IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 6, 2020, accepted May 15, 2020, date of publication May 21, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996413

Multichannel Speech Acquisition and Analysis for Computer-Aided Sigmatism Diagnosis in Children

MICHAL KRECICHWOST[1], ZUZANNA MIODONSKA[1], JOANNA TRZASKALIK[2], AND PAWEL BADURA[1]

[1]Faculty of Biomedical Engineering, Silesian University of Technology, 41-800 Zabrze, Poland
[2]Faculty of Education, Institute of Educational Sciences, Jesuit University Ignatianum in Krakow, 31-501 Kraków, Poland

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

This work was supported by the Polish National Science Centre, Poland (Hybrid System for Acquisition and Processing of Multimodal Signal in the Analysis of Sigmatism in Children) under Grant 2018/30/E/ST7/00525.

ABSTRACT A novel concept for acoustic data acquisition for computer-aided diagnosis of a common speech disorder (sigmatism) is presented in this paper. We designed and built a data acquisition device enabling repeatable speech signal acquisition in up to fifteen spatially-organized acoustic channels. The system is safe, non-invasive, comfortable, visually attractive to the user, and does not affect the articulation process. It is easy and convenient to use and transport and does not require a specialized measuring room. We collected a large speech corpus containing speech samples from 107 children aged five or six. The data were acquired according to a dedicated protocol. They consisted of multichannel acoustic recordings of selected words containing sibilants and diagnostic descriptions of articulatory features prepared by speech therapy experts. The data acquisition device was examined for responses and repeatability of individual microphones in the presence of various synthetic and human-generated acoustic stimuli. Then, it was verified for its ability to indicate distinctive patterns in spatial energy distribution in different realizations of two sibilants: /s/, /ʃ/ in three pronunciation categories each, based on collected speech-articulation corpus. The results confirm that a multichannel speech signal can be successfully employed for the analysis of the spatial distribution of airflow during normative or pathological realization of sibilant sounds in children. The method is promising for comprehensive analysis of articulatory features, which follows new trends in the description of speech disorders; such an approach was not employed in speech diagnosis or therapy so far.

INDEX TERMS Acoustic data acquisition, acoustic signal processing, computer-aided speech diagnosis, multichannel acoustic signal, sigmatism, speech analysis, speech corpus.

I. INTRODUCTION

One of the most common types of speech disorders is sigmatism (lisping), which consists in incorrect articulation of dentalized phones, called also sibilants or sibilant sounds (in Polish: /s, z, ts, dz; ʃ, ʒ, tʃ, dʒ; ɕ, ʑ, tɕ, dʑ/). Dentalized sounds appear as the very last in the child's speech development. They are considered difficult to articulate, and according to various literature reports, sigmatism may constitute 30–60% of all speech disorders among children in Poland [1], [2]. Depending on the classification criterion, multiple types of sigmatism can be distinguished. In probably the most common concept in contemporary Polish speech therapy, pathological pronunciation is analyzed by observing non-normative features in the realization of individual phonemes [3].

Articulation description may contain many different detailed features, such as manner of articulation, active and passive place of articulation, airflow direction, or sonority. Some features may be normative for selected sibilants and pathological for the others (e.g., sonority); some are not normative for any sound in specific language systems (e.g., interdentality is considered a pathology in all Polish phonemes). Additionally, many articulatory features are hard to observe, which often hampers the determination of a precise diagnosis. Moreover, there is still a lack of measurement methods to systematize and objectify the process of assessing pronunciation. The development of computer-based speech diagnosis tools for speech therapy patients can be a step that would allow improving the diagnostic process by providing additional data

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

to the therapist. Such support could be particularly helpful for less experienced diagnosticians. Such a form of articulation diagnosis support could improve the performance of speech screening tests conducted in schools and kindergartens. This would allow faster intervention, which usually leads to increased effectiveness of therapy. Finally, automated analysis of articulation could become an element of multimedia programs for speech exercises at home between meetings with a speech therapist.

A. STATE OF THE ART

1) COMPUTER-AIDED DIAGNOSIS OF SIGMATISM

Studies concerning computerized methods of pronunciation evaluation in recorded speech are conducted in different countries [4]. The methods are often based on popular speech analysis techniques. However, they focus mostly on binary evaluation of phones (norm/pathology), not on the analysis of specific types of pathology. Moreover, most of the proposed tools are designed for second-language learners [5]–[7]. Solutions dedicated to speech therapy patients can hardly be found. According to our knowledge, three propositions of speech analysis methods for sigmatism diagnosis are reported in literature [8]–[10]. The first work describes a computer application designed to support the therapy of lisping in Arabic [8]. The proposed method of phoneme binary evaluation is implemented with an artificial neural network based on Mel-frequency cepstral coefficients (MFCC). Another approach was proposed in [9] and employed the Gaussian Mixture Model, also based on MFCC coefficients. Parameters of this mixture (weights, mean value vectors, and diagonal elements of the covariance matrix of individual components of the mixture) were subsequently used as feature vectors and classified by using the support vector machine (SVM).

The solutions mentioned above were tested on speech databases of adult or teenage subjects, without data acquired from children. Such an approach is frequent in preliminary works because of a difficulty in gathering a representative corpus of children's pathological speech. Unfortunately, acoustic methods of processing normative speech dedicated to adults often prove to work less efficiently with children due to different spectral characteristics [11], [12]. Similar problems should likely occur within pathological speech experiments. According to our knowledge, there is only one work on sigmatism detection conducted on speech data collected from children. Anjos *et al.* [10] recorded articulation of four isolated sibilants in a group of 145 children with correct and incorrect pronunciation. Researchers proposed a method that employs SVMs trained to recognize correctly pronounced sibilants. No further analysis of different possible pathologies was conducted.

The main goal of these works was to evaluate the pronunciation of 2–4 selected sibilants based on speech samples annotated as correct or incorrect. However, they did not present any specific indicators of improper articulation nor analysis of different types of possible pathologies. The only measuring technique employed in these studies was single-channel speech recording, without any more data that could provide additional information on the patient's articulation. The realization of individual sibilant sounds is distinguished by a number of features, among which airflow direction during articulation is of great importance. Recording speech signal in a single channel does not enable spatial analysis. This limitation raises a question of different, more informative speech acquisition systems to be employed in a sigmatism diagnosis.

2) DATA ACQUISITION SYSTEMS

Several data acquisition techniques were employed in recent years for data collection in speech therapy support, e.g., electromagnetic articulography (EMA), electropalatography (EPG), multichannel audio recording systems, electromyography (EMG), aerodynamic methods, fiberoscopy, acoustic rhinometry, medical imaging techniques: ultrasound (US), computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), palatal videofluoroscopy (PVF).

Electromagnetic articulography [13], [14] is a method for tracking articulator motion (lips, tongue, mandible, and soft palate) by using a magnetic field. For this purpose, a set of sensors is placed on the speaker's articulatory organs by using a medical glue. The sensors are connected with wires to the central unit. The real-time position of the speech apparatus during speaking can be visualized in a three-dimensional space. EMA is mostly used to assess the position of the tongue and relationships between the tongue, lips, and mandible motion in terms of time, amplitude, and speed of movements. EMA is also used to model the articulator motion by using motion capture methods for the speech formation process [15]. Katz and Mehta [16] used EMA for speech training with the use of 3D tongue models and their real-time visualization.

Electropalatography is used to monitor tongue contact with the hard/soft palate during articulation. The examination involves placing an artificial palate containing electrodes inside the oral cavity to register contact with the tongue during the articulation of sounds in isolation, in syllables, and words. This technique was applied in teaching correct articulation patterns in speech developmental disorders and diagnosis and therapy of articulation pathology, e.g., in children with cleft palate, dysarthria, or Down syndrome [17], [18]. The high cost of the EPG examination results from the need to prepare the personalized artificial palate for each patient so that it adheres precisely to the palate. Commercial EPG-based devices are also available, e.g., the SmartPalate system [19]. The mouthpiece, with over a hundred pressure sensors, is placed in the oral cavity and follows the tongue motion during the articulation of individual sounds by using dedicated software.

The multichannel audio recording method was reported by Król and Lorence [20]. The study attempts to assess

the acoustic field distribution in the process of lateral and nasal articulation of Polish phones. The proposed device [21] consists of a 16-channel recorder, a circular microphone array, microphone amplifiers and analog-to-digital converters, and a signal processor. Mik *et al.* [22], [23] extended the system to a multimodal framework consisting of an electromagnetic articulograph, a 16-channel microphone array, and three ultrafast cameras. The system enables synchronous video acquisition of the mouth area, distribution of acoustic field intensity, and tracking trajectories of characteristic face points. The choice of the microphone array with peripherals was dictated by the physical size and structure of the parent device. Besides, the device had to be in front of the speaker during registration. This required the speaker to maintain constant focus and fixed head position during the examination. These limitations make the system hard to apply to examinations involving preschool children.

An example of a multimodal system for supporting the rehabilitation of people with motor speech disorders was described by Sebkhi *et al.* [24]. The Multimodal Speech Capture System (MSCS) enables the recording of an acoustic signal, image of articulators, and tongue movement. It consists of two microphones, a camera, and twenty-four triaxial magnetometers. The purpose of the proposed support system was to visualize the recorded data in real-time, providing feedback to the speaker. Similarly to [20]–[23], the Sebkhi's prototype is invasive and requires a fixed position of the subject's head related to the acquisition device, preventing it from application to speech diagnosis of preschool children.

The EMA-based systems affect the speech production by placing sensors on articulatory organs. Aron *et al.* [25] presented an alternative way of monitoring the speech production process. The internal articulatory organs were recorded through ultrasound imaging, and a stereo vision system observed the external articulators. An electromagnetic tracking system was used to record the movements of the US probe. The authors of [25] believe that such data acquisition methodology could be able to perform registration and fusion of the US image and other imaging techniques, e.g., MRI.

Unfortunately, due to practical reasons, probably none of the data acquisition systems mentioned above can be employed in pronunciation analysis and evaluation in children. Placing the sensors inside the mouth, e.g., on the tongue, disrupts natural articulation. It is necessary to use various preparations and tissue adhesives, which is objectionable for many parents. The sensors are wired, and the wiring is led out of the mouth, which affects the patient's comfort during the examination. Some of the measurement methods, e.g., electromagnetic articulograph, require a specialized environment during measurements which excludes application in most available venues (kindergartens or speech therapy offices). In some cases, obtained data require specialized tools or additional expert courses to be interpreted. Commercial and advanced systems often involve closed application programming interface (API) being practically inapplicable in attempts to adapt the tool to individual needs.

B. AIMS AND SCOPE

This work aimed to propose and develop a speech data acquisition method that could be used in computer-aided sigmatism diagnosis in children and would be more informative than traditionally used single-channel speech recordings. Data acquisition systems described in literature feature specific drawbacks that limit their usefulness for the considered problem. Based on the analysis of these systems and our previous preliminary studies [26], [27], we have formulated a set of assumptions that should be met by an applicable measurement method.

The assumed system should enable multichannel, spatial, and repeatable speech signal acquisition, be safe, comfortable, and visually attractive to preschool children, and should not affect the articulation process. Moreover, it should be low-cost, easy, and convenient to use and transport, and should not require a specialized measuring room. Data recording should be non-invasive, not require additional markers on the internal and external articulatory organs. Finally, the system should allow access to registered raw data.

Based on these assumptions, we designed, built, and verified a novel data acquisition device for multichannel acoustic signal recording. The data acquisition system is supposed to stand for a foundation of a measurement method that allows a more detailed analysis of articulation while maintaining low invasiveness in the context of speech therapy diagnosis. Thus, one of the goals of this study was to validate the data acquisition system from a technical and acoustic point of view. A dedicated measuring stand inside the acoustically prepared room was prepared along with a set of verification procedures compliant with the Polish Committee for Standardization requirements. Both synthetic and real (human-generated) acoustic stimuli were used for this purpose.

Another set of experiments was targeted at speech therapy analysis of selected articulation issues in sigmatism reflected in various spatial distributions of acoustic signals. We defined a 5-channel microphone array by employing signal aggregation techniques in terms of delay-and-sum beamforming. This framework was able to increase the signal-to-noise ratio compared to single-sensor recordings by a noticeable margin. The spatial energy distribution was investigated and analyzed in different realizations of selected sibilants produced by preschool children supervised by the speech therapy experts. For this purpose, a speech-articulation corpus was collected. The entire procedure and linguistic material was carefully designed and described in this paper.

C. PAPER STRUCTURE

After the introduction to the research domain and presentation of the aims and scope of the paper in Section I, the data acquisition system is described in detail in Section II. The process of speech corpus registration: linguistic material, speech examination, and data collection protocol are presented in Section III. Section IV specifies the experiments prepared and performed for both data acquisition

device validation and spatial acoustic and articulation analysis. The obtained results are discussed in Section V. Finally, Section VI concludes the paper.

II. DATA ACQUISITION DEVICE

The designed data acquisition device enables multichannel, spatial, and repeatable speech signal recording. It meets the ergonomic requirements of the target age group (preschool children). Microphone mounting enables easy regulation of sensor positions, i.a., microphone adjustment to the sound source (the subject's mouth). The designing process was carried out in consultation with speech therapy experts. The device enables recording of the speech signal in fifteen spatially arranged channels with a sampling frequency of 44.1 kHz each. It is a portable device consisting of two parts: a mask put on the head of the examined person and a processing unit. A prototype of the device is shown in Fig. 1.

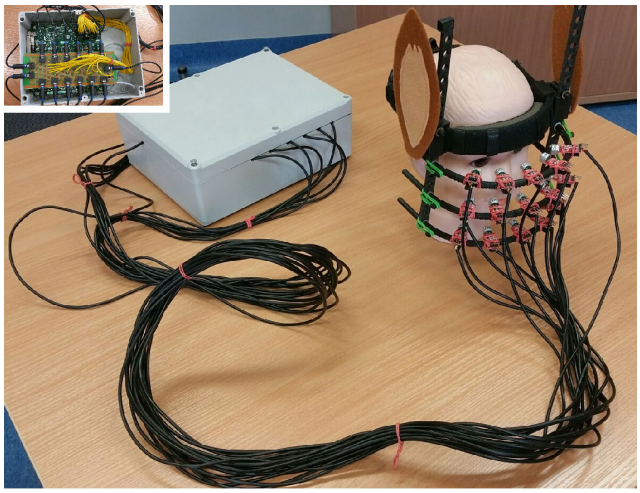

FIGURE 1. The multichannel acquisition device.

A. ACOUSTIC MASK

The mask consists of a mounting strap, three rigid arches with five microphones mounted on each one, and two connectors for fixing the arches at the desired height to the sound source. The connectors enable modifications of the distance between the arches, thereby changing the distance between individual microphones. The proposed assembly method allows adjusting the mask size to the individual needs of each subject and preserving the organization of the microphones between sessions.

A mounting strap with a knob enables a secure fitting on the speaker's head. The inner part of the strap is equipped with removable sponges to adjust the mask to a smaller head size and to secure wearing comfort. Supplementary straps are attached to the central part of the strap to additionally stabilize the position of the mask. This minimizes the risk of mask slipping and shifting even during sudden movements of the child's head. We considered the use of adjustable fastening on the chin, yet such an approach would affect the articulation too much. To secure both user-friendliness for the target group (five to six years old kids) and positive aesthetic experience, we equipped the mounting strap and connectors with Velcro straps, on which additional decorative accessories can be attached, e.g., rabbit ears or headdress (Fig. 2).

FIGURE 2. Examples of decorative accessories to be mounted on the acquisition device.

The measurement and processing unit enables powering the mask and transmitting acquired data to a computer via a USB connection. A safe voltage of 5 V powers all electronic components. It does not require an external power supply, which eliminates the problem of a so-called ground loop between the measuring device and the USB port, introducing additional noise during recording. Fig. 3 presents a diagram of the measurement system components.

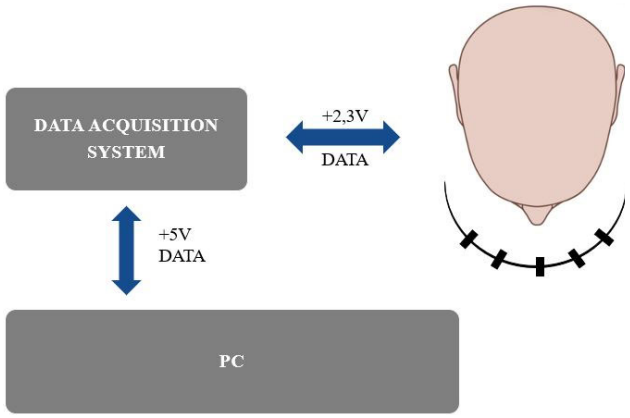

FIGURE 3. Diagram of the measurement system.

B. SIGNAL ACQUISITION AND PROCESSING UNIT

The speech signal was recorded with electret microphones Panasonic WM-61a [28]. They were chosen for several reasons. From the acoustical point of view, this microphone features high sensitivity and relatively flat amplitude characteristics in the whole acoustic band, and therefore it is willingly used for acoustic measurements [29]–[31].

The WM-61a sensor is small, which is profitable due to its placement ca. 8.5 cm from the subject's mouth. The small sensor size, along with the use of acoustic insulation, made it possible to reduce the reflected wave impact on the measurements. The microphone features also satisfying physicomechanical properties: low sensitivity to ambient temperature and mechanical shocks, high robustness, and stability of characteristics. It involves a permanently polarized electret membrane generating an electric field in the air gap. Thus, it is safer than condenser microphones with external polarization requiring phantom power of 48 V. There is no danger of electric breakdown of the air gap between the conductive plates. The WM-61a sensor has the following parameters: -35 ± 4 dB sensitivity, 20–20,000 Hz frequency band, signal-to-noise ratio (SNR) over 62 dB, and a 2.3 V supply voltage. It has an omnidirectional directivity, which makes it robust to the breath or stop consonants and also eliminates the proximity effect (increased presence of low-frequency tones) [32].

The microphone signal was amplified by using the MAX9812 [33] unit in its standard application scheme with additional peripherals. The amplifier has a small size and features a low noise level with a fixed 20 dB gain over a frequency range of up to 400 kHz. It provides 100x voltage gain and total harmonic distortion (THD) of 0.015% (-76 dB).

The system employed the high-class data acquisition board DaqBoard/3000USB Series [34]. It provides a sufficient number (sixteen) of asymmetrical single-ended analog inputs to be sampled at 44.1 kHz frequency each. The board is equipped with four 16-bit A/D converters operating at 1 MHz and a FIFO cache memory enabling data synchronization and transmission for online data storage and real-time visualization. The board is supplied with a 5 V voltage, features THD at 0.01% (-80 dB), and SNR at 72 dB. The microphone, along with the amplification system, is equipped with a shielded cable with a mini-jack plug. The wiring includes three wires: a signal wire, a 3.3 V power wire, and a ground wire, connecting to the data acquisition board via a designed adapter.

A set of procedures was designed for the verification of the acquisition system in two aspects: response repeatability of individual microphones to standardized sound stimuli and usefulness for recording non-normative airflow during articulation of specific Polish phones. The testing was divided into two protocols: synthetic testing and usability testing involving human-generated stimulation. Testing protocols and verification results are described in Section IV-A.

III. SPEECH-ARTICULATION CORPUS REGISTRATION

The described project involves the analysis of normative and pathological pronunciation patterns in children. According to our knowledge, no adequate speech databases exist for the Polish language. Therefore, a multichannel speech corpus with speech pathologists' annotations was developed as a part of our project. The registration consisted of two parts. In the first part, speech samples were recorded by using a measuring device described in Section II. The second part was a speech examination carried out by speech pathologists.

Registration of speech database was conducted in three kindergartens by an interdisciplinary team consisting of speech therapists and speech engineers. The experimental group consisted of children aged five or six. The selection of the age group resulted from the knowledge about the development of a child's articulative skills, who by the age of six should correctly articulate all phonemes of the Polish language [35]–[37].

In addition to age, the criteria for including a child in the study were:

- a written consent from parents or legal guardians and the child's oral consent to participate in the study; the child could have withdrawn this consent at any time during the recordings and speech examination,
- no respiratory tract infection. Difficulties with breathing through the nose may cause abnormal patterns in articulation, even in children with normative pronunciation. For this reason, the speech examination performed during an infection would not be fully reliable,
- having a full set of primary teeth. Missing teeth can cause an uncontrolled airflow from the mouth, resulting in distortion of speech sounds. In that case, pronunciation cannot be considered normative, but it also cannot be explicitly classified as a specific pathology.

No additional exclusion criteria were formulated.

A. SPEECH EXAMINATION PROTOCOL AND LINGUISTIC MATERIAL

A speech pathology description was prepared for each speaker. The purpose of the speech examination was to determine the type of speech disorder for a given speaker and to provide detailed information about:

- the current state of the child's pronunciation, specifically concerning dentalized sounds,
- the child's physiological features and abilities, concerning, e.g., swallowing, breathing or tongue mobility,
- anatomical features of the speech apparatus (e.g., correct length of the tongue frenulum and the upper frenulum, shape of the palate, dentition).

During speech examination, the pronunciation of dentalized sounds was defined as normative or annotated as presenting non-normative articulatory features, e.g., laterality or interdentality. The considered types of sibilants' pronunciation are presented in Table 1.

The linguistic material recorded in the speech database consisted of sixteen individual words covering two basic sibilants: /s/ and /ʃ/ (Table 2). The dictionary content was constructed taking into account several criteria. Recordings included isolated words with dentalized sounds in various articulation phases: at the beginning, in the middle, or at the end of the word. The words were illustrated in individual pictures. The child's task was to name the picture. Therefore, selected words had to be actively known by the most of preschool children. They also had to be easily represented

TABLE 1. Considered categories of pronunciation for three phone sets.

Pronunciation type		
/s, z, ts, dz/	/ʃ, ʒ, tʃ, dʒ/	/ɕ, ʑ, tɕ, dʑ/
Norm	Norm	Norm
Addentality	Addentality	Addentality
Interdentality (medial/right/left)	Interdentality (medial/right/left)	Interdentality (medial/right/left)
Dorsality	Dorsality	Dorsality
Laterality (medial/right/left)	Laterality (medial/right/left)	Laterality (medial/right/left)
Other	Dentality	Dentality
	Alveolo-palatality	Other
	Other	

TABLE 2. Word set recorded in the speech corpus. The analyzed phonemes are marked in bold.

Original word	English translation	Analyzed phoneme
samolot	*plane*	/s/
serce	*heart*	/s/
strażak	*firefighter*	/s/
pa**s**ek	*belt*	/s/
para**s**ol	*umbrella*	/s/
li**s**	*fox*	/s/
pie**s**	*dog*	/s/
szafa	*wardrobe*	/ʃ/
sznur	*cord*	/ʃ/
szufelka	*shovel*	/ʃ/
kalo**sz**e	*wellingtons*	/ʃ/
ko**sz**yk	*basket*	/ʃ/
ksią**ż**ka	*book*	/ʃ/
leka**rz**	*doctor*	/ʃ/
nó**ż**	*knife*	/ʃ/
wą**ż**	*snake*	/ʃ/

in illustrations. Thus, nominatives of common nouns were selected.

B. DATA ACQUISITION PROTOCOL

The recordings were made by using the acoustic mask described in Section II. Data registration was conducted during 3-part sessions. In the introductory part, the child was familiarized with measuring equipment. The acoustic mask was put on their head. After the initial adjustment, the position of the central microphone was verified relative to the speaker's philtrum.

In the second part of the session, the child's speech samples were recorded using the picture test. Illustrations were presented on the computer screen in front of the child. During the recording, the speaker was observed by speech therapists. At this stage, some preliminary observations on pronunciation patterns could have been made.

After the picture test, the measuring device was removed from the child's head. A team of two speech therapy specialists proceeded to the actual speech examination (the third part of the session). The results of this examination were registered according to the categories presented in Table 1.

Speech therapy examination was conducted during the same session as speech samples recording. It was crucial due to several factors. First, one of the conditions for including a child in the study was their good general health (no signs of upper respiratory tract infection). Postponing the speech therapy examination would carry the risk of developing an infection, which could make it impossible to create a reliable speech therapy description. Secondly, the pronunciation of children evolves. Changes in pronunciation may result, e.g., from patterns taken from the environment (from caregivers, but also peers in the kindergarten group), taking up speech therapy, or falling out milk teeth. A speech diagnosis distant in time from speech recordings could, therefore, not correspond entirely to the recorded material.

Overall, 107 children (51 girls and 56 boys) were recorded, examined by speech pathologists, and included in the speech-articulation corpus.

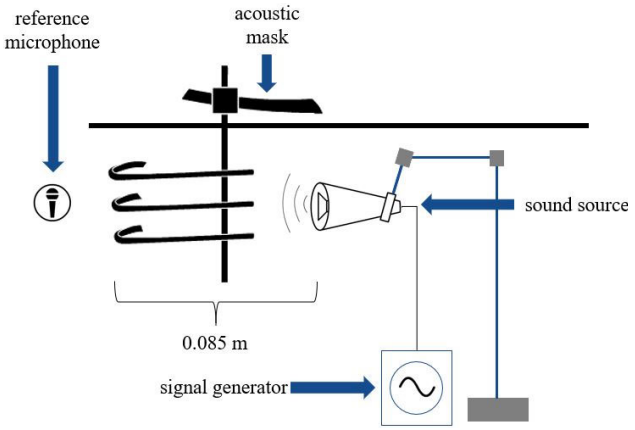

FIGURE 4. Scheme of the measuring stand.

IV. EXPERIMENTS AND RESULTS

To verify the data acquisition device and to justify the use of multichannel spatial speech recording for computer-aided speech diagnosis, two types of experiments were performed:

- technical examination of the data acquisition process by using both synthetic and human-generated signals,
- spatial acoustic analysis of normative and pathological speech.

A. VERIFICATION OF THE DATA ACQUISITION DEVICE

1) SYNTHETIC TESTING

The testing procedure was based on the Polish standard PN-EN ISO 3746: 2011 [38] specifying acoustic measurements of sound level SL in conditions close to free field. A measuring stand was proposed for synthetic testing (Fig. 4) consisting of:

- acoustic mask with microphone arrangement presented in Fig. 5,
- signal generator Rigol DG1022 [39],
- sound source – an 8 Ω speaker,
- reference microphone (sound level meter) Voltcraft SL-200 [40] for measuring the sound level in decibels. The meter belongs to the second class of accuracy in general environmental tests.

The speaker was connected to the signal generator. The sound source was located ca. 8.5 cm from the microphone under

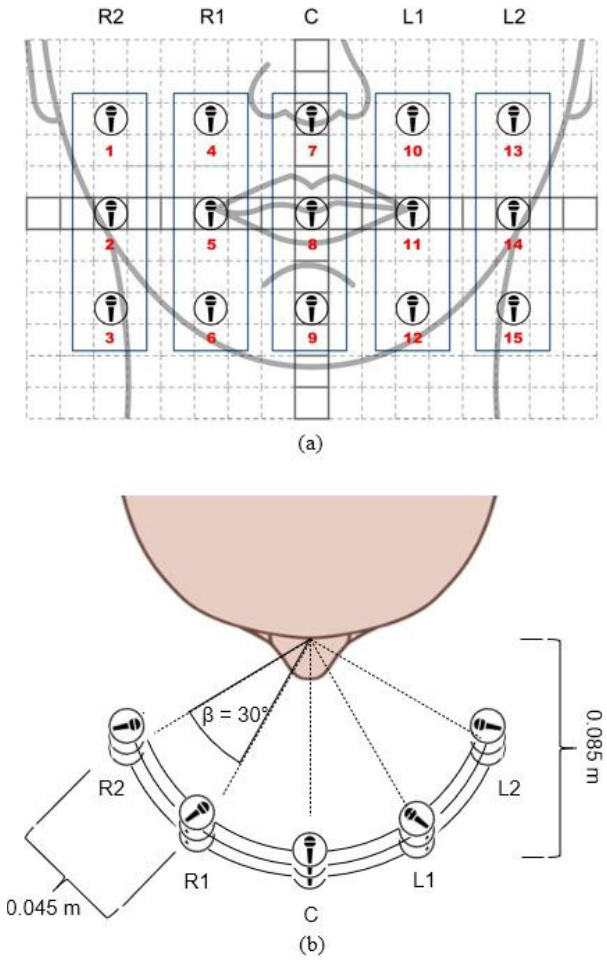

FIGURE 5. Front (a) and top (b) view at the arrangement of microphones during the study with microphone numbers (red font) and ULA IDs (black font).

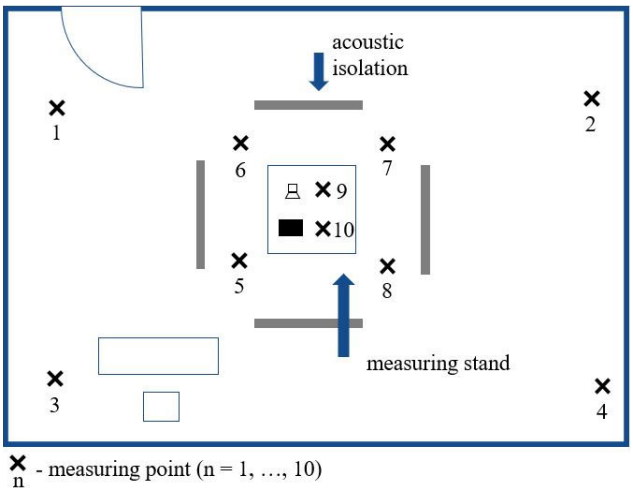

FIGURE 6. Arrangement of the room for synthetic testing.

investigation. The reference microphone was located close to the tested microphone (ca. 3 cm). The reference microphone was calibrated prior to testing by using a device producing a reference sound of a standard 94 dB level and 1 kHz frequency.

The measurements were performed in a special room (Fig. 6). The noise rating NR was in the NR 25–30 range, acceptable for recording studios [41], [42]. The background noise level was measured ten times at different room locations (Fig. 6, black crosses) by using a Voltcraft SL-200 sound level meter. The room was also verified for the presence of other sound sources audible by the human ear. For each test measurement, 10 seconds of silence was recorded for the room noise profile determination. The room was acoustically adapted by using partitions (insulating mats) surrounding the area designated for the measuring stand (Fig. 6) to reduce the level of ambient noise and possible reverberation interference.

The testing protocol covered verifying responses and repeatability of individual microphones in the presence of various acoustic stimuli. For this purpose, a 90-second acoustic test sequence was proposed, which consisted of a 10-second silence segment followed by eight separate tones with frequencies of 1, 2, 3, 4, 5, 6, 7, 8 kHz, each of a 1 V amplitude and lasting 10 s. Such sequence was recorded five times separately for each microphone yielding a total of forty individual tones recorded by each microphone plus a reference microphone. Two metrics were determined for each microphone i and tone t during the experiments:

- sound level $SL_i^{(t)}$ in decibels with all necessary adjustments related to the microphone sensitivity;
- signal-to-noise ratio $SNR_i^{(t)}$;

The reference sound level $SL_R^{(t)}$ was also directly measured for each tone by the sound level meter (reference microphone) SL-200. To compare microphone responses and assess individual sensor repeatability, the sound level detected by the i-th microphone was referred sample-by-sample to the reference sound level using a relative sound level $\Delta SL_i^{(t)}$:

$$\Delta SL_i^{(t)} = SL_i^{(t)} - SL_R^{(t)}. \tag{1}$$

Individual relative sound levels were used to compare responses of microphones. For this purpose, sets of $\Delta SL_i^{(t)}$ values were subjected to statistical analysis in terms of possible differences between pairs of microphones. For each pair $i, j = 1, \ldots, 15$; $i \neq j$ a null hypothesis H0 was formulated: differences between $\Delta SL_i^{(t)}$ and $\Delta SL_j^{(t)}$ are statistically insignificant. Based on the Shapiro-Wilk normality test, either the t-Student or U Mann-Whitney test was employed for the H0 verification. In all 840 cases (105 pairs × 8 tones), no significant difference between relative sound levels was found at $p = 0.05$. Therefore, it can be concluded that all the acoustic mask microphones record the signal in the same way.

Mean $SNR_i^{(t)}$ values are presented in Table 3. In each subtable for tone t, the microphone arrangement corresponds to the setup shown in Fig. 5. In the case of each tone, measured ratios are similar and securely acceptable for a medium-class recording equipment. The SNR parameter declared by the WM-61a manufacturer (62 dB) is preserved in most cases, with others likely decreased by the signal processing workflow.

TABLE 3. Mean $SNR_i^{(t)}$ (dB) for all frequency tones and microphones. Microphone number i in each subtable corresponds to the arrangement in Fig. 5.

t = 1 kHz				
61	63	63	63	59
62	63	61	62	62
62	63	63	64	66

t = 2 kHz				
63	65	64	64	62
62	63	62	64	64
64	64	66	65	67

t = 3 kHz				
64	63	63	64	62
64	63	63	64	65
64	63	65	64	66

t = 4 kHz				
62	62	64	63	62
63	64	62	63	64
64	64	64	65	66

t = 5 kHz				
62	59	59	61	58
61	63	63	63	62
63	63	62	63	65

t = 6 kHz				
61	61	60	63	60
59	62	62	61	60
62	62	61	62	63

t = 7 kHz				
66	64	63	66	66
62	67	67	66	65
64	65	66	64	66

t = 8 kHz				
66	64	63	66	67
63	69	67	66	66
64	65	67	64	68

2) USABILITY TESTING

The measuring stand prepared for synthetic testing was used to perform usability testing. The testing protocol was designed to assess the device's ability to detect abnormal air outflow during articulation. For this purpose, a real speaker was recorded simulating various air blows (central, left, and right outflow, each repeated three times) in ten repetitive performances. The sound source was located 8.5 cm from the central microphone. In each recording, the signals were normalized throughout all microphones into the 0–1 range, segmented, and divided into 30-millisecond frames with 15 ms overlap. Then, a root mean square value was determined over each frame in a segment S:

$$RMS_i = \sqrt{\frac{1}{B}\sum_{j=1}^{B}|b_j|^2}, \quad S = \{t_1, t_2, \ldots, t_T\}, \qquad (2)$$

where: b_j is the j-th sample of the i-th frame, B denotes the number of samples per frame, T is the number of frames within the segment S. As a result, each segment was described by a T-element set of RMS values. All microphone- and flow direction-related RMSs were grouped and analyzed, yielding distributions presented in Fig. 7.

B. SPATIAL ACOUSTIC ANALYSIS OF SPEECH

Acoustic analysis was proposed and performed with the designed data acquisition equipment and collected database. The analysis addressed the signal energy distribution for individual acoustic channels in different sibilants. For this purpose, signals were manually segmented to extract parts of recordings containing selected sibilants only. The segment duration was between 39 and 820 ms with a mean value of 172 ± 65 ms. Two sibilant sounds were analyzed, each in three different pronunciation types, most commonly found in the database (compare Table 1) – /s/:

- s_{norm} – norm,
- s_{add} – addentality,

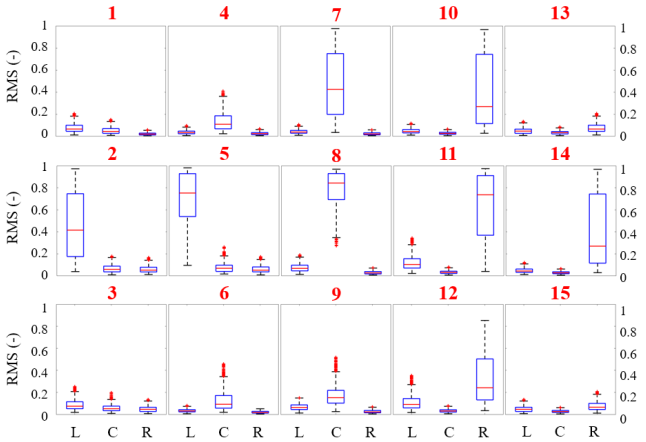

FIGURE 7. Normalized distribution of the RMS value for fifteen microphones depending on the direction of air outflow during speech: left (L), central (C), right (R). Microphone number i (red) corresponds to the arrangement in Fig. 5.

- s_{int} – interdentality

and /ʃ/:

- $ʃ_{norm}$ – norm,
- $ʃ_{int}$ – interdentality,
- $ʃ_{den}$ – dentality.

Sizes of considered groups (in terms of number of speakers and number of words) are presented in Table 4.

TABLE 4. Sizes of groups considered in the experiment.

	s_{norm}	s_{add}	s_{int}	$ʃ_{norm}$	$ʃ_{int}$	$ʃ_{dent}$
Speaker count	48	24	34	42	30	17
Word count	217	92	175	233	83	123

We decided to use a 5-channel system, as presented in Fig. 5, instead of analyzing all fifteen microphones. Our goal was to verify horizontal airflow energy distribution during pronunciation. Therefore, five uniform linear arrays (ULA) were defined to detect lateral airflow: central (C), two right (R1, R2), and two left ULAs (L1, L2). Extracted sibilant-sound-related signals from three microphones constituting a particular ULA were aggregated according to the diagram from Fig. 8.

First, each of the three ULA signals were subjected to high-pass filtering by using a 101-st order FIR filter with a cutoff frequency of 4 kHz. Such cutoff frequency was chosen to avoid near-field effect during the wave incidence angle estimation. The length of the 4 kHz wave is 8.5 cm, which reflects the closest distance between the mask's microphone and the patient's mouth. Therefore, all higher frequencies can be analyzed according to the far-field rules, which simplifies the analysis. In the far-field, the wave is considered plane, so only the wave incidence angle ought to be determined [43]. For this purpose, the time delay of arrival (TDOA) algorithm was used [44]. Its first stage employs the generalized cross-correlation with phase transform (GCC-PHAT) [45] to determine the time shift between the central and each side microphone within a ULA using maximum of their signals' correlation [46]. In the case of broadband signals, the GCC-PHAT method gives a distinct maximum and is

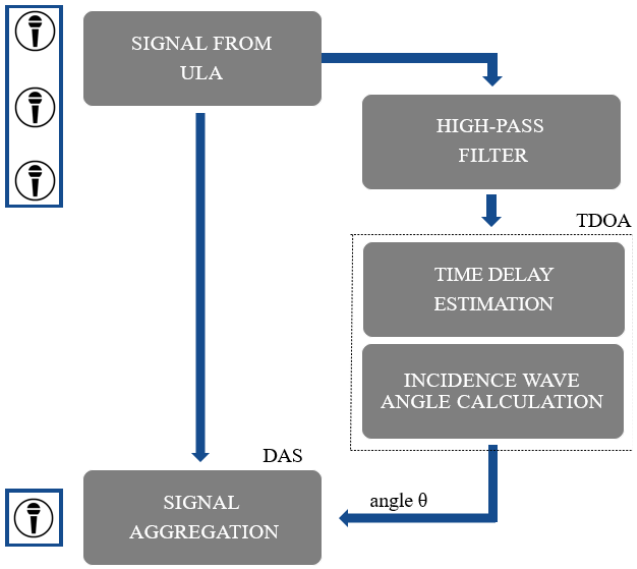

FIGURE 8. Diagram of the single ULA signal aggregation.

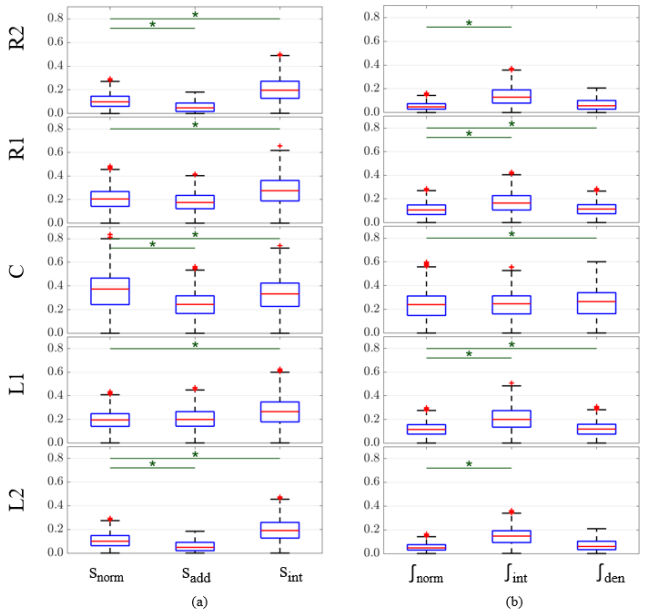

FIGURE 9. Distribution of the RMS value (-) for five ULA channels and different pronunciation types of sibilants /s/ (a) and /ʃ/ (b). ULA IDs and class IDs are explained in the text. Relation marked with a green asterisk (*) indicates a statistically significant difference between mean values at $p = 0.05$.

resistant to time-shift determination errors when processing noisy signals [47]. With a known ULA geometry, the wave incidence angle can be obtained from time shifts.

Angles calculated by the TDOA algorithm are used for signal aggregation through delay-and-sum beamforming (DAS) [48]. DAS is usually employed for shaping the sensitivity patterns of microphone arrays. Raw signals from different sensors (here, three ULA microphones) are delayed, weighted, and summed to produce a single signal. By taking into account the wave incidence angle, DAS is more sensitive to signals coming from a selected direction, and thus to attenuate unwanted sounds from other directions, including the background noise. Here, we used unit weights for all three ULA signals, limiting the aggregation to the delay and sum components. As a result, we obtained five aggregated signals for five ULAs (Fig. 5).

Each ULA signal was then subjected to pre-emphasis filtering, enhancing high frequencies (meaningful in sibilant sounds) by ca. 6 dB relative to low frequencies. Then, signals were divided into 15-millisecond frames with 10 ms overlap. The RMS values were calculated according to (2) for each sibilant-related frame of every channel. Fig. 9 presents the results. The obtained RMS sets were subjected to statistical analysis in three steps. First, distribution normality was verified by using the Kolmogorov–Smirnov test. In the case of each group, normality of distribution was confirmed at the significance level $p = 0.05$. Then, the group-to-group variance homogeneity assumption was examined by using the Brown–Forsyth test at $p = 0.05$. Finally, the H0 hypothesis about the equality of means was verified by using either one-way analysis of variance ANOVA (homogeneous variances) or Welch's ANOVA (heterogeneous variances), in both cases followed by the Tukey's range test for independent groups.

In case of the /s/ sound, statistically significant differences were noted between norm/pathology for most channels (Fig. 9(a)). No differences were found only in case of norm/addentality for the R1 and L1 channels. Less differences were found in case of the /ʃ/ sound (Fig. 9(b)). Mean RMS values for norm and dentality were statistically different in three middle channels, whereas for norm and interdentality – in all except the central one (though in this case, both variances were statistically different).

TABLE 5. Mean $SNR_c^{(t)}$ (dB) for all frequency tones and five channels. Channel order (*c* index) in each subtable corresponds to the arrangement in Fig. 5. SNR values exceeding mean value for corresponding ULA microphones (Table 3) marked with a bold green font.

t = 1 kHz				
64	**64**	**63**	**64**	**65**

t = 2 kHz				
65	**65**	**65**	**67**	**66**

t = 3 kHz				
67	**64**	**64**	**67**	**65**

t = 4 kHz				
66	62	62	**67**	**65**

t = 5 kHz				
65	**62**	**63**	**65**	**63**

t = 6 kHz				
63	**63**	**64**	**65**	**63**

t = 7 kHz				
63	62	64	**66**	63

t = 8 kHz				
66	65	**66**	**68**	**67**

Additionally, we performed the SNR experiment described for all fifteen channels in Section IV-A1 (note Table 3) for the 5-channel system. Synthetic signals from three channels constituting each ULA were aggregated with the incidence wave angle calculation based on higher-frequency tones, and the SNR was determined. Table 5 presents the results for all tones. In 33/40 cases the ULA signal aggregation yielded higher SNR than the mean SNR of its individual microphones with the mean increase equal +1.12 dB. In 20/40 cases the ULA SNR was higher than the maximum SNR over the ULA microphones.

V. DISCUSSION

As shown in Section I, speech data acquisition methods described in the literature feature various properties that can be useful in specific tasks within the speech recording and assessment domain. However, they have a number of drawbacks in terms of pronunciation evaluation in children. The availability and cost of these systems remains an issue. Moreover, interfering with the articulatory organs of the speaker is usually considered unacceptable.

Therefore, we designed and built a speech acquisition system dedicated to non-invasive, multichannel registration of speech for pronunciation evaluation. The system was validated in several experiments, proving its ability to reliably record speech signals in multiple channels with a satisfying signal-to-noise ratio indicating sufficient reduction of noise from other sound sources, echo, and reverberation effects. All WM-61a sensors were successfully examined for appropriate and comparable signal acquisition. The device meets ergonomic and safety requirements of the target age group – preschool children. The system was consulted and approved by speech therapy experts experienced in childcare. The microphone mounting is flexible to the subject's mouth. It allows adjusting the mask size to individual needs while maintaining the position between sessions. It does not interfere with the motion of the child's articulators, being also visually friendly and attractive. However, the stability of the mask's fixation on the child's head remains an issue, as well as its acoustic sensitivity to loud sounds or touching its components.

The developed measuring device was employed to record the speech signal in the process of creating a speech-articulation corpus. The speech corpora development is a labor-consuming and time-consuming process. Thus, audio resources for speech analysis systems are frequently acquired from already existing data, e.g., radio or television recordings. However, as far as pathological speech could also be collected this way, it could not be reliably annotated with pathology description. Therefore, in most cases, non-normative speech corpora are registered for the needs of specific research problems and very rarely contain children's speech. According to our knowledge, except our research, there is only one published study on sigmatism based on data acquired from children [10]. The database employed for that research contained single-channel recordings of four Portuguese sibilants pronounced by children and annotated as correct or incorrect. The corpus collected as a part of our study contains specific diagnoses provided by a team of speech pathologists; more detailed annotation provides opportunities for more in-depth acoustic analysis. Moreover, speech samples were registered with the multichannel measuring device, which allows the use of spatial processing techniques and inference.

The device was examined for signal energy distribution over acoustic channels. The experiment involved sibilant sound speech samples produced by children with different pronunciation characteristics: normative and pathological. As a result, in multiple acoustic channels, we obtained significantly different mean values of signal energy in different realizations of sibilants. Thus, a spatial speech signal can serve as an indicator of abnormal pronunciation or air outflow – pronunciation characteristics which are not present in a single channel speech signal.

The results of the acoustic analysis can be explained by articulatory features related to the realization of sibilants /s/ and /ʃ/. The tongue apex plays a key role in the correct pronunciation. In particular, its position is important in forming a gap in relation to the palate [49]. The narrow gap guarantees correct pronunciation of /s/ and /ʃ/ sounds, which is reflected in the central direction of the airflow with higher (/s/) or lower (/ʃ/) energy (note, however, that the generated noise falls in different high-frequency bands). The energy recorded by side channels is much lower. The lack of dentalization (close positioning of the jaws) and the abnormal position of the tongue apex prevent the gap from being formed correctly, and the air comes out with a wide stream, stimulating the microphones of the side channels to a greater extent. This is particularly noticeable during the interdental and addental realization of sibilant /s/ (s_{int}, s_{add}), especially in the former case the outflow is wide and the external channels (R2, L2) feature high energy.

Obtained results confirm the initial assumption and goal of the study and encourage us to continue data collection, processing, and analysis. The signal energy expressed by the RMS value of the signal can indicate differences between realizations of a given sibilant. So, it is likely that way more valuable information can be found in other signal features, e.g., spectral, cepstral, or spatiotemporal. The significant part of the sibilant sound spectrum lies over 1 kHz, thus the spectral preprocessing is reasonable for limiting the impact of a possible vowel or consonant co-articulation. Furthermore, automated extraction of features from different signal representations can be assumed employing deep learning techniques, initially proposed and investigated in previous studies [50]. The latter can be applied and trained over either raw or preprocessed signals, their 2D representations (spectrograms), or multidimensional data incorporating, e.g., spatial information on acoustic channels. The ultimate goal is to formulate conclusions and develop a speech therapy articulation standard as well as to develop a computer-aided speech therapy diagnostic tool for practical use.

Presented experiments were performed on Polish speech data. Sibilant sounds and sigmatism occur in most of the existing languages. However, there are interlanguage differences in articulatory patterns as well as in the occurrence of specific sibilants. Despite that analyzed sounds (/s/ and /ʃ/) are often considered basic and are commonly found, they may be pronounced differently in different countries [51]. It should be noted, though, that general characteristics of sibilants and the most frequent pathologies in their pronunciation remain the same. Therefore, the proposed

measurement and analysis techniques are highly probable to be successfully employed for other languages than Polish.

VI. CONCLUSION

Our data acquisition device is able to provide data to prepare spatial models of articulation for different purposes, e.g., identification and redefinition of phone articulation stages, or pronunciation pathology detection and binary or multiclass classification. Various temporal, spectral, or hybrid representations of the spatial speech signal can be used. We can think of multiple advanced data processing techniques to be employed, e.g., machine learning or deep learning tools. Such models can support both linguistic and speech therapy research. The above thoughts set directions for our future research.

REFERENCES

[1] E. Minczakiewicz, "Dyslalia in the context of other speech defects and disorders in preschool and school children, (PL) Dyslalia na tle innych wad i zaburzeń mowy u dzieci w wieku przedszkolnym i szkolnym," *Konteksty Pedagogiczne*, vol. 1, no. 8, pp. 149–169, 2017.

[2] J. Trzaskalik, *Respiratory System Diseases as a Cause of Dysplasia—Study of Children From the Silesian Agglomeration, (PL) Choroby Układu oddechowego jako przyczyna dyslalii—na przykładzie dzieci z aglomeracji śląskiej*. Wrocław, Poland: Oficyna Wydawnicza ATUT—Wrocławskie Wydawnictwo Oświatowe, 2012, pp. 67–85.

[3] E. Skorek, *Faces of Speech Disorders, (PL) Oblicza Wad Wymowy*. Warszawa, Poland: Wydawnictwo Żak, 2001.

[4] G. Demenko, A. Wagner, and N. Cylwik, "The use of speech technology in foreign language pronunciation training," *Arch. Acoust.*, vol. 35, no. 5, pp. 309–330, 2010.

[5] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, and K. Hirose, "Automatic chinese pronunciation error detection using SVM trained with structural features," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Kyoto, Japan, Dec. 2012, pp. 473–478. [Online]. Available: https://ieeexplore.ieee.org/document/6424270

[6] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Proc. INTERSPEECH ISCA*, 2007, pp. 1837–1840. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2007.html#Stri%kTWC07

[7] S. Xu, J. Jiang, Z. Chen, and B. Xu, "Automatic pronunciation error detection based on linguistic knowledge and pronunciation space," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Washington, DC, USA, Apr. 2009, pp. 4841–4844, doi: 10.1109/ICASSP.2009.4960715.

[8] 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.

[9] C. Valentini-Botinhao, S. Degenkolb-Weyers, A. Maier, E. Noeth, U. Eysholdt, and T. Bocklet, "Automatic detection of sigmatism in children," in *Proc. Workshop Child, Comput. Interact. (WOCCI)*, Portland, OR, USA, 2012, pp. 1–4.

[10] I. Anjos, M. Grilo, M. Ascensão, I. Guimarães, J. Magalhães, and S. Cavaco, "A model for sibilant distortion detection in children," in *Proc. DMIP*, 2018, pp. 42–47.

[11] S. M. Mirhassani and H.-N. Ting, "Fuzzy-based discriminative feature representation for children's speech recognition," *Digit. Signal Process.*, vol. 31, pp. 102–114, Aug. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200414001407

[12] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Commun.*, vol. 49, no. 12, pp. 861–873, Dec. 2007, doi: 10.1016/j.specom.2007.05.004.

[13] 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.

[14] 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

[15] I. Steiner, K. Richmond, and S. Ouni, "Speech animation using electromagnetic articulography as motion capture data," *Comput. Res. Repository*, vol. abs/1310.8585, 2013. [Online]. Available: http://arxiv.org/abs/1310.8585

[16] M. L. Berthier and I. Moreno-Torres, "Commentary: Visual feedback of tongue movement for novel speech sound learning," *Frontiers Hum. Neurosci.*, vol. 10, p. 612, Dec. 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2015.00612

[17] 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.

[18] 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.

[19] Complete Speech. (Apr. 5, 2020). *Visual Speech Therapy SmartPalate*. Accesssed: May 21, 2020. [Online]. Available: http://completespeech.com/

[20] D. Król and A. Lorenc, "Acoustic field distribution in speech with the use of the microphone array," *Sci., Technol. Innov.*, vol. 4, no. 3, pp. 9–16, Oct. 2017.

[21] D. Król, A. Lorenc, and R. Święciński, "Detecting laterality and nasality in speech with the use of a multi-channel recorder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5147–5151.

[22] L. Mik, R. Wielgat, D. Król, R. Jedryka, A. Lorenc, and R. Święciński, "Multimodal speech data acquisition with the use of EMA, fast-speed video cameras and a dedicated microphone array," in *Proc. 23rd Int. Conf. Mixed Design Integr. Circuits Syst. (MIXDES)*, Jun. 2016, pp. 415–418.

[23] L. Mik, A. Lorenc, D. Król, R. Wielgat, R. Święciński, and R. Jedryka, "Fusing the electromagnetic articulograph, high-speed video cameras and a 16-channel microphone array for speech analysis," *Bull. Polish Acad. Sci. Tech. Sci.*, vol. 66, no. 3, pp. 257–266, 2018.

[24] 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.

[25] M. Aron, M.-O. Berger, E. Kerrien, B. Wrobel-Dautcourt, B. Potard, and Y. Laprie, "Multimodal acquisition of articulatory data: Geometrical and temporal registration," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. 636–648, Feb. 2016, doi: 10.1121/1.4940666.

[26] M. Kręcichwost, Z. Miodońska, J. Trzaskalik, J. Pyttel, and D. Spinczyk, "Acoustic mask for air flow distribution analysis in speech therapy," in *Information Technologies in Medicine*. Cham, Switzerland: Springer, 2016, pp. 377–387.

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

[28] Panasonic. (Apr. 5, 2020). *Omnidirectional Back Electret Condenser Microphone Cartridge, Series: WM-61A, WM-61B*. Accesssed: May 21, 2020. [Online]. Available: http://konektor.nazwa.pl/serwisowe/panasonic-wm-61a.pdf

[29] G. Danavaras, "Testing Panasonic's WM-61A mike cartridge," in *audioXpress*, 2007. [Online]. Available: https://issuu.com/150176/docs/audioexpress_-_testing_panasonic_wm

[30] M. J. Schloneger and E. J. Hunter, "Assessments of voice use and voice quality among college/university singing students ages 18–24 through ambulatory monitoring with a full accelerometer signal," *J. Voice*, vol. 31, no. 1, p. 124–e21, 2017.

[31] Z. Šarić, M. Subotić, R. Bilibajkić, and M. Barjaktarović, "Bidirectional microphone array with adaptation controlled by voice activity detector based on multiple beamformers," *Multimedia Tools Appl.*, vol. 78, no. 11, p. 15 235–15 254, 2019.

[32] E. Milanova and E. Milanov, "Proximity effect of microphone," in *Audio Engineering Society Convention 110*. May 2001. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=9940

[33] Maxim Integrated Products. (Apr. 5, 2020). *MAX9812/MAX9813 Datasheets*. Accesssed: May 21, 2020. [Online]. Available: https://datasheets.maximintegrated.com/en/ds/MAX9812-MAX9813L.pdf

[34] Measurement Computing Corporation. (Apr. 5, 2020). *DaqBoard/3000USB Series*. Accesssed: May 21, 2020. [Online]. Available: https://www.mccdaq.com/products/db3000usbs

[35] D. Antos, G. Demel, and I. Styczek, *How to Treat Lisp Other Speech Disorders, (PL) Jak usuwac seplenienie i inne wady wymowy*. Warszawa, Poland: Wydawnictwa Szkolne i Pedagogiczne, 1978.

[36] E. Słodownik-Rycaj, *O Mowie Dziecka. Jak Zapobiegac Powstawaniu Nieprawidlowosci Wjej Rozwoju*. Warszawa, Poland: Wydawnictwo Akademickie Żak, 2000.
[37] A. Sołtys-Chmielowicz, *Articulation Disorders. Theory and Practice, (PL) Zaburzenia artykulacji. Teoria i praktyka*. Warszawa, Poland: Oficyna Wydawnicza Impuls, 2016.
[38] *Acoustics—Determination of Sound Power Levels and Sound Energy Levels of a Noise Source on the Basis of Sound Pressure Measurements—An Indicative Method Using the Surrounding Measuring Surface Above the Reflecting Plane, (PL) Akustyka—Wyznaczanie poziomów mocy akustycznej i poziomów energii akustycznej źródel hałasu na podstawie pomiarów ciśnienia akustycznego—Metoda orientacyjna z zastosowaniem otaczającej powierzchni pomiarowej nad płaszczyzną odbijającą dźwięk*, Standard PN-EN ISO 3746:2011, Polish Committee for Standardization, 2017. [Online]. Available: https://sklep.pkn.pl/fileuploader/download/download/?d=0&fileId=7727
[39] RIGOL. *DG1022 Series Function/Arbitrary Waveform Generators Datasheet*. [Online]. Available: https://www.rigolna.com/products/waveform-generators/dg1000/
[40] Voltcraft. *Decibel Meter SL-200 Datasheet*. [Online]. Available: https://www.conrad.com/p/voltcraft-sound-level-meter-sl-200-30-130-db-3%15-hz-8-khz-100805
[41] *Acoustics—Determination of Sound Power Levels of Noise Sources—Guidelines for the Use of Basic Standards*, Standard ISO 3740:2019(en), International Organization for Standardization, 2019. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:3740:ed-3:v1:en
[42] J. Zhang, Y. Zhang, Z. Zhou, and M. Fang, "Design and analysis of acoustic reforms of studio," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 399, Sep. 2018, Art. no. 012060, doi: 10.1088%2F1757-899x%2F399%2F1%2F012060.
[43] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Cham, Switzerland: Springer, Jun. 2001.
[44] N. M. Kwok, J. Buchholz, G. Fang, and J. Gal, "Sound source localization: Microphone array design and evolutionary estimation," in *Proc. IEEE Int. Conf. Ind. Technol.*, Dec. 2005, pp. 281–286.
[45] M. Imran, A. Hussain, N. M. Qazi, and M. Sadiq, "A methodology for sound source localization and tracking: Development of 3D microphone array for near-field and far-field applications," in *Proc. 13th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2016, pp. 586–591.
[46] M. Rhudy, B. Bucci, J. Vipperman, J. Allanach, and B. Abraham, "Microphone array analysis methods using cross-correlations," in *Proc. Int. Mech. Eng. Congr. Expo. (ASME)*, vol. 15, Nov. 2009, pp. 281–288, doi: 10.1115/IMECE2009-10798.
[47] B. Kwon, Y. Park, and Y. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *Proc. Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2010, pp. 2070–2073.
[48] M. Omologo, M. Matassoni, and P. Svaizer, *Speech Recognition With Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 331–353, doi: 10.1007/978-3-662-04619-7_15.
[49] T. Yoshinaga, K. Nozaki, and S. Wada, "A simplified vocal tract model for articulation of [s]: The effect of tongue tip elevation on [s]," *PLoS ONE*, vol. 14, no. 10, pp. 1–11, Oct. 2019, doi: 10.1371/journal.pone.0223382.
[50] A. Woloshuk, M. Kręcichwost, Z. Miodońska, D. Korona, and P. Badura, "Convolutional neural networks for computer aided diagnosis of interdental and rustling sigmatism," in *Proc. Inf. Technol. Biomed. (ITIB)*. Cham, Switzerland: Springer, 2019, pp. 179–186.
[51] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages* (Phonological Theory). Hoboken, NJ, USA: Wiley, 1996.

MICHAL KRECICHWOST was born in Bielsko-Biala, Poland, in 1990. He received the B.S. and M.S. degrees in biomedical engineering from the Silesian University of Technology, Gliwice, Poland, in 2014. He is currently pursuing the Ph.D. degree in biomedical engineering with the Silesian University of Technology.

From 2015 to 2017, he was a Programmer with the Salus Publica—the Foundation for Public Health, Krakow, Poland. Since 2019, he has been a Research Assistant with the Faculty of Biomedical Engineering, Silesian University of Technology. 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 20 articles. His research interests include computer-aided speech diagnosis and therapy, signal processing, deep learning, and electronics.

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

From 2014 to 2019, she was a Research Assistant with the Faculty of Biomedical Engineering, Silesian University of Technology. From 2015 to 2017, she was a Research Assistant with the Salus Publica—the Foundation for Public Health, Krakow, Poland. Since 2019, she has been an Assistant Professor 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.

JOANNA TRZASKALIK received the M.S. degree in linguistics from the University of Gdansk, Poland, in 1984, and the Ph.D. degree in linguistics from the University of Silesia, Katowice, Poland, in 2009.

From 1985 to 1998, she was a Research Assistant with the University of Silesia. From 1998 to 2016, she worked for the Upper Silesian Pedagogical School, as a Research Assistant, until 2009, and as an Assistant Professor since then. Since 2016, she has been an Assistant Professor with Jesuit University Ignatianum, Krakow, Poland. She is the author of a monograph and multiple research articles. Since 1986, she has been an active speech therapist. She focuses on the theory and practice of speech disorders in dyslalia, in particular lateral sigmatism. Her research interests include also delayed speech development, reading and writing disorders. She is currently an Editor-in-Chief of the journal *Forum Logopedyczne (Logopedic Forum)*, for years she was an Editor of the journal *Nauczyciel i Szkoła (Teacher and School)*.

Since 2015, she has been a Chair of the Silesian Branch of the Polish Speech Therapy Society.

PAWEL BADURA was born in Katowice, Poland, in 1980. He received the B.S. and M.S. degrees in electronics and telecommunication,, the Ph.D. degree in biomedical engineering, and the D.Sc. degree 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 worked for the Faculty of Biomedical Engineering, Silesian University of Technology, Gliwice, Poland, as an Assistant Professor, until 2018, and as an Associate Professor since then. He took part in nine research projects. He currently 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 60 articles. His research interests include image analysis, signal processing, computer-aided diagnosis systems, artificial intelligence, and machine learning methods. He is the Guest Editor of the journal *Applied Sciences* and a Co-Editor of five books.

Prof. Badura was a recipient of the distinction in the biennial contest for the best Ph.D. thesis in the domain of image processing, from 2006 to 2007, organized by the Association for Image Processing.