE transactions on medical imaging **24**(8), 946–956 (2005)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
8. Viola, P., Jones, M., et al.: Rapid object detection using a boosted cascade of simple features. CVPR (1) **1**, 511–518 (2001)
9. Zhang, B., Kumar, B.V., Zhang, D.: Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features. IEEE transactions on biomedical engineering **61**(2), 491–501 (2014)
10. Zhang, H., Zuo, W., Wang, K., Zhang, D.: A snake-based approach to automated segmentation of tongue image using polar edge detector. International Journal of Imaging Systems and Technology **16**(4), 103–112 (2006)
11. Zhongxu, Z., Aimin, W., Lansun, S.: The color tongue image segmentation based on mathematical morphology and his model [j]. Journal of Beijing Polytechnic University **2** (1999)