ASSISLT: Computer-aided speech therapy tool

Zuzana Bílková, Michal Bartoš, Adam Dominec, Šimon Greško,
Adam Novozámský, Barbara Zitová, Markéta Paroubková
Czech Academy of Sciences
Institute of Information Theory and Automation
Prague, Czech Republic
{bilkova, bartos, dominec, gresko, novozamsky, zitova, paroubkova}@utia.cas.cz

***Abstract*—This work proposes a new software system ASSISLT to support speech therapy for children and adults using deep learning approaches. The application offers an adjustable set of exercises recommended by a speech therapist and aims to motivate and help with regular home practice. Augmented reality is employed to lead the exercise moves and to appraise the effort. The core of the system is automatically evaluating exercises using a webcam and developing image processing and neural network methods for face, lips, teeth, and tongue detection. The pipeline is shown together with solutions of subtasks and with demonstrations of the functionality. The statistical validation of the ASSISLT is provided, comparing the performance of speech therapy specialists and the software.**

***Index Terms*—medical imaging, tongue detection, lips detection, speech therapy**

I. Introduction

One of the keys of successful speech therapy is a regular practice. Together with the speech therapy specialist, we have developed new software system to support speech therapy for children and adults with inborn or acquired motor speech disorders such as dysarthria or dyslalia. The base component of our system is a module for automatic evaluation of exercises recommended by a speech therapist. The course of the exercise is recorded by a webcam, so no special equipment is needed. Although a near-infrared camera could provide a data stream that is less sensitive to changes in lighting, we have decided to use a standard webcam to ensure wide availability. The data stream is automatically evaluated in real time by image processing and neural network methods. The application offers an adjustable set of exercises designed to help and motivate regular home practice as we have previously outlined in [1]. The system is complemented with the augmented reality module leading the exercise moves and introducing the gamification element into the therapy.

To our knowledge, there is no other computer-aided speech therapy applications offering an automatic assessment of face muscles or tongue motions. Other applications only passively show pictures or videos of exercises but do not offer an individual feedback and/or evaluation. One of the best of such applications is Speech Tutor [2] which uses visualizations how sounds are created in the mouth and throat. Splingo's Language Universe [3] is an interactive game aiming to improve children's listening and receptive language skills by following spoken instruction. There are applications, such as Articulation station [4], that focus on pronunciation of individual sounds and give a score for audio recording.

As opposed, our application is unique in providing a direct real-time evaluation of exercise performance based on image processing of standard webcam data allowing widespread use. Moreover, the proposed methods for detection of the tip of the tongue and lips is applicable in other areas such as a human–computer interface for disabled people, as described in [5].

II. Software system

The software system offers three operational modes—a mode for therapists and two modes for patients with or without adjustments. In the mode for therapists, the expert is able to define a sequence of exercises together with their difficulty levels for each patient and to check a patient's progress via saved results from home practice. Patients can run the application and save the results of their progress either in the mode without adjustments, which opens the sequence of exercises recommended by a therapist, or in the mode enabling adjustments. In this mode the patient is able to modify the exercise settings and their order. The application workflow is shown in Fig. 1.

III. Automatic evaluation

In this section, we describe an automatic evaluation of various speech therapy exercises. Based on expert knowledge of speech therapists, we divide the recommended exercises into four groups processing cheeks, lips, tongue, and teeth.

All of the processing begins by precisely detecting the mouth location obtained from tracking facial landmarks [6]. This method with necessary performance operating in real-time is provided by the well-known Dlib library [7] (modification [8]). Eye corner landmarks are also used in order to detect any motion of the whole head, which would be undesirable from a therapeutical standpoint.

A. Cheeks processing

A couple of speech therapy exercises focuses on puffing cheeks. The cheeks can be puffed out either by air or by a tongue. This corresponds to two different exercises, but the processing follows the same scheme. The aim of the processing is to determine whether a cheek is puffed or not.

This work has been supported by TACR, grant No. TJ01000181, and the grant SVV–2017–260452.

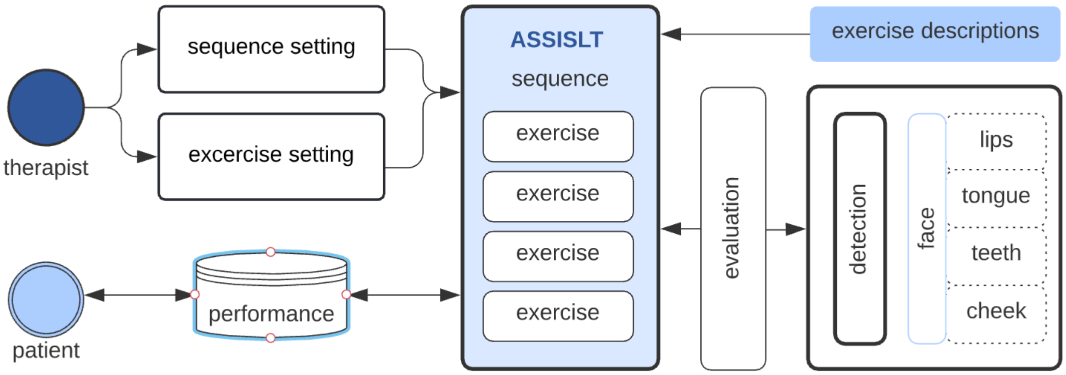


Fig. 1. The application workflow: the sequence of exercises is composed and configured by a speech therapist, choosing from a library of available exercises with descriptions and guidelines. While the patient completes the whole sequence, his/her performance in each of the exercises is automatically evaluated and stored for later review.

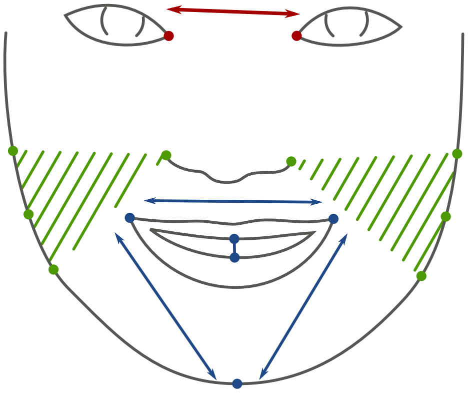


Fig. 2. An overview of facial landmarks used for tracking of facial expressions. Landmarks on inner eye corners (red dots) are used for stabilization and for prevention of a head movement. Landmarks on the chin and the nose (green shaded area) are used for evaluation of cheek exercises. Landmarks on the lips (blue dots) are used for exercises based on lips, tongue and teeth.

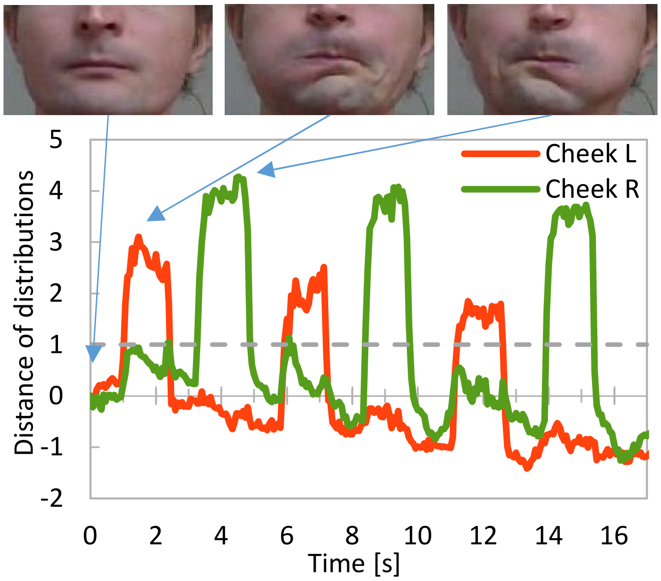


Fig. 3. Example of computed distances of brightness distributions between reference and processed cheeks during an exercise with **alternate puffing of cheeks**. The gray dashed line marks a threshold. If the value is above the threshold for at least one second, the respective cheek is marked as puffed. The images above demonstrate a reference image, a puffed left cheek and a puffed right cheek, respectively, at specified time points.

The processing is based on a measurement of brightness changes in regions of interest placed on each cheek defined by facial landmarks (Fig. 2, green areas; similar areas and the landmark detection is independently described in [9]). The idea is that a puffed cheek has a smooth surface and reflects more light towards a camera than in the case of neutral expression. In each cheek region of interest, a mean μ and a standard deviation σ of brightness are estimated. These statistics per cheek are also computed for a reference image taken at the beginning of the exercise when the patient is instructed to have a neutral expression. From the reference (r) and actual (a) statistics, the distance of the actual and the reference brightness distributions is measured using: $D = (\mu_a - \mu_r)/\sqrt{\sigma_a^2 + \sigma_r^2}$, which is related to the mean-difference term of the Bhattacharyya distance [10, p. 99]. An example of the computed distances (divided by sensitivity/difficulty coefficient) for the left and the right cheeks during an exercise is shown in Fig. 3.

The success of the proposed evaluation approach is based on following assumptions: the patient is steadily illuminated and pixel values in cheeks are far from saturation in the case of neutral expression; the actual face position is not far from the reference one. This last assumption goes in hand with the definition of the correct exercise.

B. Lips processing

There are many methods for lips localization [11], the most popular are the active appearance model [12] and Dlib. Our evaluation of lips exercises is based on the latter approach, deriving features from the facial landmarks from Dlib. For more detail about individual features and exercises, see [1]. A specific combination of the features for each speech therapy exercise results in a value that is thresholded in the ASSISLT

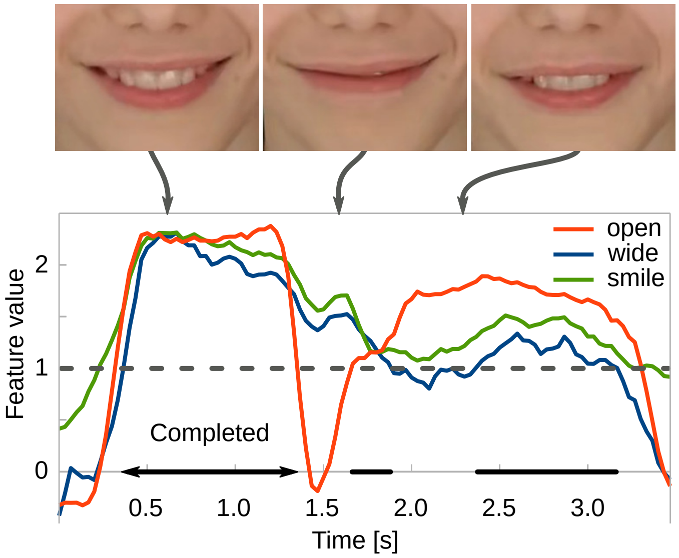


Fig. 4. Example of features computed during a **smiling exercise**. The x-axis represents time in the video sequence. The three features, open, wide, smile, are evaluated for each video frame. The dashed horizontal line marks the default threshold value for all of the features. The time intervals, when requirements of the exercise were met, are marked on the x-axis. Note that the patient closed his mouth briefly during the recording and has only barely completed the exercise.

application to decide whether the exercise is performed correctly.

As an example, see Fig. 4 showing a smiling exercise where features indicate whether lips are wide and open with visible teeth. In order to complete the exercise successfully, all of the features must remain above a threshold for a specified period of time (1 second in the default setting), and then they must remain below the threshold for the same duration.

C. Tongue processing

The group of tongue exercises is evaluated by detecting the tongue position and its tip. Estimated tongue positions and its motion between frames is then compared to the required set of moves (directions, range of motion). Each exercise has predefined areas where the tongue should be and how much it should be stretched (see Fig. 5).

Detection of a tongue and its tip is a very complex task. In the literature, there are several methods of tongue segmentation using various color schemes, tongue color, or texture, such as [13], [14]. Sage et al. [15] use a convolutional neural network and two cameras for computer-aided speech diagnosis and therapy. We have trained the U-Net [16] convolutional neural network on cropped images of a mouth to find a segmented tongue and a position of its tip. We used 1252 different frames from 60 videos of children and adults for training and 165 frames from 10 videos as a validation set. The videos are of different image quality which ensures the robustness of the method. The data were annotated manually. The output frames from the neural network are filtered so as to avoid potential outliers and smoothed with a Kalman filter.

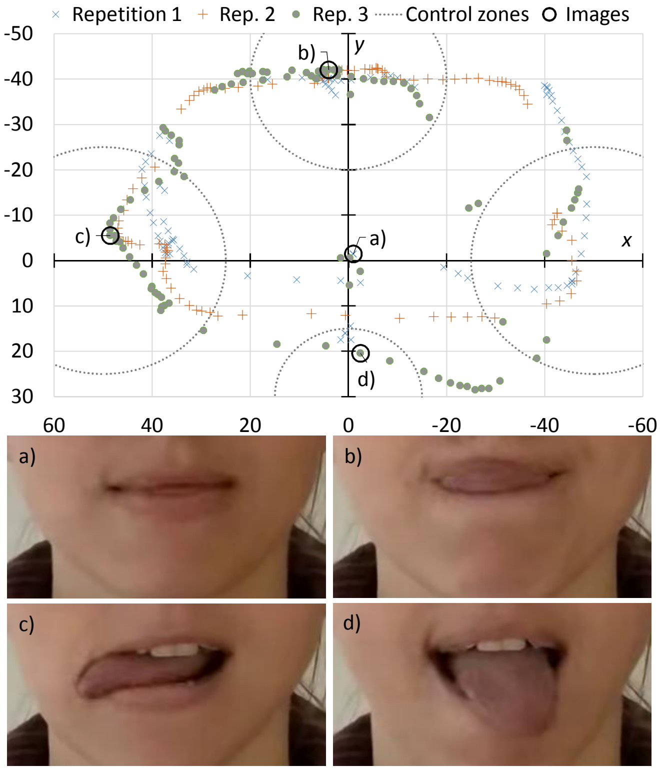


Fig. 5. Plot of the estimated position of the tongue tip for three consecutive repetitions of the exercise **tongue around**. The axis are reversed to correspond with natural interpretation. The coordinate center lies on the bottom lip. The repetition is evaluated as passed if the tip of the tongue passes consecutively all predefined control zones (dotted ellipses).

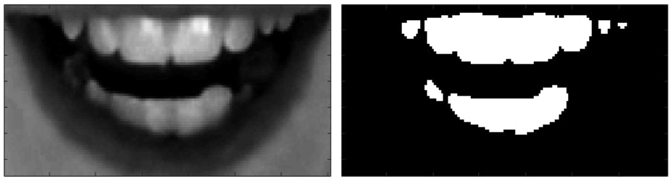


Fig. 6. Image of a mouth in custom color space (left) and its teeth segmentation using thresholding with masking to inner lips (right).

D. Teeth processing

Some speech therapy exercises require to distinguish situations where the teeth are clenched or open. This is equivalent to detecting a gap between the upper and lower teeth. There are works oriented on teeth detection [17] however we have implemented our own solution, based on image analysis. Firstly, the input mouth image is rotated (if necessary) and cropped to get a stable view of the mouth. The processing takes place in a non-linear color space, specifically designed to emphasize the flesh tones [18]. The image is then masked using facial landmarks on the inside of the lips (Fig. 2) and thresholded to obtain a mask of the teeth (Fig. 6). We then compute the sum of each row and threshold this function (by 1/6 or 1/4 of the maximum depending on the particular

TABLE I
STATISTICAL EVALUATION OF EXERCISES. SEE DESCRIPTION IN SEC. IV

Exercise	Therapists – cross			App x Therapists			Theoretical limits				Succ. Rate
	Hits	from	%	Hits	from	%	Min Hits	Max Hits	Min %	Max %	
Tongue L/R	962	1416	68	118	180	66	43	137	24	76	**0,80**
Tongue U/D	1006	1464	69	70	117	60	32	85	27	73	**0,72**
Cheeks L/R	1036	1272	81	122	192	64	24	168	13	88	**0,68**
Cheeks both	1110	1368	81	122	219	56	28	191	13	87	**0,58**
Smile (sml)	1191	1539	77	157	234	67	40	194	17	83	**0,76**
Revert sml	1015	1491	68	119	219	54	58	161	26	74	**0,59**
Crooked sml	1144	1368	84	138	237	58	20	217	8	92	**0,60**
Opened sml	1171	1491	79	110	270	41	42	228	16	84	**0,37**
Sml w teeth	1154	1464	79	126	243	52	39	204	16	84	**0,53**
Kiss	1227	1491	82	115	219	53	24	195	11	89	**0,53**

exercise) so as to ignore possible reflections occurring on lips or tongue. This procedure results in a binary vector indicating rows with visible teeth. To detect a gap between the upper and lower teeth, the distance between the descending and the ascending edge (found by forward differences) is computed, if they are present. To successfully perform a speech therapy exercise with open teeth, the minimum predefined width of the gap has to be maintained.

IV. STATISTICAL EVALUATION

To validate the results, we have recorded 31 videos of patients practising all 10 exercises with 3 repetitions for each exercise. Speech therapy specialists were divided into groups and each group evaluated all exercises of three chosen patients to reduce time needed for evaluation. All the videos were also evaluated by our software. However, a pairwise comparison of the scores given by the individual speech therapists showed a low level of agreement, see Fig. 7. These findings show that two different specialists can have two opposing opinions and we can also see that each therapist focuses on different aspects of the exercises. This knowledge leads us to incorporate more features to evaluate in the future version. Moreover, every group of experts evaluated different patients. To deal with these problems, accuracy was computed separately for each exercise through all the therapists, see Table I.

The table is based on simple statistic based on the number of agreement among the therapists for particular repetition of the exercise. The columns represent: Therapists - cross shows number of agreements among all therapists (Hits), maximal number of possible agreements (from) and percentage. Apps x Therapists shows number of agreements between the application and each therapists (hits), maximal number of possible agreements (from) and percentage. Theoretical limits show minimal possible number of agreements (Min Hits), maximal possible number of agreements (Max Hits) and their percentages to compute success rate (Succ. Rate) defining if the application is in agreement with the majority of the therapists (close to 1) or in a disagreement (close to 0). The Success rate simply scales the accuracy of the application (Percentages in Apps x Therapists) based on the theoretical limits. The limits take into account situations where the therapists were in a disagreement and thus the application cannot be in perfect agreement with all of them.

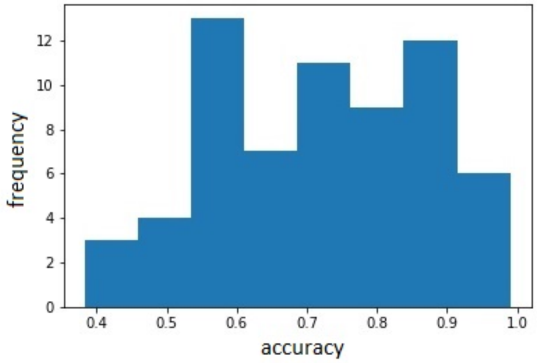
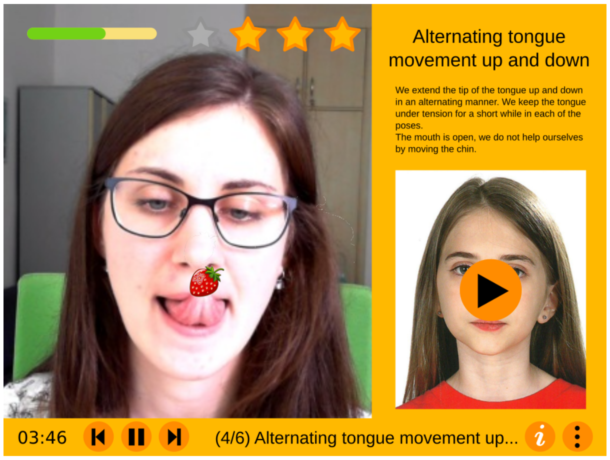


Fig. 7. Accuracy histogram for pairs of speech therapists showing low correspondence between experts.

V. ASSISLT APPLICATION


Alternating tongue movement up and down

We extend the tip of the tongue up and down in an alternating manner. We keep the tongue under tension for a short while in each of the poses.
The mouth is open, we do not help ourselves by moving the chin.

03:46 (4/6) Alternating tongue movement up...

Fig. 8. Screenshot from the ASSISLT application while performing a tongue exercise. The cues in the camera view window show the progress bar and a total number of repetitions (the stars). On the right side, an assistant panel is displayed with the exercise name, its description, and guidelines, and with a short video showing the correct procedure. The strawberry icon motivates a patient and illustrates the desired tongue tip position.

Fig. 8 shows the application with the main window showing the webcam stream with the patient practicing an exercise. The camera stream is enriched with the augmented reality (the strawberry icon) which motivates a patient and illustrates the desired tongue tip position. The stars in the right top corner indicate the number of successful repetitions of a given exercise. The green-yellow progress bar illustrates achieved and required duration of the exercise. The bottom panel shows the total exercise time, control buttons (pause, skip), the name of the current exercise, and its position in the sequence. A patient can unroll exercise details in the assistant panel, as is demonstrated. There are a short exercise description and guidelines together with an instructional video. In the bottom bar, the difficulty of the exercise can be adjusted. The patient performance can be saved for further assessment by a speech therapist.

We have introduced augmented reality to guide and motivate users to practice regularly. Fig. 8 shows the guidance using the inserted object (a strawberry icon), an example of a static method showing an icon image in the desired position of the tip of the tongue. A dynamic approach is represented by arrows appearing at the key part of the face, for example in the mouth corners when the patient is struggling with an exercise, e.g. closed smile, illustrating the desired lip movement. The last, an interactive approach is defined by removing image icons by tongue movement in the desired areas (licking virtual ice cream from the lips).

VI. Conclusion

A new software system ASSISLT is introduced, the aim of which is to support speech therapy for children and adults using deep learning approaches. The system is automatically evaluating an exercising patient recorded by the webcam. The detection and evaluation algorithms are based on image processing and neural network methods employed to detect a patient's face, lips, teeth, and tongue. The new set of features for exercises evaluation has been designed, implemented, and their threshold values were experimentally derived. The ASSISLT pipeline is shown together with chosen solutions for detection subtasks, and with examples for all implemented methods. The novel cheek, tongue and, teeth detectors are applicable in other application areas such as an alternative human-computer interface design.

References

[1] Z. Bílková, A. Novozámský, A. Dominec, Š. Greško, B. Zitová, and M. Paroubková, "Automatic evaluation of speech therapy exercises based on image data," in *International Conference on Image Analysis and Recognition*, pp. 397–404, Springer, 2019.

[2] Speech Tutor, "Speech tutor." www.speechtutor.org/app. Accessed: 2021-01-12.

[3] The Speech and Language Store, "Splingo's language universe." https://speechandlanguagestore.com/our-apps/splingos-language-universe/. Accessed: 2021-01-12.

[4] H. Hanks and C. Hanks, "Articulation station." http://littlebeespeech.com/articulation_station.php. Accessed: 2021-01-12.

[5] Z. Bílková, A. Novozámský, M. Bartoš, A. Dominec, Š. Greško, B. Zitová, M. Paroubková, and J. Flusser, "Human computer interface based on tongue and lips movements and its application for speech therapy system," *Electronic Imaging*, vol. 2020, no. 1, pp. 389–1, 2020.

[6] V. Kazemi and J. Sullivan, "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.

[7] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[8] C. Álvarez Casado and M. Bordallo López, "Real-time face alignment: evaluation methods, training strategies and implementation optimization," *Journal of Real-Time Image Processing*, vol. 18, no. 6, pp. 2239–2267, 2021.

[9] M. Frackiewicz, H. Palus, and K. Radlak, "Automatic cheek detection in digital images," in *Computational Vision and Medical Image Processing V - Proceedings of 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing, VipIMAGE 2015*, pp. 49–56, 2016.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd Ed.)*. USA: Academic Press Professional, Inc., 1990.

[11] . B. S. Priyanka, P.K., "Lip feature extraction and movement recognition methods: A review.," *International Journal of Scientific Technology Research*, no. 8, pp. 50–55, 2019.

[12] X. L. X. Gao, Y. Su and D. Tao, "A review of active appearance models," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 2, pp. 145–158, 2010.

[13] Z. Zhongxu, W. Aimin, and S. Lansun, "The color tongue image segmentation based on mathematical morphology and his model [j]," *Journal of Beijing Polytechnic University*, vol. 2, 1999.

[14] J.-q. Du, Y.-s. Lu, M.-f. Zhu, K. Zhang, and C.-h. Ding, "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.

[15] A. Sage, Z. Miodońska, M. Kręcichwost, J. Trzaskalik, E. Kwaśniok, and P. Badura, "Deep learning approach to automated segmentation of tongue in camera images for computer-aided speech diagnosis," in *Information Technology in Biomedicine*, pp. 41–51, Springer, 2020.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[17] J. Lee, S.; Kim, "Gingiva and facial landmarks for 2d digital smile design using real-time instance segmentation network," *J. Clin. Med.*, no. 11, p. 852, 2022.

[18] A. Novozámský, J. Flusser, I. Tachecí, L. Sulík, J. Bureš, and O. Krejcar, "Automatic blood detection in capsule endoscopy video," *Journal of biomedical optics*, vol. 21, no. 12, p. 126007, 2016.