IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2022, accepted 28 August 2022, date of publication 2 September 2022, date of current version 12 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3203572

RESEARCH ARTICLE

4D Multimodal Speaker Model for Remote Speech Diagnosis

MICHAL KRECICHWOST, AGATA SAGE, ZUZANNA MIODONSKA, AND PAWEL BADURA
Faculty of Biomedical Engineering, Silesian University of Technology, 41-800 Zabrze, Poland

Corresponding author: Pawel Badura (pawel.badura@polsl.pl)

This work was supported in part by the Polish National Science Centre (Hybrid System for Acquisition and Processing of Multimodal Signal in the Analysis of Sigmatism in Children), Poland, under Grant 2018/30/E/ST7/00525; and in part by the Polish Ministry of Science and Silesian University of Technology under Grant 07/010/BKM22/1015 (BKM-528/RIB1/2022).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Biomedical Committee at the Academy of Physical Education in Katowice under Application No. 3/2021.

ABSTRACT This paper presents a concept of a 4D multimodal speaker model (4D-MSM) for asynchronous remote speech diagnosis. Recording and archiving diagnostically significant articulation material remain an issue in computer-aided speech diagnosis. Therefore, we propose a workflow for preparing and storing reliable and easily interpretable multimodal data regarding pronunciation. According to our assumptions, data acquisition should be non-invasive, comfortable for both the patient and therapist, not interfere with the articulation process, and provide essential data of high quality. We developed and employed a dedicated device, obtaining a 15-channel spatially distributed audio signal and stable stereovision stream from two cameras focused on the lower part of the face. Our framework for data preprocessing covers digital beamforming of the multichannel audio signal, audio-video synchronization, and segmentation of words in the audio signal. Then, we use stereo data to calculate and adjust the depth map and prepare point clouds. Simultaneously, we delineate the mouth in video frames using a dedicated semi-automated segmentation algorithm. The point clouds are then textured with the camera images with superimposed mouth regions. Finally, we add the audio track to constitute the 4D-MSM. In the paper, we show the concept and detailed specification of the model and present experiments to justify the methodology. Proposed 4D-MSMs may be employed in remote speech diagnosis for objectifying and archiving diagnoses, conducting asynchronous consultations, and documenting the progress in therapy.

INDEX TERMS Articulation data acqusition, audio-video processing, computer-aided speech diagnosis, remote speech diagnosis and therapy, stereovision.

I. INTRODUCTION

Understandable communication between people is the basis of society's functioning. Therefore, speech disorders constitute a significant social problem, especially for children. Poorly developed or disordered speech may cause rejection by peers, alienation, low self-esteem, and later difficulties in getting an adequate job [1]. Therefore, speech screenings in kindergartens and schools are crucial, as they allow for the early detection of communication difficulties [2]. Subsequently, the children in need can undergo a more profound speech and hearing diagnosis and benefit from appropriate therapy. This process depends mainly on the availability of speech and language pathologists (SLPs). In Poland, 6-year-old children should be provided with speech diagnosis and therapy in their preschools. However, still over 50% (7,789 of 15,552) of Polish kindergartens do not hire a speech therapist [3]. SLPs often work simultaneously as teachers, so it is impossible to provide profound therapy to all children in need, especially those younger than 6. For this reason, parents often decide to choose private consultations and speech therapy. Such solutions are available

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

mainly to people living in cities. Access to qualified speech therapists can be difficult or impossible in rural areas and small towns. The problem of worse availability of speech therapists in rural areas is a global problem also reported, e.g., in the United States [4] and Australia [5].

The popularization of remote consultations, diagnosis, and speech therapy is a recent attempt to solve the expert availability problem. Supported by telemedicine tools, this field has been consistently developing for years. Nevertheless, only the outbreak of the COVID-19 pandemic accelerated the process of opening the speech therapy community to new technologies. For many therapists and their patients, using computer tools was the only way to continue their work since 2020 [6].

A. STATE OF THE ART

Speech diagnosis and therapy can be assisted with various computer-based methods. Videoconferences of a therapist and a patient (and possibly additional external experts) can be easily conducted with standard teletransmission tools [4], [7]. Internet platforms providing the patient with access to speech therapy exercises are technologically one step further. In some cases, the therapist can adjust the therapy programming module apart from standard access to the patient's performance or statistics [8], [9]. The gathered diagnostic material can be subjected to comprehensive analysis with automated computerized methods to provide measurable parameters of speech and hearing. Finally, the data can be used to estimate the probability of selected disorders (e.g., [10], [11], [12], [13], [14], [15]).

Remote speech therapy and diagnosis can be carried out synchronously, with the simultaneous presence of the speech therapist and the patient, or asynchronously based on mutually transmitted data [16]. Synchronous mode usually employs videoconferencing software activated on a laptop or smartphone. It is considered comparable to stationary therapy, e.g., Coufal *et al.* [17] found no significant differences between the therapeutic effects in a group of over 1,700 children with dyslalia. Other studies reported similar conclusions [4]. Raman *et al.* [7] described synchronous remote speech screening in rural India, demonstrating comparable effects of in-person and telemedicine studies. Despite the significant development of the field, both technical issues (computer equipment access and stable Internet connection) and the skeptical attitude of some speech therapists remain an issue [6]. Hence, the ease of installation, user-friendliness, and reliability of the designed solutions are of particular importance [18].

Asynchronous mode covers primarily therapy, not the diagnosis. It is employed mainly in cases where the diagnosis is already known, and the patient performs online exercises ordered by the SLP. Multiple studies describe asynchronous remote speech therapy based on Internet platforms containing multimedia training material for both children [8], [9] and adults with various disorders (e.g., the Polish platform "Afast! Say it" for aphasia patients [19]).

Computer-aided speech diagnosis (CASD) is a less explored topic. Automated diagnosis based on speech mainly concerns the detection of early signs of Parkinson's disease [10], [11], [20] or autism [21]. There are reports on the detection of mispronunciations and pronunciation disorders [13], [22], [23], nasality [24], stuttering [12], and automated recognition of incorrect pronunciation patterns within selected pathology [14], [25]. Other papers also describe methods for aiding speech therapy using audio and video processing [26], [27], [28], [29]. However, the proposed solutions feature two main drawbacks: they rely on relatively small groups of speakers, and they are rarely implemented in practice so far.

To document the patient's pronunciation, speech therapists often use notes or videos recorded with a smartphone during the examination. They can use such recordings to verify the patient's diagnosis, consult questionable issues with other experts, and monitor the therapy progress over time. Such a solution has several drawbacks. Simultaneous video recording with the phone and examining the patient is inconvenient for the therapist. The image is often unstable, and the data covers only short parts of the examination, so not all essential elements of the articulation are captured. In turn, recording using a tripod reduces the quality and diagnostic content of the data due to the greater distance between the camera/microphone and the articulators. The lack of convenient and reliable techniques for registration, archiving, and sharing of articulation data forms a significant gap in the current speech diagnosis and therapy, both remote and stationary.

The articulation registration methods proposed in the literature include electromagnetic articulation (EMA) [30], [31], electropalatography (EPG), [32], [33], electromyography (EMG), and various imaging methods (computed tomography, magnetic resonance imaging, ultrasound, fluoroscopy). These methods are expensive and hardly available and, in most cases, also invasive, so their use impacts the patient's pronunciation. Another disadvantage is also the problematic interpretability of the data. Without appropriate software and expert knowledge of the measurement specificity, it is impossible to conclude from the obtained results.

During an in-person speech diagnosis, a speech therapist can analyze how sounds are produced using the sense of hearing, touch, and sight. A single-camera view hardly allows for comprehensive observation of the speaker's articulation organs during pronunciation. A stereovision system enables 3D mapping of the surroundings thanks to a point cloud based on the images from a pair of cameras. Stereoscopic vision is the most natural way for humans to perceive three-dimensional images, as it allows feeling the depth of the observed scenes [34]. 3D mapping finds many applications, i.a., in monitoring, tracking, robotics control, terrain reconstruction based on aerial photography, or generating models for virtual reality [34], [35], [36]. Therefore, the use of spatial animations generated based on stereoscopic data collected while speaking may be a promising direction for articulation archiving. Steiner *et al.* [37] proposed to process the EMA measurements as motion-capture data. The animated 3D models presented the tongue motion during articulation. Busso *et al.* [38] reported high-quality speaker models produced using markers attached to the face. Xie *et al.* [39] generated three-dimensional VSA (visual speech animation) face models based on a video. However, there are no reports that this idea has been implemented or used to present articulation data. Also, the literature review suggests that, at the moment, there are no non-invasive methods for the acquisition, archiving, and visualization of 3D articulation data. Such a solution could immensely increase the possibilities of asynchronous remote speech diagnosis and consultations.

B. AIMS AND SCOPE OF THE CURRENT STUDY

In our previous works, we described our portable device that aims at a spatial recording of the speech [40]. The device is placed on the patient's head. It does not affect the articulation process but slightly reduces the child's face visibility for the therapist. We used 15 spatially distributed microphones to acquire the data and reported our analyses and results on sigmatism detection and recognition in various setups [14], [15], [25]. However, as that device recorded audio signals only, specific information on articulatory movements in time could not be obtained.

In this paper, we propose a workflow for generating a 4D multimodal speaker model (4D-MSM) for remote speech diagnosis and therapy. We developed an enhanced version of the acquisition device to record speech signals synchronized with a stereovision stream of the articulators. The 15-channel microphone array is supported by two cameras that cover the image of the lower part of the speaker's face. The data are then processed in a novel approach to prepare a virtual, multimodal representation of the speaker, which provides new opportunities for remote speech diagnosis and articulation archiving. The paper presents a set of experiments concerning the image analysis and parameter settings for the point cloud generation. We also provide exemplary 4D-MSM animations available as a supplementary material.

The paper is structured as follows. After the introduction in Section I, we present our 4D-MSM in Section II. That covers the description of a dedicated multimodal data acquisition device and a whole audio/video data processing workflow that produces the model (digital beamforming, synchronization, word extraction in the audio signal, mouth segmentation in the video stream, generation and texturing of point clouds from disparity maps, and combining the data to a 4D-MSM). Section III describes the speech therapy examination and data acquisition protocol as well as a series of experiments justifying our 4D-MSM generation. First, we quantitatively assess the mouth segmentation algorithm. Then, we describe, illustrate, and qualitatively validate the parameter setting process in multiple stages of the point cloud generation and adjustment. The essential part of the paper is included in Section IV, where we profoundly discuss our concept and provide perspectives for its application and development. Finally, Section V concludes the study.

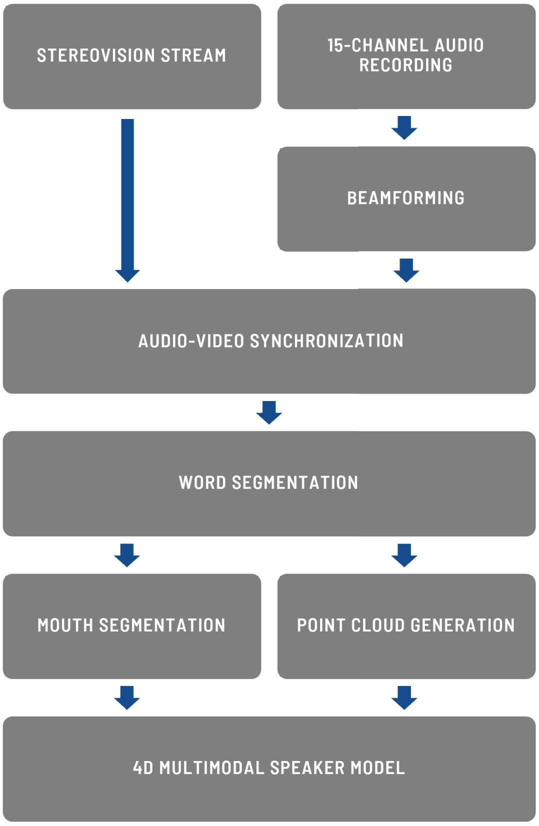


FIGURE 1. General scheme of the 4D multimodal speaker model generation.

II. METHODS

Our 4D multimodal speaker model is prepared according to the general scheme presented in Fig. 1. The method relies on audio and video data recorded using a dedicated multimodal data acquisition device. We use a newly designed device based on our experiences with the former equipment [40]. Double-camera stereovision module is the main addition, though we have also redesigned other features for feasibility and data quality. More details are given in Section II-A. The processing starts with digital beamforming applied to the spatial audio signals from 15 channels. Then, both audio and video paths are connected with the time synchronization module. After that, we segment individual words based on the audio signal. These three procedures constitute the preprocessing stage of our method. Then, we simultaneously generate a cloud of points based on videos from two stereo cameras and perform mouth segmentation in the video frames within words. Finally, we texture the point cloud using the left camera image and highlight the mouth area. The entire animated model (point cloud, video, segmentation mask, audio) of a single word is stored in a file for further analysis.

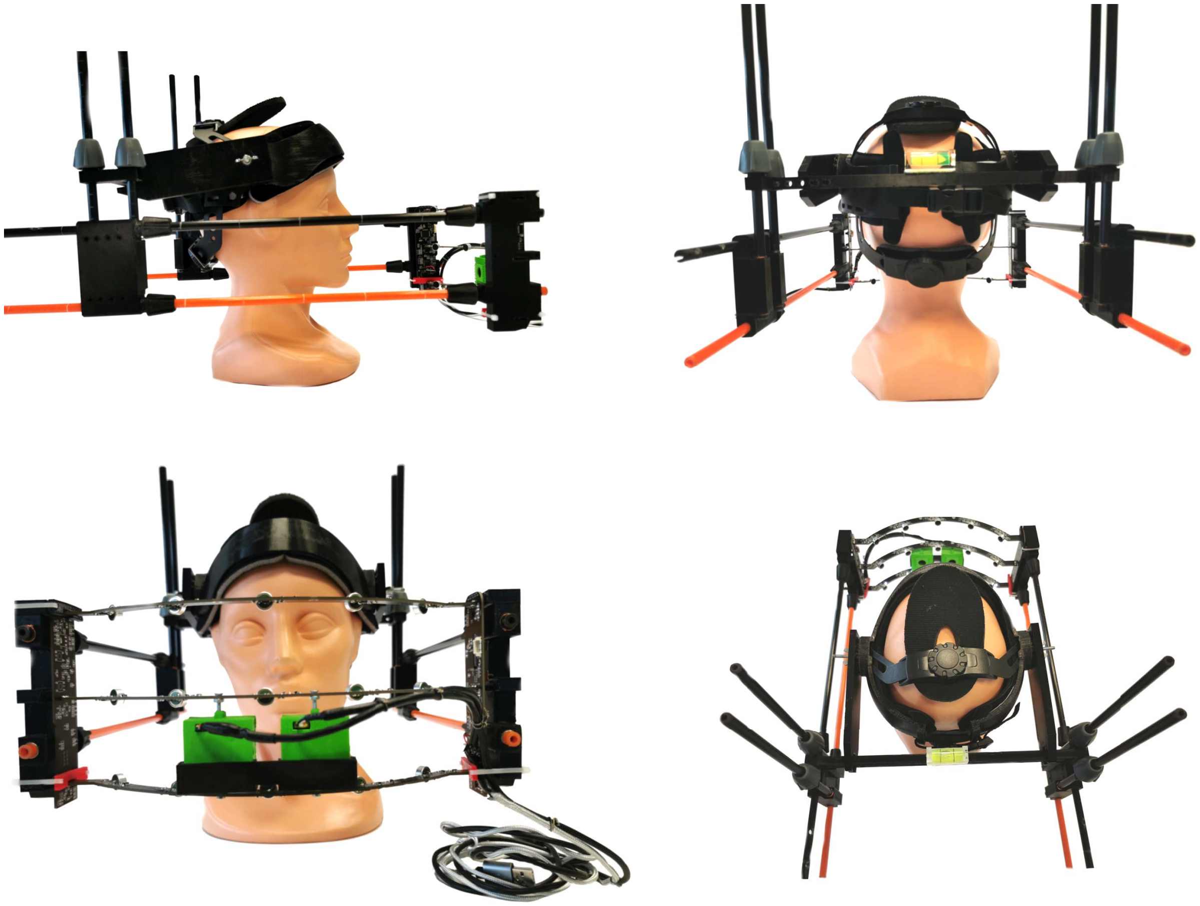


FIGURE 2. Illustration of the multimodal data acquisition device.

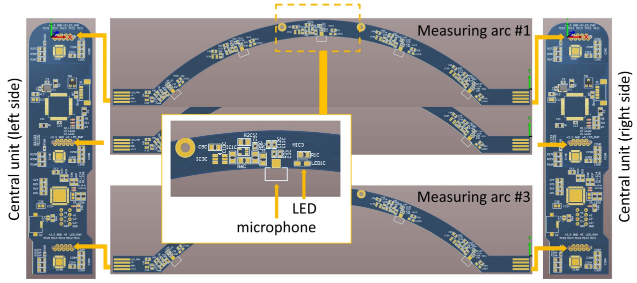


FIGURE 3. Illustration of a central unit and measuring arcs.

A. MULTIMODAL DATA ACQUISITION DEVICE

As mentioned, the concept of the acoustic mask described in [40] was verified and upgraded by a stereovision module. The current multimodal device is presented in Fig. 2.

The device consists of a central unit (CU) and three measuring arcs (MA). CU is powered by 5 volts and communicates with the computer via the USB interface. The MAs are connected to the CU as they exchange data using the serial peripheral interface (SPI). Two printed circuit boards (PCB) of the CU are also a mechanical frame for measuring arcs (note the illustration of the device's PCBs in Fig. 3).

Each MA records an audio signal using five microphones WM-61a [41] with omnidirectional characteristics, each equipped with a preamplifier TS472 and an amplifier TLV6741. Fifteen microphones form a 3 × 5 semicylindrical array with a 5-centimeter distance between the mics. The device records acoustic signals synchronized in time, with a sampling frequency of 44.1 kHz.

Finally, two cameras (Arducam 8MP 1080P Auto Focus [42]) are installed between two bottom MAs, constituting a stereovision optical system. To illuminate the speaker's face and increase the quality of the video, each MA is equipped with LEDs. With such a setup, we get a direct, unobstructed, and relatively stable view of the articulators during pronunciation from a short distance (ca. 15 cm) regardless of the head movements.

Major data registration parameters of the device are given in Table 1.

The construction is made of light materials with good mechanical properties, mainly through rapid prototyping (3D printing). The element put on the head is equipped with additional sponges from the inner side to increase the speaker's comfort. We can easily adjust the position of the mobile part of the mask and the sensor's distance from the sound source.

TABLE 1. Technical parameters of the multimodal data acquisition device.

Device	
Number of audio channels	15
Sampling frequency	44.1 kHz
Number of cameras	2
Microphone	
Type	Panasonic WM-61A (electret)
Bandpass	20 Hz – 16 kHz
Sound pressure level	120 dB
Signal-to-noise ratio	62 dB
Sensitivity (1 kHz, 94 dB SPL):	-35 ± 4 dB
Camera	
Type	Arducam 8MP 1080P Auto Focus
Resolution	640×480 VGA
Frames per second	30

Although being adjustable, the device ensures mechanical stability during the registration.

We designed the device to enable repeatable interspeaker and intraspeaker data acquisition. For this purpose, we adjust the mask position on a subject's head by superimposing reference lines on the camera images (Fig. 4). We use them to align the stereovision viewpoint with the characteristic points of the face, e.g., the philtrum.

B. CALIBRATION OF STEREO CAMERAS

We estimated the extrinsic and intrinsic geometric parameters of the stereo system for calibration purposes. It was performed by finding the geometrical relationship between two cameras by observing the same point (Fig. 5) [43], [44]. We captured 342 shots of a template with known dimensions and geometry: a 6 × 7 chessboard pattern with a grid size of 0.9 cm. Each projection presented a different view of the chessboard regarding viewpoint position and angle.

The calibration was divided into two stages. First, we calibrated individual cameras (left and right) separately using the position of the vertices of the chessboard fields. Then, we determined the translation and rotation matrices between the cameras and obtained an average calibration error of 0.39 pixels [45]. The optical system calibration was performed using OpenCV library tools (OpenCV: Camera Calibration) [46].

C. DIGITAL BEAMFORMING AND AUDIO-VIDEO SYNCHRONIZATION

We applied digital beamforming to 15 audio signals recorded at different points in space. As a result, we obtained a single signal with an improved signal-to-noise ratio (SNR). We employed the delay-and-sum beamforming (DAS) [47] that can reduce noise coming from non-central directions (Fig. 6). Since that point, the methodology involves the beamformed single-channel audio signal.

The audio-video synchronization in time is done by adjusting the beamformed audio signal to match the video timescale. There are several reasons for the desynchronization (delays caused by the components, electronics, or multithreaded software), yet we found the time shift between signals relatively constant. We measured it in multiple experiments to be 670±78 ms (Table 2). The measurement is based on the difference in the start times of the two software threads for communicating with the device: one for handling the audio data stream and the other for the video data. For certainty, we employed an additional expert assessment of the synchronization outcome and confirmed that the automated approach provides correct synchronization. Thus, we apply the measured latency to the audio signal before combining it with the stereovision stream.

TABLE 2. Summary of video-audio delay measurements.

Speaker	Video-to-audio delay
Speaker #1	601 ms
Speaker #2	746 ms
Speaker #3	698 ms
Speaker #4	733 ms
Speaker #5	574 ms
Average	670 ± 78 ms

D. WORD SEGMENTATION IN AUDIO SIGNAL

One of the assumptions for the 4D-MSM is its association with diagnostically important speech segments. Thus, we use a framework for extracting words or possibly other sections related to speech therapy exercises. This step affects the following time-consuming procedures by avoiding unnecessary computations. For word segmentation in a beamformed signal, we employed the method described by Giannakopoulos [48] based on the statistical analysis of the acoustic spectrum. Word boundaries are then applied to indicate and extract video frames for further analysis (Fig. 6).

E. MOUTH SEGMENTATION IN VIDEO STREAM

We prepared a semi-automated framework for mouth segmentation in video frames (Fig. 7). Each frame is processed individually, though consecutive images use information from previous iterations. The workflow described below applies to a single camera stream.

The data preprocessing begins with reducing image dimensions by embracing the mouth area. We determine the region of interest (ROI) by indicating the middle point on the upper lip in the first frame. Since the mouth appearance does not vary widely among children, the size and position of the ROI in the first frame are constant regarding the seed point (80 × 120 pixel size; seed point in the middle of the ROI, horizontally, and in the 7/8 of the height, vertically). Note that the ROI size and location vary during iterating through video frames, as it follows the segmentation results from the previous image. The cropped ROI is subjected to preprocessing, including RGB color space suppression to grayscale by modifying the I3 feature for lips enhancement [49] and image filtering using morphological opening for reducing minor artifacts while preserving edges (disk-shaped structuring element with radius equal to 5).

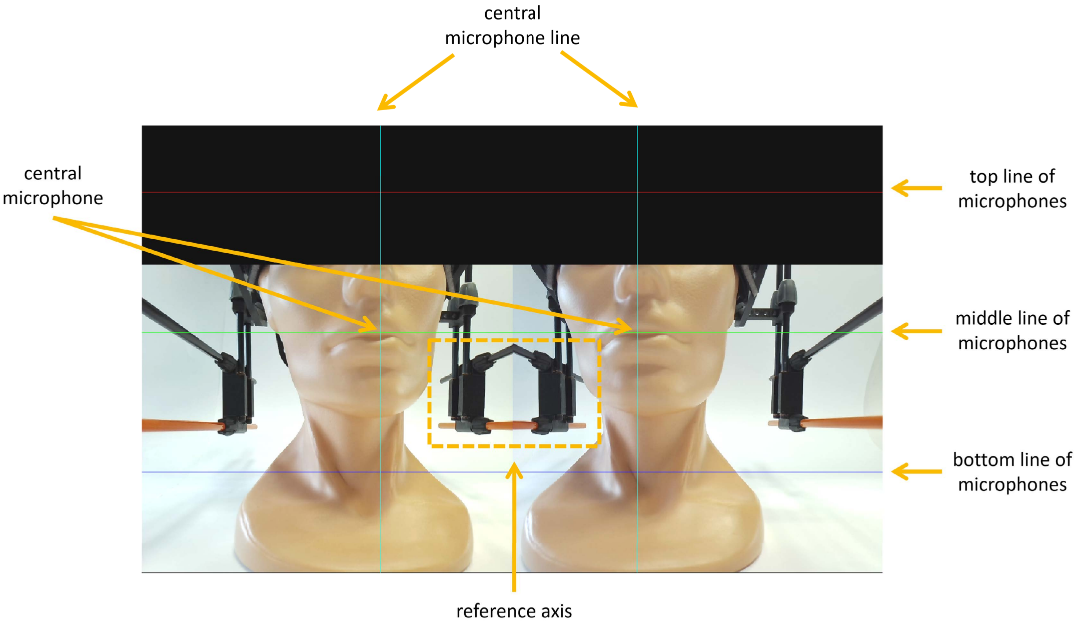


FIGURE 4. Illustration of the adjustment interface used to position the data acquisition device on the subject's head. Left and right images are produced by the stereo cameras.

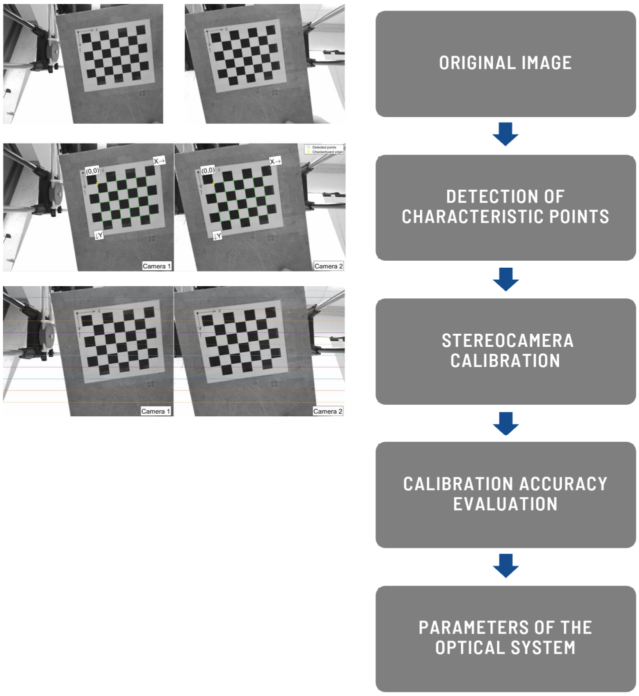

ORIGINAL IMAGE
DETECTION OF CHARACTERISTIC POINTS
STEREOCAMERA CALIBRATION
CALIBRATION ACCURACY EVALUATION
PARAMETERS OF THE OPTICAL SYSTEM
(0,0)
X→
Y
Camera 1
Camera 2

FIGURE 5. Workflow of the stereo cameras calibration.

Then, we transform the ROI intensities using a Gaussian fuzzy membership function [50]. The Gaussian mean and standard deviation follow the lips intensity and homogeneity retrieved from the ROI. The resulting fuzzy scene reinforces the mouth region and attenuates the background. The binarization followed by morphological corrections constitutes the initial contour for the fine segmentation using distance-regularized level set evolution (DRLSE) [51] over the fuzzy ROI scene. The segmentation result serves as the initial contour for the next frame. It also adjusts the corresponding ROI bounding box, assuming that the DRLSE segmentation can robustly chase the frame-to-frame differences. We determined the DRLSE parameters experimentally to $\alpha = -3.0$, $\lambda = 5.0$, and the number of iterations to 5.

F. POINT CLOUD GENERATION

The point cloud generation diagram is shown in Fig. 8. Based on the optical system parameters (Section II-B), we first rectify corresponding pairs of frames from both stereo cameras. The images are transformed so that the related epipolar lines become collinear and parallel to the horizontal edges of the frames. Rectification also significantly reduces the computational cost of the following alignment stages [52], [53]. Then, we determine the face ROI through manual delineation to preliminarily reduce artifacts in the resulting point cloud.

Then, we compute the disparity (depth) map from a rectified pair of grayscale stereovision frames by using the stereo semi-global block matching (StereoSGBM) algorithm [54]. StereoSGBM is one of the most widely used stereovision algorithms since the OpenCV library provides a fast and robust implementation [46].

Various matching errors can appear in the disparity map. They are usually concentrated in uniform texture-less areas, half-occlusions, and regions near depth discontinuities [55]. To reduce this effect, we apply filtering by using a left-right disparity-difference threshold and obtain a significant reduction of alignment errors. Another filtering technique used to restore the map continuity is the disparity weighted least squares filter (WLS) [56]. It removes holes

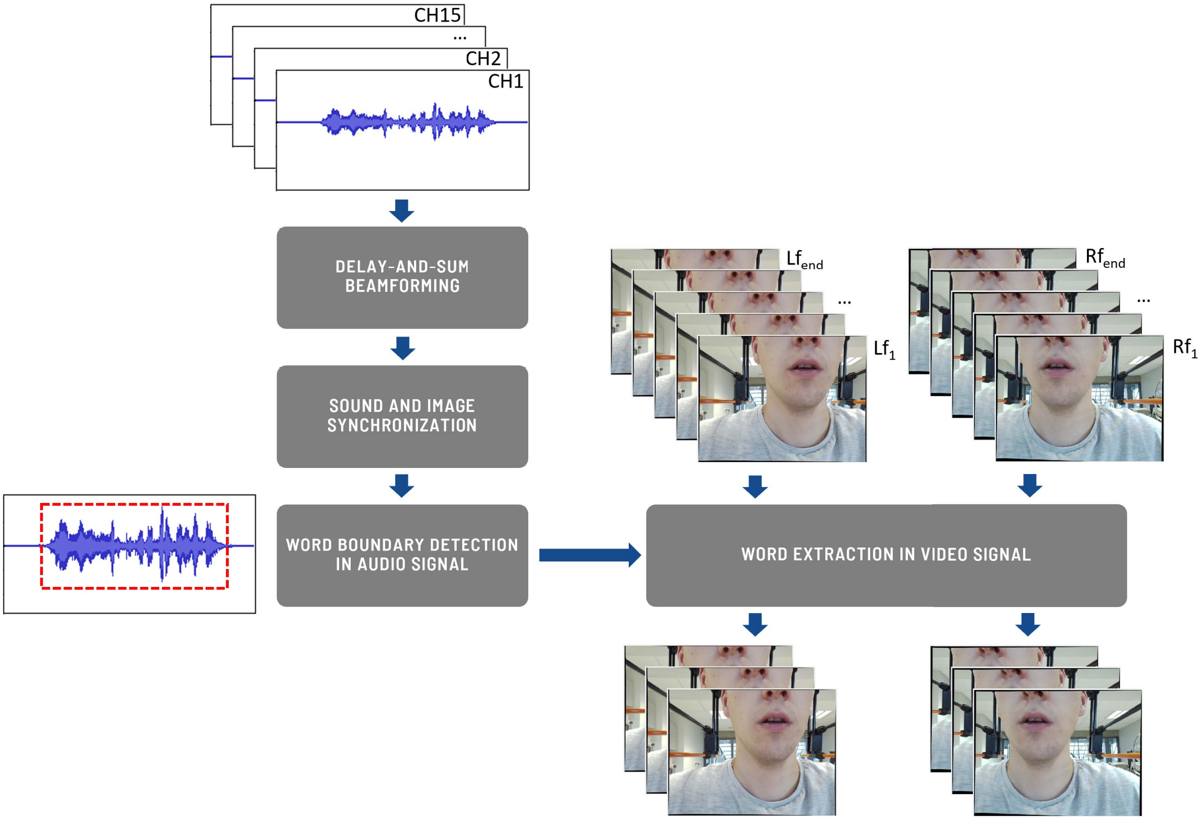

CH15
...
CH2
CH1
DELAY-AND-SUM BEAMFORMING
SOUND AND IMAGE SYNCHRONIZATION
WORD BOUNDARY DETECTION IN AUDIO SIGNAL
WORD EXTRACTION IN VIDEO SIGNAL
Lf_end
...
Lf_1
Rf_end
...
Rf_1

FIGURE 6. Workflow for preparation of audio and video data for the 4D-MSM. The scheme covers digital beamforming, audio-video synchronization, and word segmentation in audio signal.

from half-occlusions while preserving edges, as it calculates smoothing weights for pixels based on their isotropy and gradients [57], [58].

The depth map is transformed into the point cloud using camera parameters and the acquisition system geometry [36], [59]. In the final step, the cloud is limited to the face ROI and textured with the image data from the left camera. In a stereovision system, one of the cameras serves as a reference data source, and the other is a side camera [34]. The individual disparity map points correspond to the reference camera view. Here, we use the left camera as a reference.

G. 4D SPEAKER MODEL GENERATION

To generate a 4D-MSM, we combine the textured point clouds with the remaining data (Fig. 9). Each frame of the model consists of a single point cloud with a texture image and an overlaid mouth segmentation result plus an audio frame. The resulting 4D-MSMs can be stored as the .gltf files (graphics language transmission format binary file) [60] to allow flexible viewing angle or, for presentation purposes, as mp4 files.

III. EXPERIMENTS AND RESULTS

A. MATERIALS

We used the multimodal data acquisition device to record a dataset containing samples of the speech signal and video data in one of the preschools in the Silesia region. The recording team included two speech engineers and one SLP. Participating children were included in the study based on the inclusion criteria: age 4 to 7 years old and oral consent to participation in the experiment. Exclusion criteria included epilepsy and seizure states and ongoing respiratory tract infection. All speakers provided written consent to participate in the study, signed by their legal guardians. As a result, we obtained speech and video data from five speakers for this study (four girls and one boy aged 4 to 7). The study protocol was approved by the Biomedical Committee at the Academy of Physical Education in Katowice (decision No. 3/2021), as it followed all the required legal and ethical standards.

During the recording session, the child was invited to the room, and the recording team member presented the measuring device to them. If the child consented to participate in the study, the acquisition device was placed on their head and adjusted to fit securely. The team member made sure that the speaker was comfortable and proceeded with the recordings.

The recorded material consisted of two parts. In the first one, the speaker's task was to name the pictures presented on the screen. The selected pictures included everyday objects, professions, and animals that could be easily recognized and named by a preschool child. In the second part of the recording, the SLP asked the participant to repeat after them different facial expressions, including smiling and tongue

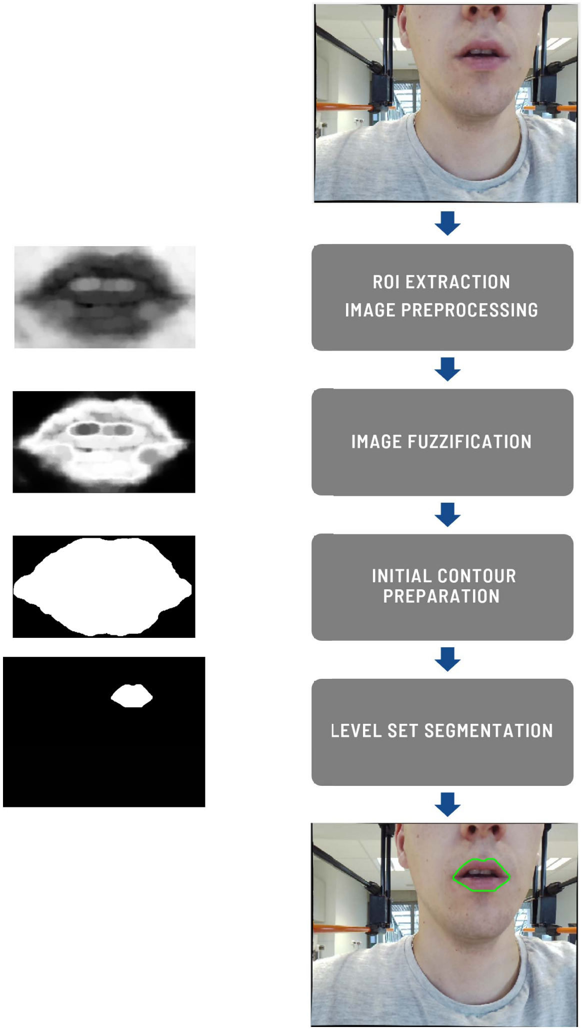


FIGURE 7. Mouth segmentation workflow.

TABLE 3. Summary of mouth segmentation performance.

Speaker	No. of frames	Camera	DI	IoU
Speaker #1	412	Left	0.99 ± 0.03	0.97 ± 0.05
		Right	0.99 ± 0.02	0.97 ± 0.03
Speaker #2	416	Left	0.98 ± 0.03	0.96 ± 0.05
		Right	0.98 ± 0.03	0.96 ± 0.05
Speaker #3	244	Left	0.94 ± 0.10	0.90 ± 0.16
		Right	0.90 ± 0.15	0.85 ± 0.20
Speaker #4	294	Left	0.98 ± 0.05	0.96 ± 0.08
		Right	0.97 ± 0.04	0.95 ± 0.07
Speaker #5	420	Left	0.97 ± 0.04	0.94 ± 0.08
		Right	0.97 ± 0.04	0.94 ± 0.08
All	1 786	Left	0.97 ± 0.05	0.96 ± 0.08
		Right	0.97 ± 0.07	0.95 ± 0.10

exercises. This set included movements that are useful for visual assessment of the articulators.

B. MOUTH SEGMENTATION ASSESSMENT

We evaluated mouth segmentation performance using Dice index (DI) and intersection-over-union (IoU) over a total of 3,572 images (1,786 for both left and right camera) taken from five speakers. Tab. 3 gathers both speaker-wise and overall results for either camera independently. Differences in metrics values between cameras are minor. Overall, the left camera segmentation performs slightly better (DI $= 0.97 \pm 0.05$ and IoU $= 0.96 \pm 0.08$ vs. DI $= 0.97 \pm 0.07$ and IoU $= 0.95 \pm 0.10$). Possible reasons for decreased effectiveness in some speakers (mainly Speaker #3) can be found in poor data quality, illumination issues, and rapid movements of the child's articulators and head.

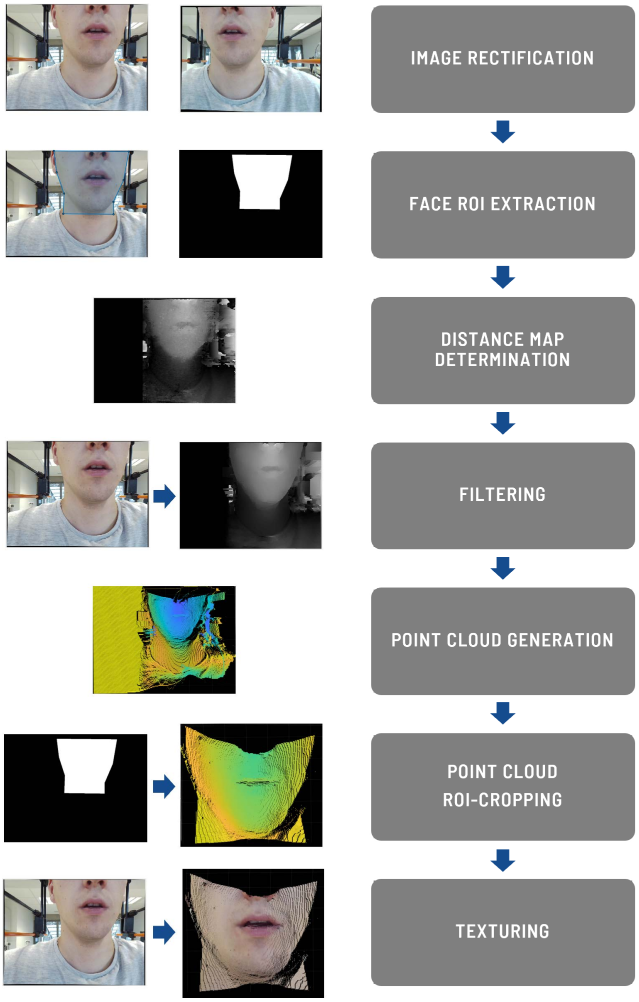


FIGURE 8. Diagram of the point cloud generation and texturing.

TABLE 4. Parameter settings for the StereoSGBM algorithm.

Parameter	Value	Tested values
Block size (BS)	3	1, 3, 5, 7, 9, 11, 13, 15
Max. disparity difference (d_{max})	128	1, 2, 4, 8, 16, 64, 128, 256
Speckle window size (SWS)	300	50, 300, 1000, 3000
Speckle range (SR)	32	16, 32, 48

C. PARAMETER SETTINGS FOR POINT CLOUD GENERATION

We conducted a series of experiments to select the appropriate StereoSGBM and WLS filtering parameters securing low

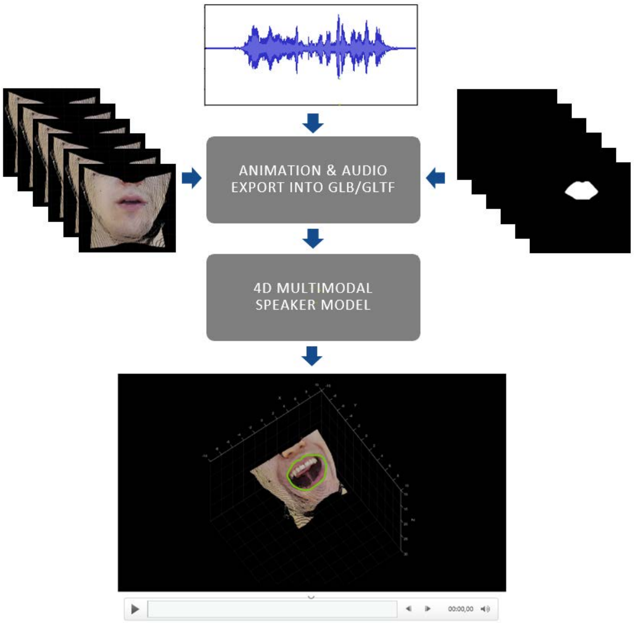

ANIMATION & AUDIO EXPORT INTO GLB/GLTF
4D MULTIMODAL SPEAKER MODEL

FIGURE 9. 4D multimodal speaker model workflow.

rates of depth map matching errors (Tables 4 and 5). With a relatively close distance between the stereo cameras and mouth (15 cm) and a known fixed arrangement of the optical system, we experimentally set the minimum disparity for the StereoSGMB to 125 and the disparity levels to 96 (maximum disparity at 221).

First, we investigated the alignment block size *BS* and, simultaneously, two penalty factors, P_1 and P_2, controlling the disparity map smoothness [54]. A small block size produces a detailed disparity map for a price of more matching errors. An increasing block size reduces the noise, but the smoothed disparity map can lose essential details, e.g., depth edges. We tested multiple *BS* values from the range of 1–15 (Table 4) and eventually set it to 3. Penalties P_1 and P_2 can depend on the number of chromatic channels N_{ch} and the block size [54], [61], and we used the proposed formulae in our study:

TABLE 5. Parameter settings for the WLS filtering.

Parameter	Value	Tested values
Regularization factor (λ)	1000	10, 100, 1000, 10000
Smoothing factor (σ)	1.5	0.2, 0.5, 1.0, 1.5, 2.0, 2.5

$$P_1 = 8 \cdot N_{ch} \cdot BS^2 \tag{1}$$

$$P_2 = 32 \cdot N_{ch} \cdot BS^2 \tag{2}$$

Disparity maps produced by different block sizes are shown in Fig. 10.

Then, we tested the maximum allowed difference d_{max} between the left and right disparity maps. It specifies the threshold in pixels above which the disparity is filtered from the resulting map. We obtained the optimal value of 128 from a 0–256 range. The d_{max}-dependent illustrations are shown in Fig. 11.

We also verified two properties of the speckle filter used to handle noise blobs: the speckle window size *SWS* and speckle range *SR*. *SWS* is the window size for smooth disparity regions to be checked for noise speckles, whereas *SR* specifies the maximum disparity variation within a connected component. Fig. 12 presents the effects of disparity map filtering with pairs of *SWS* and *SR* from 50–3000 and 16–48 ranges, respectively.

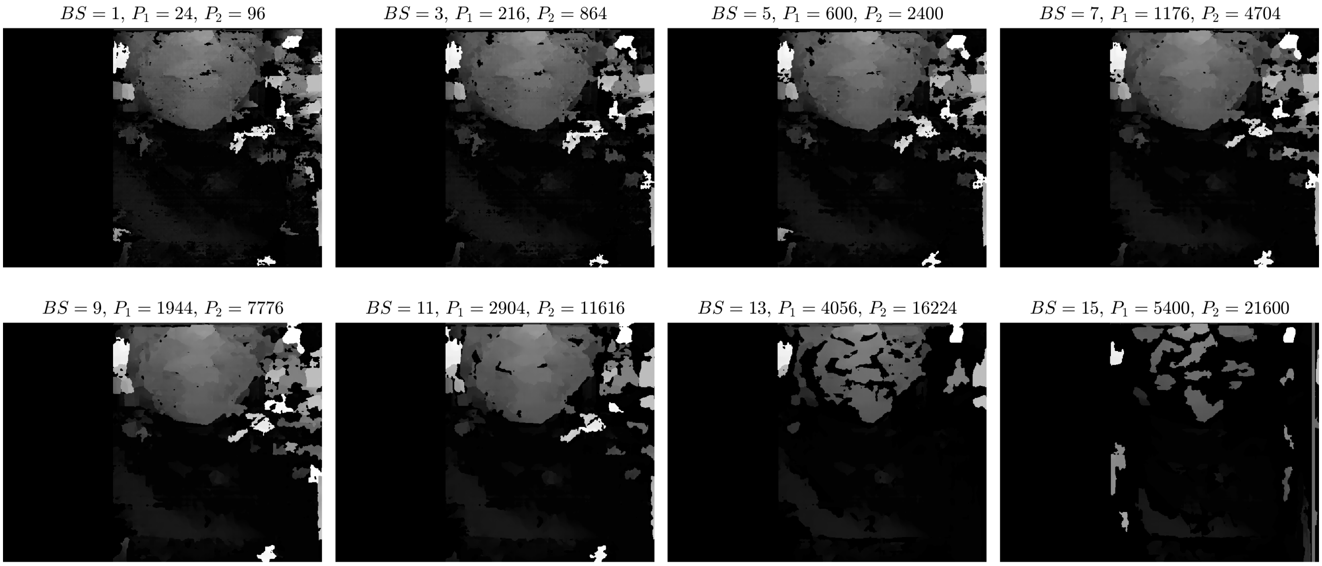

$BS = 1$, $P_1 = 24$, $P_2 = 96$
$BS = 3$, $P_1 = 216$, $P_2 = 864$
$BS = 5$, $P_1 = 600$, $P_2 = 2400$
$BS = 7$, $P_1 = 1176$, $P_2 = 4704$
$BS = 9$, $P_1 = 1944$, $P_2 = 7776$
$BS = 11$, $P_1 = 2904$, $P_2 = 11616$
$BS = 13$, $P_1 = 4056$, $P_2 = 16224$
$BS = 15$, $P_1 = 5400$, $P_2 = 21600$

FIGURE 10. Disparity maps produced by different block size *BS* and corresponding penalties P_1, P_2.

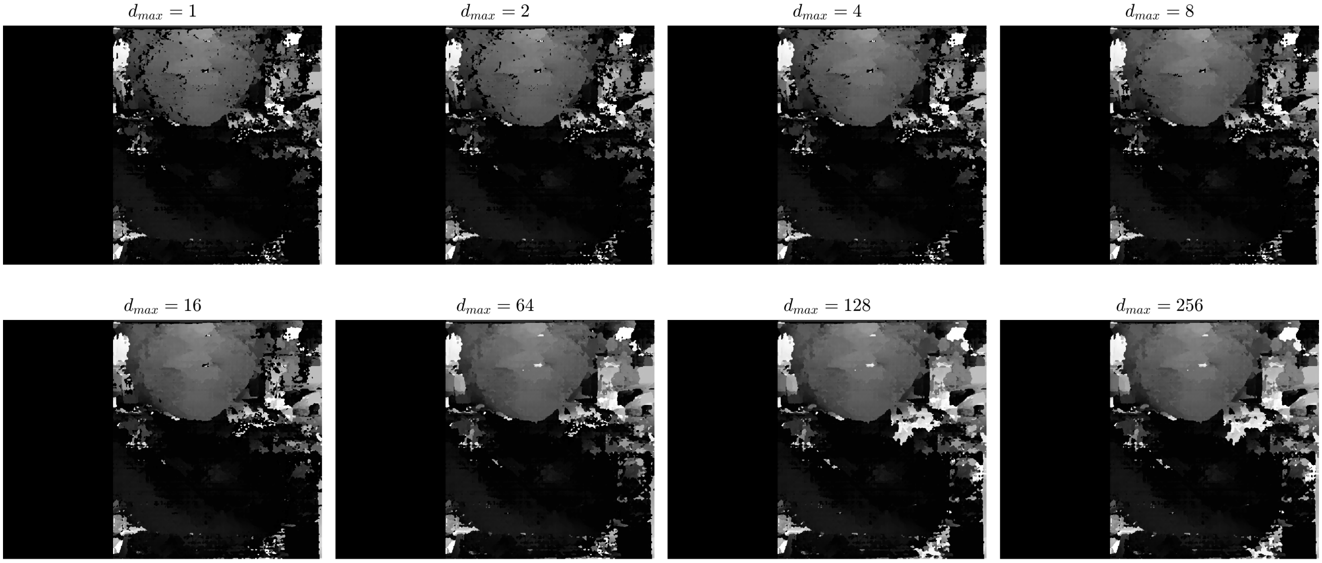


FIGURE 11. Disparity maps produced by different maximum disparity distance d_{max}.

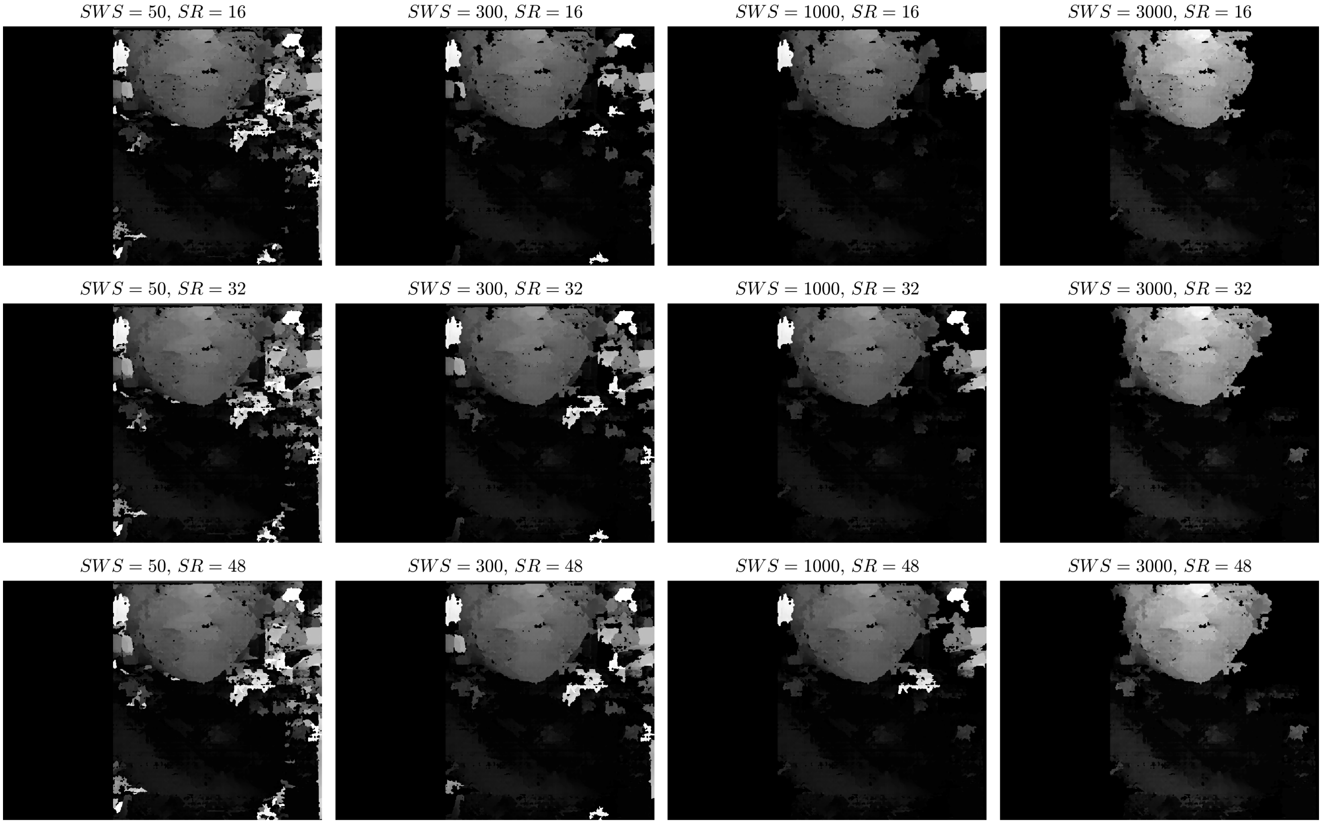
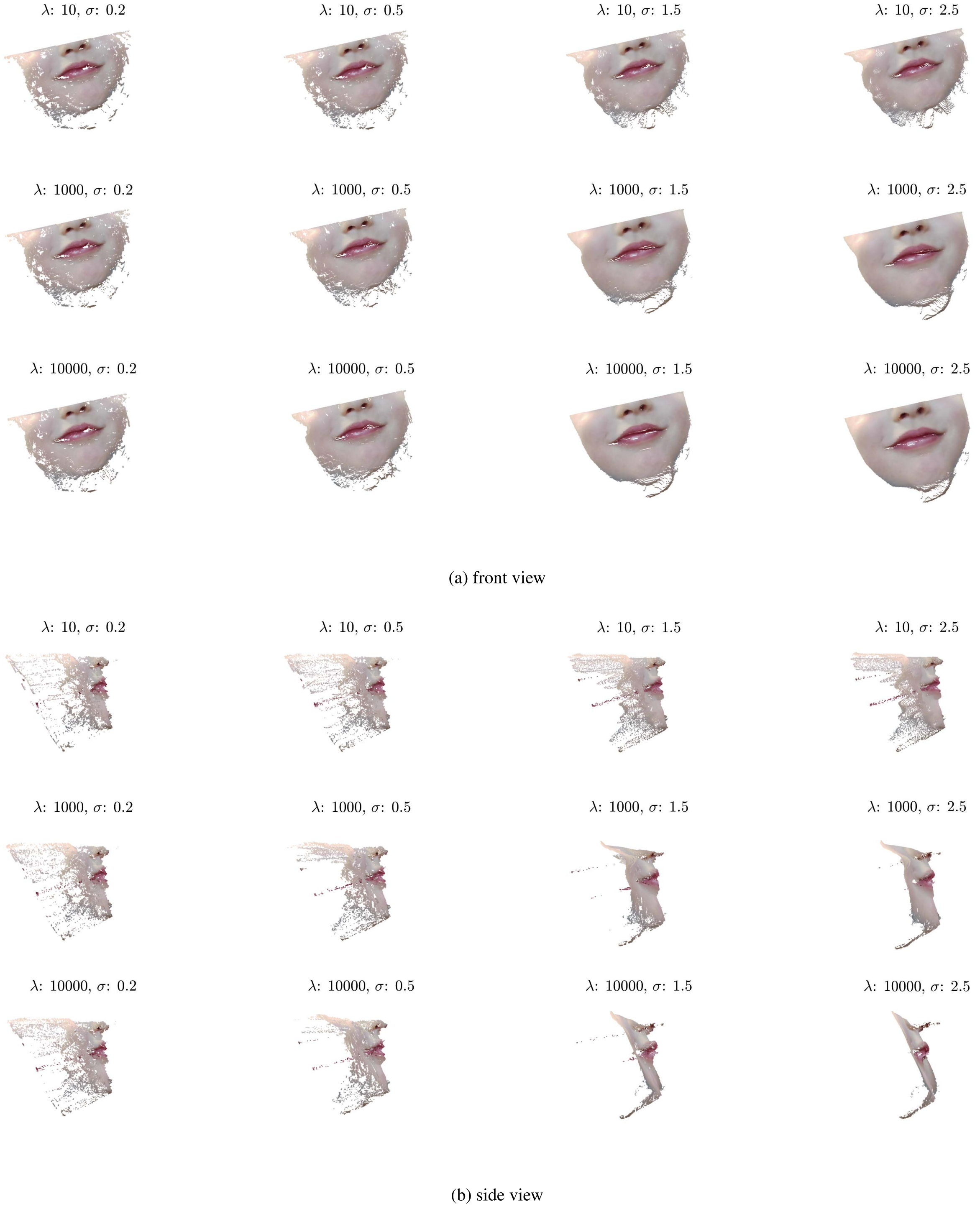


FIGURE 12. Disparity maps produced by different speckle window size *SWS* (fixed in columns) and speckle range *SR* (fixed in rows).

Verification of the WLS filtering involved two main parameters: λ and σ. λ controls regularization during filtering to match the disparity map edges to the image edges. The smoothing factor σ sets the filtering sensitivity to image edges. Large σ may cause disparity leakage through low-contrast edges, while small σ leaves noise and textures in



(a) front view

(b) side view

FIGURE 13. Textured point clouds produced from disparity maps by different σ (fixed in columns) and λ (fixed in rows) in the front (a) and side (b) views.

homogeneous regions. Our experiments confirmed the above general rules. To show the effects obtained over disparity maps more clearly, we use two views of the point clouds eventually textured with the image of the speaker's face in Fig. 13. Parametrized double-filtering effects are shown in Fig. 14.

Based on the above experiments, we selected a set of parameters for calculating the disparity maps. Then, we assessed the algorithm's robustness by evaluating the generated disparity maps with and without WLS filtering for individual speakers. Since we operate on videos of real objects moving in time, it was not possible to directly

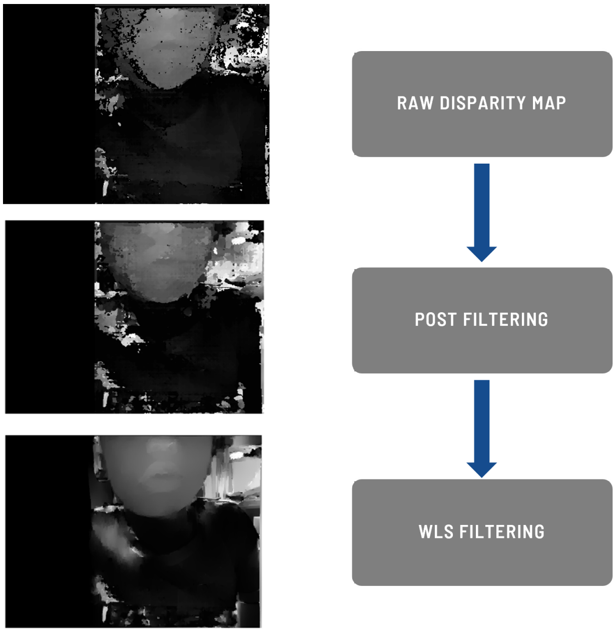


FIGURE 14. Illustration of the disparity map generation and filtering with parameters set in our experiments.

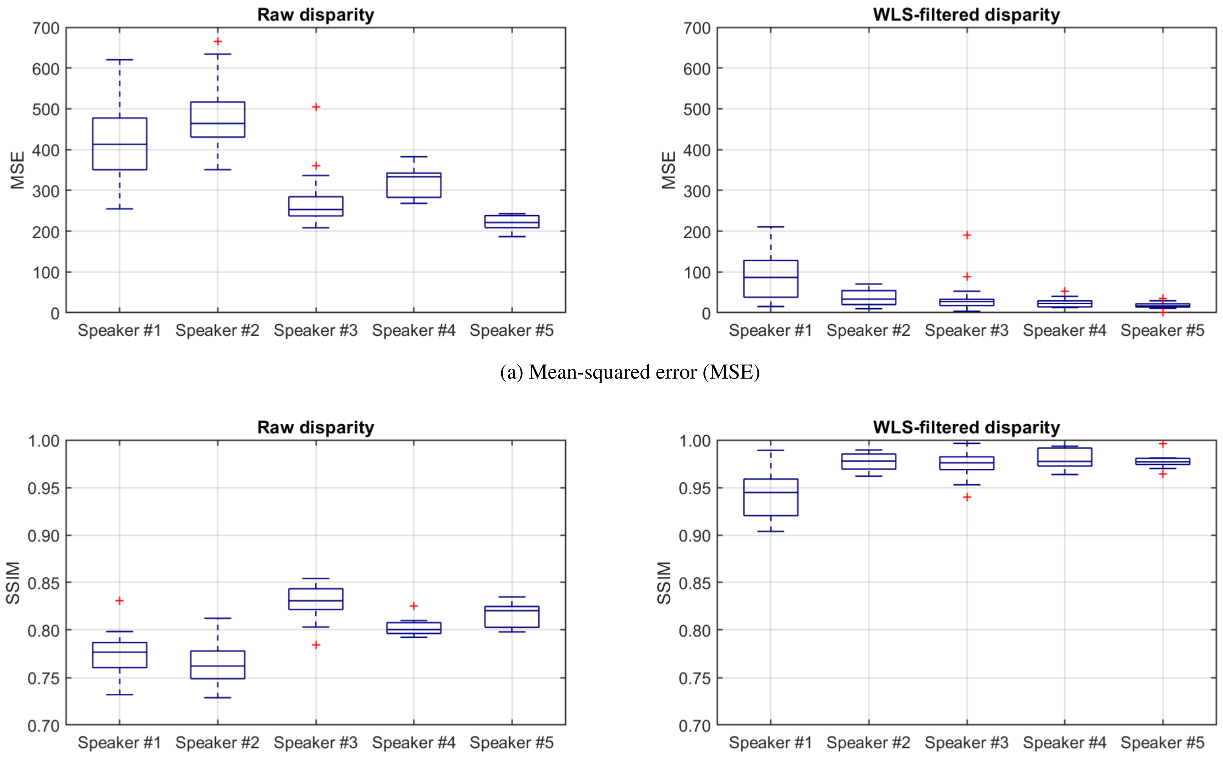
obtain reference images (ground truth) with methods proposed in the literature, e.g., by using data produced with Blender [62], [63]. Therefore, ground-truth images were prepared by an expert by manually removing artifacts and alignment errors for individual disparity maps. Then, using the reference images, we determined the mean-squared error (MSE) and the structural similarity index (SSIM) for every fifth frame for subsequent speakers [64]. Before calculating the errors, the images were cropped to the face area. The results are presented in Fig. 15.

Finally, some illustrations of a textured 4D-MSM are shown in Fig. 16. Note that some views of complete 4D-MSMs can be found as the supplementary online material.

IV. DISCUSSION

Despite the significant increase in the popularity and availability of telemedicine solutions in speech diagnosis and therapy during the COVID-19 pandemic, some problems remain unresolved [6]. Supporting speech therapy with remote exercises has already been well researched, and there are no significant differences in the effectiveness of this type of synchronous therapy compared to stationary treatment. However, doubts about the possibility of a reliable diagnosis without an in-person examination of the patient remain an essential issue [65].

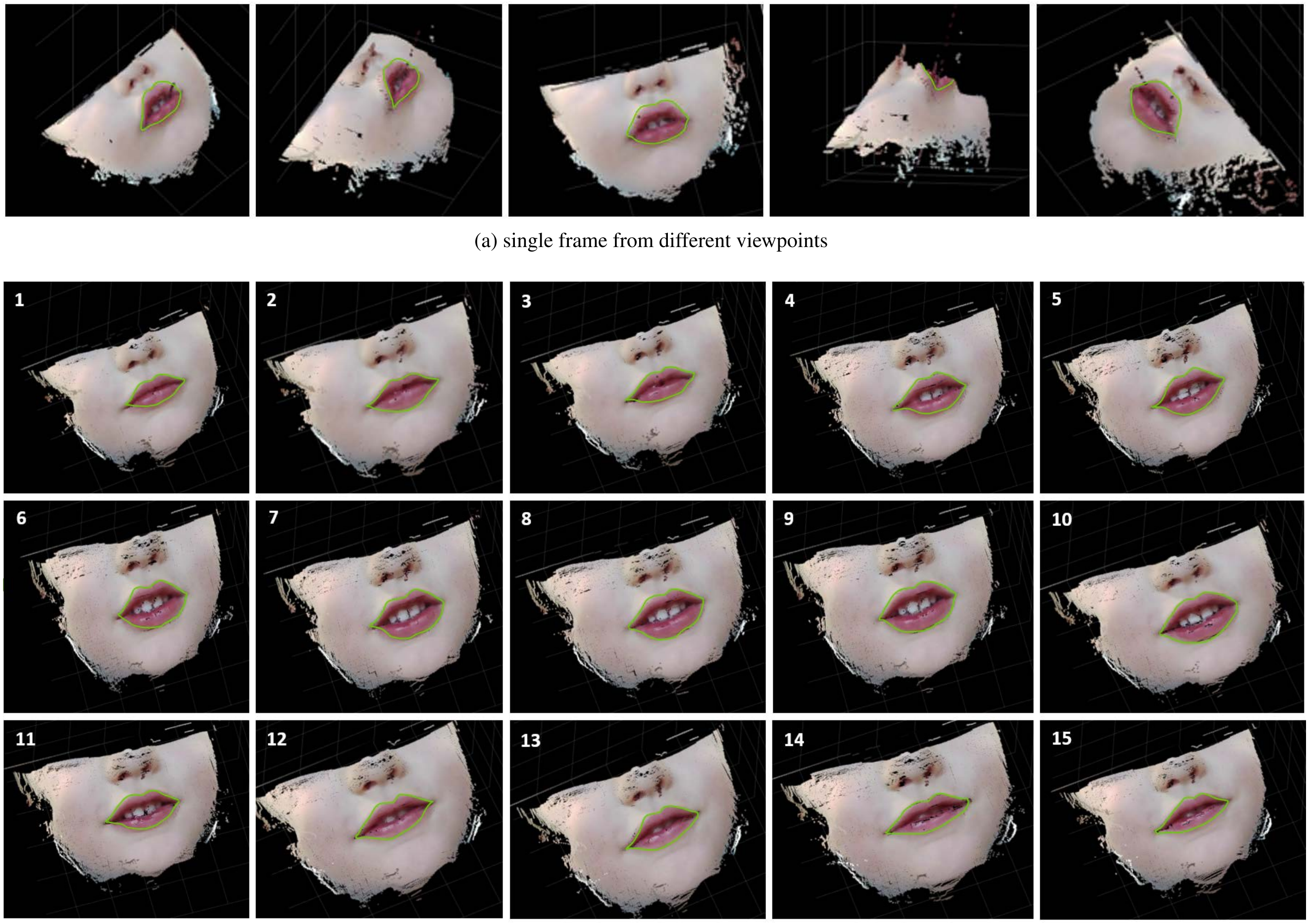
The proposed speaker model can be generated based on one short measurement session and then viewed and analyzed at any time. Thanks to that, additional verification and



(a) Mean-squared error (MSE)

(b) Structural similarity index (SSIM)

FIGURE 15. Summary of disparity map determination assessment. Mean-squared error (a) and structural similarity index (b) without (left) and with (right) WLS filtering. Each box covers 25th to 75th percentile interval with a median indicated by a central line. Whiskers refer to 1.5 times the interquartile range. Outliers indicated with red +.



(a) single frame from different viewpoints

(b) frame-by-frame animation showing the production of the vowel *a*

FIGURE 16. Illustration of the 4D-MSM in a child speaker: a single frame from different angles (a) and frames covering the production of a vowel *a* (b).

reevaluation of the diagnosis are possible, and more complex or questionable cases may be consulted with another specialist. Recorded material can be archived and constitute the basis for tracking the patient's progress over time. Archive models can support the diagnosis of more challenging health problems, which only show visible symptoms over time. The data acquisition device we use in this study can be considered non-invasive for articulation or ease of speaking. Also, we did not experience any problems with children's willingness to try the device on. Note that the same conclusions came from our previous study with the former version of the device, used to examine over 100 children aged 5–6 using the picture-naming protocol [40]. Both devices share main architecture concepts and solutions, e.g., removable sponges inside the mask or bicycle-helmet-like head mount.

The innovation of the proposed solution lies mainly in the possibility of documenting the depth of the articulator's image, which is not possible with the typical documentation of the diagnosis using videos recorded with a smartphone. Stereovision techniques and high-resolution dynamic spatial models of the speaker can provide diagnostically important information. A vital contribution to a remote diagnosis is brought by watching the movement of the speech organs from the front and side and listening to the high-quality denoised audio data. It is particularly important when observing the pathological features of the articulation.

The resolution and preservation of the actual depth of the model and the possibility of its free observation (from any viewing angle) in real-time are crucial for the speech diagnosis. Providing this kind of effect without the need for additional markers applied to the face opens new perspectives to the process, especially since it is considered a challenging issue [66], [67]. The number of depth levels and the precision of their extraction from the stereoscopic image depends on the cameras' resolution and the distance between the cameras' optical axes. Higher resolution provides more levels of depth for the price of increasing time consumption and stereo set expense. The greater distance between the cameras offers higher precision of location in space. However, too wide spacing of the cameras does not allow for the reconstruction of the close plan due to the disparity exceeding the camera's resolution. Our optical system allows for an accurate reconstruction of the face surface, including the mouth. Consequently, the observation angle range is limited while maintaining the appropriate number of details. We consider increasing the number of stereo cameras in future studies, likely allowing for a more accurate representation of the recorded speaker at any observation angle.

The segmentation of articulatory organs, e.g., mouth, lips, tongue, or the detection of the frenulum of the tongue along with a synchronized audio signal may be crucial in systems for screening child articulation. Quick, automated detection of abnormal movements of the tongue or lips during articulation can enable early detection of speech disorders and abnormalities in the articulation apparatus development. Our mouth segmentation method yields accurate results, with the Dice index at 0.97 for both left and right cameras. Several factors impact the video data quality, leading to lower segmentation metrics (e.g., Speaker #3 in Tab. 3). These include external lumination and shadows in the examination room, interspeaker differences in the anatomical structure of the bottom part of the speaker's face, lips color, or rapid head movements. With possible goals of extending the segmentation scope to other anatomical structures relevant to speech therapy, we consider employing the machine learning tools when developing the 4D-MSM in the future.

Based on the literature review, we conclude that there are no works on the Polish language combining articulatory features with those of a video image. For the development of CASD systems, linking the image features describing segmented speech organs with the acoustic signal features and articulatory cues becomes essential. The proposed 4D-MSM may provide input data for CASD systems based on image analysis and artificial intelligence. Robust segmentation of different articulators, which is the main direction for our future research, will enable the analysis of anatomical and physiological features, such as the mandible size and position, occlusion, or asymmetry in the lips position. Also, computer-aided analysis of the lingual frenulum behavior during speech exercises can be a promising direction, as the shortening of the frenulum (ankyloglossia) significantly reduces the tongue motor skills, influencing speech development, occlusion, and physiological functions, e.g., swallowing. The SLPs must have adequate experience and training to evaluate the frenulum, and they often order further consultation with an ankyloglossia expert. A computer system supporting the evaluation of the frenulum could significantly shorten the diagnosis and decision on possible cutting the frenulum (frenotomy).

In the future, we plan to create an online platform for therapists, teachers, and parents for remote and computer-assisted speech diagnosis and therapy using 4D-MSM. The multimodal data eventually processed with the artificial intelligence techniques will allow reporting on the speaker's articulation state. This solution can be used to conduct screening tests in schools and kindergartens efficiently. The model of the speaker generated and available on the diagnostic platform will enable further consultations by a multidisciplinary team, including orthodontists, neurologists, and physiotherapists. As a result, the in-depth diagnostic process can be accelerated, which is particularly important for children living in areas lacking appropriate specialists.

V. CONCLUSION

This paper presents the concept and framework for recording multimodal data and generating a 4D multimodal speaker model, which can be widely used in remote speech diagnosis and therapy. A novel device allows for repeatable registration of the multichannel acoustic signal and stereovision stream of the face part during the speech therapy examination. Our data processing workflow leads to an animated, spatial model of the speaker with a segmented mouth area. 4D-MSMs may become the essential tool for objectifying and archiving diagnoses, conducting asynchronous expert consultations, and documenting the progress in therapy. In the future, we plan to build a computer-aided speech diagnosis system – an expert system linking the audio and video features with the occurrence of selected speech disorders.

ACKNOWLEDGMENT

The authors would like to thank Joanna Trzaskalik, Ph.D., Ewa Kwaśniok, M.A., and Natalia Moćko, Ph.D. for their valuable help in consulting speech therapy issues.

REFERENCES

[1] P. T. Fogle, *Essentials of Communication Sciences & Disorders*. Burlington, MA, USA: Jones & Bartlett, 2022.

[2] K. G. Shipley and J. G. McAfee, *Assessment in Speech-Language Pathology: A Resource Manual*. San Diego, CA, USA: Plural, 2019.

[3] Ministry of Science and Higher Education, (PL) Ministerstwo Edukacji i Nauki. (2022). *Register of Schools and Educational Institutions, (PL) Rejestr Szkól i Placówek Oswiaty*. Accessed: May 19, 2022. [Online]. Available: https://rspo.gov.pl/

[4] C. Scheideman-Miller, P. Clark, S. Smeltzer, J. Carpenter, B. Hodge, and D. Prouty, "Two year results of a pilot study delivering speech therapy to students in a rural Oklahoma school via telemedicine," in *Proc. 35th Annual Hawaii Int. Conf. Syst. Sci.*, 2002, p. 9, doi: 10.1109/HICSS.2002.994136.

[5] G. C. Fairweather, M. A. Lincoln, and R. Ramsden, "Speech-language pathology teletherapy in rural and remote educational settings: Decreasing service inequities," *Int. J. Speech-Lang. Pathol.*, vol. 18, no. 6, pp. 592–602, Nov. 2016, doi: 10.3109/17549507.2016.1143973.

[6] D. R. Campbell and H. Goldstein, "Evolution of telehealth technology, evaluations, and therapy: Effects of the COVID-19 pandemic on pediatric speech-language pathology services," *Amer. J. Speech-Lang. Pathol.*, vol. 31, no. 1, pp. 271–286, Jan. 2022, doi: 10.1044/2021_AJSLP-21-00069.

[7] N. Raman, R. Nagarajan, L. Venkatesh, D. S. Monica, V. Ramkumar, and M. Krumm, "School-based language screening among primary school children using telepractice: A feasibility study from India," *Int. J. Speech-Lang. Pathol.*, vol. 21, no. 4, pp. 425–434, Jul. 2019, doi: 10.1080/17549507.2018.1493142.

[8] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Apraxia world: A speech therapy game for children with speech sound disorders," in *Proc. 17th ACM Conf. Interact. Design Children*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 119–131, doi: 10.1145/3202185.3202733.

[9] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna, and K. J. Ballard, "Speech-driven mobile games for speech therapy: User experiences and feasibility," *Int. J. Speech-Lang. Pathol.*, vol. 20, no. 6, pp. 644–658, Oct. 2018, doi: 10.1080/17549507.2018.1513562.

[10] A. Rueda and S. Krishnan, "Feature analysis of dysphonia speech for monitoring Parkinson's disease," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2308–2311, doi: 10.1109/EMBC.2017.8037317.

[11] J. Rusz, J. Rusz, J. Hlavnička, T. Tykalová, M. Novotný, P. Dušek, K. Šonka, and E. Råžčka, "Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1495–1507, Aug. 2018, doi: 10.1109/TNSRE.2018.2851787.

[12] S. Alharbi, M. Hasan, A. J. H. Simons, S. Brumfitt, and P. Green, "A lightly supervised approach to detect stuttering in children's speech," in *Proc. Interspeech*, 2018, pp. 3433–3437, doi: 10.21437/Interspeech.2018-2155.

[13] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Comput. Speech Lang.*, vol. 50, pp. 62–84, Jul. 2018, doi: 10.1016/j.csl.2017.12.006.

[14] M. Krecichwost, Z. Miodonska, P. Badura, J. Trzaskalik, and N. Mocko, "Multi-channel acoustic analysis of phoneme /s/ mispronunciation for lateral sigmatism detection," *Biocybernetics Biomed. Eng.*, vol. 39, no. 1, pp. 246–255, Jan. 2019, doi: 10.1016/j.bbe.2018.11.005.

[15] Z. Miodonska, P. Badura, and N. Mocko, "Noise-based acoustic features of Polish retroflex fricatives in children with normal pronunciation and speech disorder," *J. Phonetics*, vol. 92, May 2022, Art. no. 101149, doi: 10.1016/j.wocn.2022.101149.

[16] K. Weidner and J. Lowman, "Telepractice for adult speech-language pathology services: A systematic review," *Perspect. ASHA Special Interest Groups*, vol. 5, no. 1, pp. 326–338, Feb. 2020, doi: 10.1044/2019_PERSP-19-00146.

[17] K. Coufal, D. Parham, M. Jakubowitz, C. Howell, and J. Reyes, "Comparing traditional service delivery and telepractice for speech sound production using a functional outcome measure," *Amer. J. Speech-Lang. Pathol.*, vol. 27, no. 1, pp. 82–90, Feb. 2018, doi: 10.1044/2017_AJSLP-16-0070.

[18] A. J. Hill and H. M. Breslin, "Refining an asynchronous telerehabilitation platform for speech-language pathology: Engaging end-users in the process," *Frontiers Hum. Neurosci.*, vol. 10, Dec. 2016, doi: 10.3389/fnhum.2016.00640.

[19] Salus Publica Foundation. *Afast! Say it. (PL) Afast! Powiedz to*. Accessed: Jul. 24, 2022. [Online]. Available: https://afast.pl/

[20] A. Gaodida, H. Koppisetty, K. Potdar, and A. Biwalkar, "Aiding speech therapy using audio and video processing," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2020, pp. 1–5, doi: 10.1109/CSDE50874.2020.9411576.

[21] T. Qiu, H. Zhang, C. Zhou, Q. Tang, L. Wang, and X. Ke, "Application of telemedicine for preliminary screening of autism spectrum disorder," *Frontiers Pediatrics*, vol. 9, Jan. 2022, Art. no. 745597, doi: 10.3389/fped.2021.745597.

[22] Z. Miodonska, M. D. Bugdol, and M. Krecichwost, "Dynamic time warping in phoneme modeling for fast pronunciation error detection," *Comput. Biol. Med.*, vol. 69, pp. 277–285, Feb. 2016, doi: 10.1016/j.compbiomed.2015.12.004.

[23] A. Hair, G. Zhao, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions," in *Proc. Interspeech*, 2021, pp. 2936–2940, doi: 10.21437/Interspeech.2021-69.

[24] K. Young, T. Sweeney, R. R. Vos, F. Mehendale, and H. Daffern, "Evaluation of noise excitation as a method for detection of hypernasality," *Appl. Acoust.*, vol. 190, Mar. 2022, Art. no. 108639, doi: 10.1016/j.apacoust.2022.108639.

[25] M. Krecichwost, N. Mocko, and P. Badura, "Automated detection of sigmatism using deep learning applied to multichannel speech signal," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102612, doi: 10.1016/j.bspc.2021.102612.

[26] Z. A. Benselama, M. Guerti, and M. A. Bencherif, "Arabic speech pathology therapy computer aided system," *J. Comput. Sci.*, vol. 3, no. 9, pp. 685–692, Sep. 2007, doi: 10.3844/jcssp.2007.685.692.

[27] K. Dhaky, M. Bulsara, and B. Sethna, "Speech therapy and assessment (via multimedia devices for cleft lip and palate Patients)," in *Proc. IEEE Global Humanitarian Technol. Conf.*, Oct. 2011, pp. 415–418, doi: 10.1109/GHTC.2011.46.

[28] N. Sebkhi, D. Desai, M. Islam, J. Lu, K. Wilson, and M. Ghovanloo, "Multimodal speech capture system for speech rehabilitation and learning," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2639–2649, Nov. 2017, doi: 10.1109/TBME.2017.2654361.

[29] Z. Bílková, A. Novozámský, A. Domínec, V. S. Greško, B. Zitová, and M. Paroubková, "Automatic evaluation of speech therapy exercises based on image data," in *Image Analysis and Recognition*, F. Karray, A. Campilho, and A. Yu, Eds. Cham, Switzerland: Springer, 2019, pp. 397–404, doi: 10.1007/978-3-030-27202-9_36.

[30] W. Katz, S. Mehta, M. Wood, and J. Wang, "Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation," *J. Acoust. Soc. Amer.*, vol. 141, no. 1, pp. 57–63, 2017, doi: 10.1121/1.4973907.

[31] C. Kroos, "Evaluation of the measurement precision in three-dimensional electromagnetic articulography (carstens AG500)," *J. Phonetics*, vol. 40, no. 3, pp. 453–465, May 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S009544701200023X

[32] S. Wood, J. Wishart, W. Hardcastle, J. Cleland, and C. Timmins, "The use of electropalatography (EPG) in the assessment and treatment of motor speech disorders in children with Down's syndrome: Evidence from two case studies," *Develop. Neurorehabilitation*, vol. 12, no. 2, pp. 66–75, Jan. 2009, doi: 10.1080/17518420902738193.

[33] J. Cleland, C. Timmins, S. E. Wood, W. J. Hardcastle, and J. G. Wishart, "Electropalatographic therapy for children and young people with Down's syndrome," *Clin. Linguistics Phonetics*, vol. 23, no. 12, pp. 926–939, Dec. 2009, doi: 10.3109/02699200903061776.

[34] A. Bhatti, *Stereo Vision*. Rijeka: IntechOpen, 2008. [Online]. Available: https://doi.org/10.5772/89

[35] G. D. Fulvio, E. Frontoni, A. Mancini, and P. Zingaretti, "A stereovision system for dimensional measurements in industrial robotics applications," in *Proc. IEEE/ASME 10th Int. Conf. Mech. Embedded Syst. Appl. (MESA)*, Sep. 2014, pp. 1–6, doi: 10.1109/MESA.2014.6935618.

[36] Y. Xu, V. John, S. Mita, H. Tehrani, K. Ishimaru, and S. Nishino, "3D point cloud map based vehicle localization using stereo camera," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 487–492, doi: 10.1109/IVS.2017.7995765.

[37] I. Steiner, K. Richmond, and S. Ouni, "Speech animation using electromagnetic articulography as motion capture data," 2013, *arXiv:1310.8585*.

[38] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1075–1086, Mar. 2007, doi: 10.1109/TASL.2006.885910.

[39] L. Xie, L. Wang, and S. Yang, *Visual Speech Animation*. Cham, Switzerland: Springer, 2016, pp. 1–30, doi: 10.1007/978-3-319-30808-1_1-1.

[40] M. Krecichwost, Z. Miodonska, J. Trzaskalik, and P. Badura, "Multi-channel speech acquisition and analysis for computer-aided sigmatism diagnosis in children," *IEEE Access*, vol. 8, pp. 98647–98658, 2020, doi: 10.1109/ACCESS.2020.2996413.

[41] Panasonic. *Omnidirectional Back Electret Condenser Microphone Cartridge, Series: WM-61A, WM-61B. Panasonic*. Accessed: Jul. 24, 2022. [Online]. Available: http://konektor.nazwa.pl/serwisowe/panasonic-wm-61a.pdf

[42] ArduCam. *Arducam 8MP 1080P Auto Focus USB Camera Module With Microphone*. Accessed: Jul. 24, 2022. [Online]. Available: https://www.arducam.com/product/b0197arducam-8mp-1080p-auto-focus-usb-camera-module-with-microphone-1-3-2-CMOS-imx179-mini-uvc-usb2-0-webcam-board-with-3-3ft-1m-cable-for-windows-linux-android-and-mac-os/

[43] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000, doi: 10.1109/34.888718.

[44] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 1106–1112, doi: 10.1109/CVPR.1997.609468.

[45] A. Bier and L. Luchowski, "Error analysis of stereo calibration and reconstruction," in *Computer Vision/Computer Graphics CollaborationTechniques*, A. Gagalowicz and W. Philips, Eds. Berlin, Germany: Springer, 2009, pp. 230–241, doi: 10.1007/978-3-642-01811-4_21.

[46] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools*, vol. 120, pp. 122–125, 2000.

[47] M. Omologo, M. Matassoni, and P. Svaizer, *Speech Recognition With Microphone Arrays*. Heidelberg, Germany: Springer, 2001, pp. 331–353, doi: 10.1007/978-3-662-04619-7_15.

[48] T. Giannakopoulos. (2009). *A Method for Silence Removal and Segmentation of Speech Signals, Implemented in MATLAB*. Accessed: May 19, 2022. [Online]. Available: https://uk.mathworks.com/help/audio/ref/detectspeech.html

[49] A. D. Gritzman, D. Rubin, and A. Pantanowitz, "Comparison of colour transforms used in lip segmentation algorithms," *Signal, Image Video Process.*, vol. 9, pp. 947–957, Jan. 2015, doi: 10.1007/s11760-014-0615-x.

[50] P. Badura, W. Wieclawek, and B. Pycinski, "Automatic 3D segmentation of renal cysts in CT," in *Information Technologies in Medicine*, E. Piętka, P. Badura, J. Kawa, and W. Wieclawek, Eds. Cham, Switzerland: Springer, 2016, pp. 149–163, doi: 10.1007/978-3-319-39796-2_13.

[51] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Jun. 2010, doi: 10.1109/MediVis.2008.12.

[52] G.-Y. Lin, X. Chen, and W.-G. Zhang, "A robust epipolar rectification method of stereo pairs," in *Proc. Int. Conf. Measuring Technol. Mechatronics Autom.*, Mar. 2010, pp. 322–326, doi: 10.1109/ICMTMA.2010.220.

[53] K. Jawed, J. Morris, T. Khan, and G. Gimel'farb, "Real time rectification for stereo correspondence," in *Proc. Int. Conf. Comput. Sci. Eng.*, vol. 2, 2009, pp. 277–284, doi: 10.1109/CSE.2009.473.

[54] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008, doi: 10.1109/TPAMI.2007.1166.

[55] The OpenCV Library. (2022). *Disparity map post-filtering with OpenCV*. Accessed: May 19, 2022. [Online]. Available: https://docs.opencv.org/4.x/d3/d14/tutorial_ximgproc_disparity_filtering.html

[56] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, Dec. 2014, doi: 10.1109/TIP.2014.2366600.

[57] H. Wang, J. Cao, X. Liu, J. Wang, T. Fan, and J. Hu, "Least-squares images for edge-preserving smoothing," *Comput. Vis. Media*, vol. 1, no. 1, pp. 27–35, Mar. 2015, doi: 10.1007/s41095-015-0004-6.

[58] W. Liu, X. Chen, C. Shen, Z. Liu, and J. Yang, "Semi-global weighted least squares in image filtering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5862–5870, doi: 10.1109/ICCV.2017.624.

[59] T. S. Sheikh and I. M. Afanasyev, "Stereo vision-based optimal path planning with stochastic maps for mobile robot navigation," in *Intelligent Autonomous Systems*, vol. 15, M. Strand, R. Dillmann, E. Menegatti, and S. Ghidoni, Eds. Cham, Switzerland: Springer, 2019, pp. 40–55, doi: 10.1007/978-3-030-01370-7_4.

[60] Khronos Group. *Graphics Language (GL) Transmission Format (glTF) Family*. Accessed: Jul. 24, 2022. [Online]. Available: https://www.khronos.org/gltf/

[61] The OpenCV Library. (2022). *Stereo SGBM in OpenCV*. Accessed: May 19, 2022. [Online]. Available: https://docs.opencv.org/3.4/d2/d85/classcv_1_1StereoSGBM.html

[62] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Computer Vision—ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 510–523, doi: 10.1007/978-3-642-15558-1_37.

[63] J. He, E. Zhou, L. Sun, F. Lei, C. Liu, and W. Sun, "Semi-synthesis: A fast way to produce effective datasets for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2878–2887, doi: 10.1109/CVPRW53098.2021.00323.

[64] S. Merrouche, M. Andrić, B. Bondžulic, and D. Bujaković, "Objective image quality measures for disparity maps evaluation," *Electronics*, vol. 9, no. 10, p. 1625, Oct. 2020, doi: 10.3390/electronics9101625.

[65] The American Speech-Language-Hearing Association. (2022). *Considerations for Speech, Language, and Cognitive Assessment Via Telepractice*. Accessed: May 19, 2022. [Online]. Available: https://www.asha.org/slp/clinical/considerations-for-speech-language-and-cognitive-assessment-via-telepractice/

[66] N. Smolyanskiy, C. Huitema, L. Liang, and S. E. Anderson, "Real-time 3D face tracking based on active appearance model constrained by depth data," *Image Vis. Comput.*, vol. 32, no. 11, pp. 860–869, Nov. 2014, doi: 10.1016/j.imavis.2014.08.005.

[67] S. Tulyakov, R.-L. Vieriu, E. Sangineto, and N. Sebe, "FaceCept3D: Real time 3D face tracking and analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 28–33, doi: 10.1109/ICCVW.2015.13.

MICHAL KRECICHWOST was born in Bielsko-Biala, Poland, in 1990. He received the B.S., M.S., and Ph.D. degrees in biomedical engineering from the Silesian University of Technology, Gliwice, Poland, in 2013, 2014, and 2020, respectively.

From 2015 to 2017 he was a Programmer with the Salus Publica—the Foundation for Public Health, Krakow, Poland. Since 2019, he has been with the Faculty of Biomedical Engineering, Silesian University of Technology, as a Research Assistant, and since 2020, as an Assistant Professor. He was one of the originators and developers of a computer system supporting the aphasia therapy "Afast! Say it." He took part in four research projects. He is the author of more than 30 articles. His research interests include computer-aided speech diagnosis and therapy, signal processing, deep learning, and electronics.

AGATA SAGE received the B.S. and M.S. degrees in biomedical engineering from the Silesian University of Technology, Gliwice, Poland, in 2018 and 2019, respectively, where she is currently pursuing the Ph.D. degree in biomedical engineering with the Joint Doctoral School.

She took part in two research projects. She publishes her research in recognized journals and international scientific conferences. Her research interests include image processing, artificial intelligence, deep learning, and their applications to computer-aided diagnosis systems.

ZUZANNA MIODONSKA received the B.S., M.S., and Ph.D. degrees in biomedical engineering from the Silesian University of Technology, Gliwice, Poland, in 2013, 2014, and 2019, respectively.

From 2014 to 2019, she was a Research Assistant with the Faculty of Biomedical Engineering, Silesian University of Technology, and from 2015 to 2017, a Research Assistant with the Salus Publica—the Foundation for Public Health, Krakow, Poland. Since 2019, she has been a Postdoctoral Researcher with the Faculty of Biomedical Engineering, Silesian University of Technology. She was one of the originators and developers of a computer system supporting aphasia therapy "Afast! Say it." She participated in three research projects. She is the author of more than 20 articles. Her research interests include computer-aided speech diagnosis and therapy, signal processing, acoustic, and articulation analysis.

PAWEL BADURA was born in Katowice, Poland, in 1980. He received the B.S. and M.S. degrees in electronics and telecommunication, and the Ph.D. and D.Sc. degrees in biomedical engineering from the Silesian University of Technology, Gliwice, Poland, in 2004, 2007, and 2017, respectively.

From 2007 to 2011, he was an Assistant Professor with the Faculty of Automatic Control, Electronics and Computer Science. Since 2011, he has been working with the Faculty of Biomedical Engineering, Silesian University of Technology, as an Assistant Professor, until 2018, and as an Associate Professor, since then. He took part in nine research projects. Currently, he leads a research project of the Polish National Science Centre: Hybrid system for acquisition and processing of multimodal signal in the analysis of sigmatism in children. He is the author of more than 70 articles. His research interests include image analysis, signal processing, computer-aided diagnosis systems, artificial intelligence, and machine learning methods. He is a Guest Editor of the journal *Applied Sciences* and a co-editor of six books.