scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Polish multichannel audio-visual child speech dataset with double-expert sigmatism diagnosis

Michal Krecichwost[1], Zuzanna Miodonska[1], Agata Sage[1]✉, Joanna Trzaskalik[2], Ewa Kwasniok[3] & Pawel Badura[1]

The paper introduces PAVSig: Polish Audio-Visual child speech dataset for computer-aided diagnosis of Sigmatism (lisp). The study aimed to gather data on articulation, acoustics, and visual appearance of the articulators in different child speech patterns, particularly in sigmatism. The data was collected in 2021–2023 in six kindergarten and school facilities in Poland during the speech and language therapy examinations of 201 children aged 4–8. The diagnosis was performed simultaneously with data recording, including 15-channel spatial audio signals and a dual-camera stereovision stream of the speaker's oral region. The data record comprises audiovisual recordings of 51 words and 17 logotomes containing all 12 Polish sibilants and the corresponding speech and language therapy diagnoses from two independent speech and language therapy experts. In total, we share 66,781 audio-video segments, including 12,830 words and 53,951 phonemes (12,576 sibilants).

Background & Summary

Computer-aided speech diagnosis (CASD) or computer-aided speech therapy (CAST) is an emerging field with access to data acquisition and processing capabilities. It is strongly connected to particular languages, as it relies on acoustics and articulation, which are significantly different in various language groups[1]. The distinction can also concern the speaker's age, as the therapy can involve children, adolescents, or adults. Nonetheless, the earlier the diagnosis occurs, the more efficient the therapy can be. Therefore, it is essential to support the speech and language therapist/pathologist (SLP) in treating even very young children[2]. Sigmatism (lisp) is one of the most common types of speech sound disorders[3–12]. It refers to an incorrect articulation of sibilant sounds, different across languages. In Polish, there are 12 sibilants: four denti-alveolars /s/, /z/, /t͡s/, /d͡z/, four retroflexes /ʂ/, /ʐ/, /t͡ʂ/, /d͡ʐ/; and four alveolo-palatals /ɕ/, /ʑ/, /t͡ɕ/, /d͡ʑ/[13].

Many detailed features can describe articulation, e.g., the manner of articulation, active and passive articulators, airflow direction, or voicing[14–17]. The measurement methods to systematize and objectify the process of assessing speech production still need to be better defined. CASD tools can support the therapy by providing additional information to the therapist or improve speech screening tests in schools and kindergartens. Finally, automated articulation analysis can be implemented in applications for speech exercises that can be performed autonomously at home between speech and language therapy (SLT) sessions.

To be applicable in practice, CASD systems have to be based on data that are reliable and easily recordable without disrupting natural articulation. That excludes many specialized invasive systems used in articulation research, like electromagnetic articulography (EMA)[18,19] or electropalatography (EPG)[20,21]. On the other hand, acoustic analysis has been performed in this area for years[11,16,17,22–25]. The other, less common idea involves a video recording of the speaker's face to monitor the appearance of articulatory organs (articulators): tongue, lips, or teeth[26–30]. Both audio and video can be recorded every day using devices that offer sufficiently good quality. Access to annotated audiovisual (AV) data related to normal and distorted child speech is necessary to make the analysis reliable and repeatable for possible training and validation of automated CASD models.

This study was a part of research project no. 2018/30/E/ST7/00525: "Hybrid System for Acquisition and Processing of Multimodal Signal in the Analysis of Sigmatism in Children", financed by National Science Center, Poland, in 2019–2024. It aimed at finding relationships between articulation, acoustics, and visual appearance of

[1]Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800, Zabrze, Poland. [2]Jesuit University Ignatianum in Krakow, Kopernika 26, 31-501, Kraków, Poland. [3]Center for Human Health and Development "Therapy", Klodnicka 2, 41-706, Ruda Slaska, Poland. ✉e-mail: agata.sage@polsl.pl

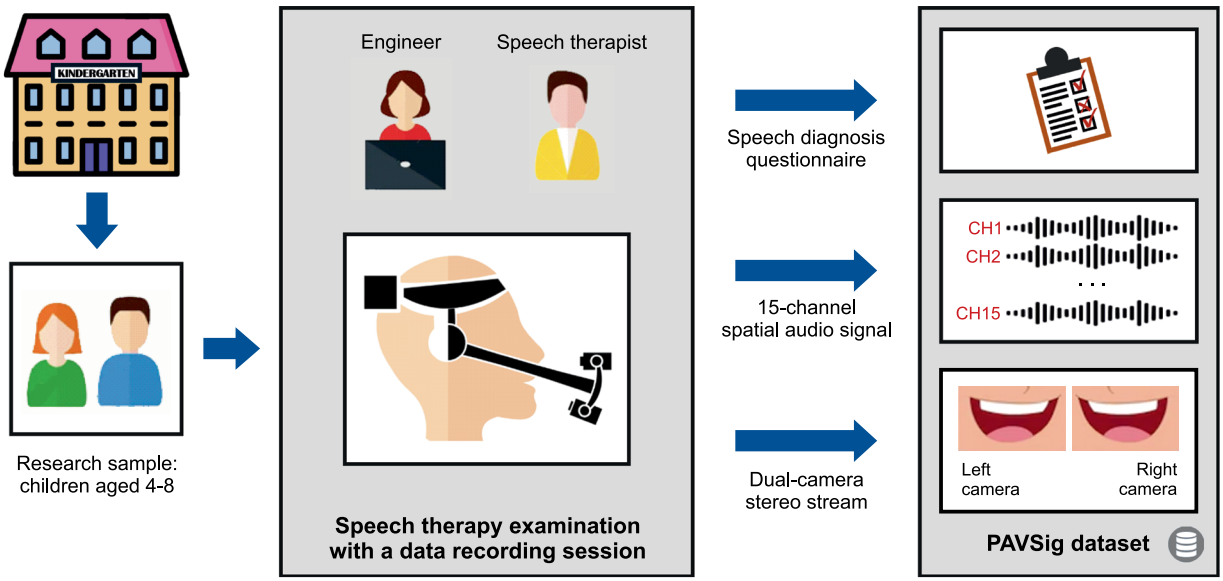


Fig. 1 Schematic overview of the study.

Age (years)	Average age (months)	Number of children		
		Girls	Boys	Total
4–5	58.7 ± 0.6	1	2	3
5–6	66.6 ± 3.4	37	23	60
6–7	77.7 ± 3.6	48	48	96
7–8	87.9 ± 3.5	21	21	42
Aggregated	76.2 ± 8.6	107	94	201

Table 1. Specification of the research sample.

the articulators in different child speech patterns. A literature review and available solutions showed a need for an adequate dataset for the Polish language. Therefore, we prepared a detailed framework for the SLT examination with a data recording session, including multichannel spatial audio signals and a dual-camera stereovision stream of the speaker's oral region (Fig. 1). As a result, we collected an extensive multimodal PAVSig (Polish Audio-Visual speech dataset for computer-aided diagnosis of Sigmatism) dataset of 201 children aged 4–8, along with the corresponding SLT diagnoses from two independent experts.

Methods

Research sample. Our interdisciplinary research team, including biomedical engineers and SLPs, performed the SLT examinations and data recording sessions in six kindergarten and school facilities in Myslowice, Katowice, Ruda Slaska, and Zabrze, Poland, from October 2021 to June 2023. The research sample covered 208 children, but several factors (non-native Polish speakers, data acquisition failures, others) limited it to 201 (107 girls and 94 boys) aged 4–8 (see Table 1). Including the child in the research sample required written consent from their parents or legal guardians to participate in the study and share the data as described in this paper. The child also had to agree verbally to participate in the study. The exclusion criteria covered: (1) diagnosed disabilities, including hard of hearing, deafness, low vision, visual impairment, aphasia, autism spectrum disorder, intellectual disability, and (2) epilepsy record. The study received a positive recommendation from the Bioethics Committee for Scientific Research at the Jerzy Kukuczka University of Physical Education in Katowice, Poland (Decision No. 3/2021).

Speech and language therapy assessment with a data recording session. The assessment consisted of three stages, with two involving data recording:

1. In the first part, a dedicated multimodal data acquisition device (MDAD) registered the child's speech while naming pictures visible on the screen (Fig. 2a).
2. In the second part, the speaker was recorded while repeating selected words and one- or two-syllable logotomes following the SLP. This stage also involved various tongue movements, swallowing, or smiling (Fig. 2b).
3. The third part was the SLT examination according to the dedicated diagnostic protocol for sigmatism-related speech assessment (Fig. 2c). It was performed by the SLP, and no data was recorded at this point.

(a) part 1 **(b)** part 2 **(c)** part 3

Fig. 2 Illustration of the SLT examination with a data recording session: (**a**) part 1: data recording while naming graphics visible on the screen; (**b**) part 2: data recording while repeating words or logotomes and undergoing SLT assessment; (**c**) part 3: SLT examination. The individuals in the pictures or their legal guardians consented to publishing their image in the manuscript.

Each session produced a record of multimodal data (15-channel spatial audio and dual-camera video stream) and a filled diagnostic questionnaire of the case. At least two biomedical engineers and one SLP were present during each recording session. The second SLP prepared their independent diagnosis another day without data recording and with no access to the previous assessment outcomes.

Speech corpus. The sibilant-related linguistic material prepared and collected in the study consisted of 51 words and 12 one-syllable logotomes containing all Polish sibilants (see Table 2). The corpus organized isolated words with sibilants in different phases of articulation (word positions): at the beginning, in the middle, and at the end of the phrase[31,32] (final positions only for voiceless sibilants). Our assumption was to use words where sibilants are surrounded by a vowel /a/ wherever possible. However, the priority was that the words were known and unambiguous to a preschool-age child and easily graphically represented, as the child's task was to name the object they saw in a picture. During the language selection process, we encountered a disproportion concerning the presence of different sibilants in words applicable in picture naming. We reviewed picture tests available for Polish children and considered their use, but the sounds' distributions were diversified as well. Some tests (e.g., from Krajna and Bryndal[33]) employed words featuring sibilants in different neighborhoods or featuring facultative pronunciation (e.g., jam, Polish: *dżem*, which may be produced with an affricate or asynchronously: /d͡ʑɛm/ or /dʑɛm/). Additionally, we have included a set of words that did not follow the described selection criteria, but we have used them in our previous experiments, and we consider them an added value to the core dictionary. We assume that the language material covered in the database may be used for different purposes and filtered according to the needs.

Most words (38) were displayed graphically on the screen in part 1 of the examination—a single illustration accompanied each word. Due to their difficulty and graphic ambiguity, the remaining words (13) and logotomes (12) were produced by the SLP in part 2 of the examination, while the speaker's task was to repeat the phrase. The word order was the same in all measurements. Since four words contain two sibilants each (/ˈkɕɔw̃ʂka/—book, /ˈstrazʲak/—firefighter, /ˈt͡ɕastka/—cookies, /saˈd͡zafka/—pond), the total number of unique occurrences of sibilants in the speech corpus is 67.

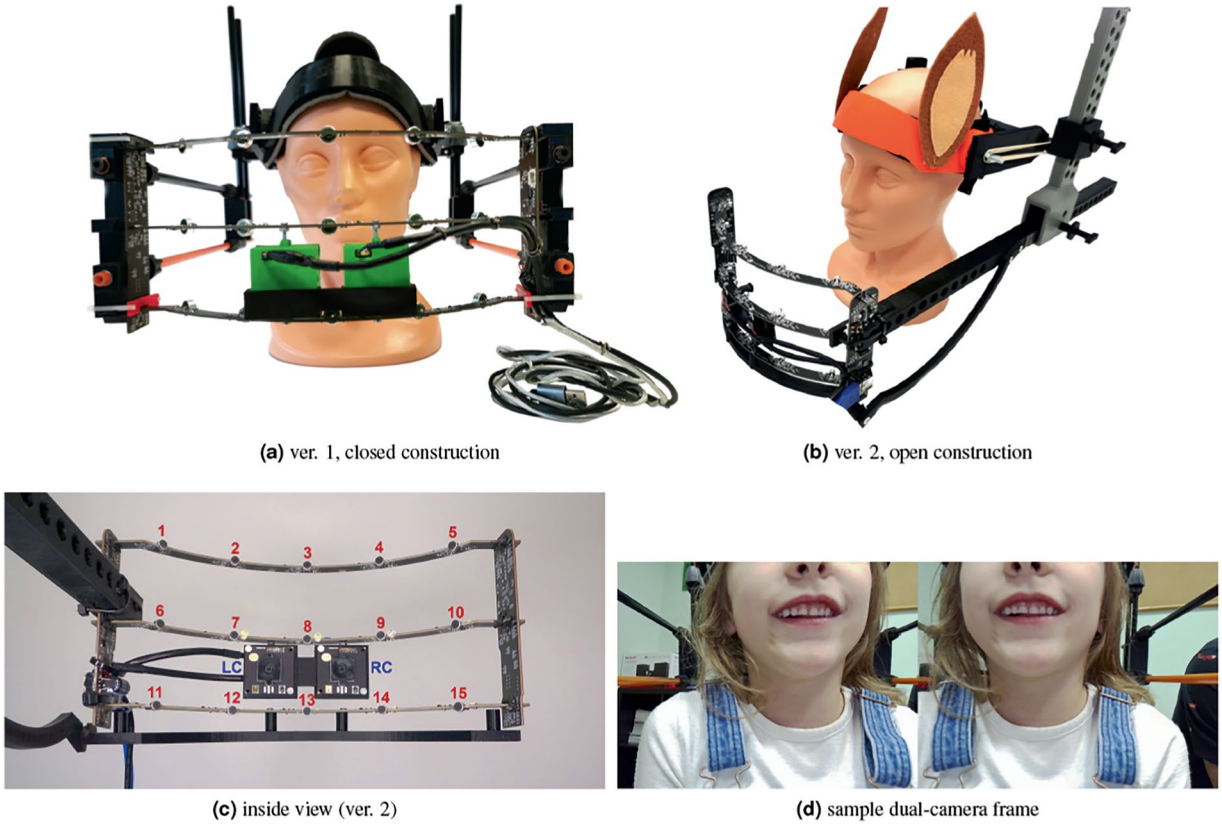
There were five more types of logotomes with vowels only in the speech corpus (Table 2, bottom section). They can be used as an additional, sibilant-free resource for the articulation assessment.

Multimodal data acquisition device. We collected the data using a dedicated, self-designed multimodal data acquisition device (MDAD, Fig. 3). It was invented and redesigned multiple times with the milestone versions described and validated in[34,35] (Fig. 3a). The only adjustments introduced in the most recent device were the construction updates, e.g., to reduce weight or improve visual user-friendliness (Fig. 3b). The equipment records the audio signal from 15 spatially distributed channels (a semicylindrical microphone array) and captures the video data using a dual-camera stereovision module (Fig. 3c).

Word (PL)	IPA	Word (EN)	P	#	Word (PL)	IPA	Word (EN)	P	#	Word (PL)	IPA	Word (EN)	P	#
/s/					/ʂ/					/ɕ/				
pie**s**	/ˈpʲɛs/	dog	1	182	**sz**afa	/ˈʂafa/	wardrobe	1	191	k**si**ążka	/ˈkɕɔw̃ʂka/	book	1	190
strażak	/ˈstraʐak/	firefighter	1	186	**sz**ufelka	/ʂuˈfɛlka/	scoop	1	145	**si**atka	/ˈɕatka/	net	1	167
samolot	/saˈmɔlɔt/	airplane	1	196	ko**sz**yk	/ˈkɔʂik/	basket	1	195	w pa**si**e	/fˈpaɕɛ/	in waist	2a	198
sałata	/saˈwata/	lettuce	1	161	kalo**sz**e	/kaˈlɔʂɛ/	rain boots	1	190	pa**ź**	/ˈpaɕ/	pageboy	2a	199
para**s**ol	/paˈrasɔl/	umbrella	1	196	nó**ż**	/ˈnuʂ/	knife	1	196	**si**a	/ˈɕa/	—	2b	197
la**s**	/ˈlas/	forest	1	180	wą**ż**	/ˈvɔw̃ʂ/	snake	1	195					
ciastka	/ˈt͡ɕastka/	cookies	1	146	książ**k**a	/ˈkɕɔw̃ʂka/	book	1	190					
sadzawka	/saˈd͡zafka/	pond	2a	196	leka**rz**	/ˈlɛkaʂ/	physician	1	142					
sa	/ˈsa/	—	2b	198	**sz**nurek	/ˈʂnurɛk/	string	1	142					
					szalik	/ˈʂalik/	scarf	1	193					
					kuchа**rz**	/ˈkuxaʂ/	cook	1	192					
					ka**sz**a	/ˈkaʂa/	groats	2a	198					
					sza	/ˈʂa/	—	2b	194					
/z/					/ʐ/					/ʑ/				
ko**z**a	/ˈkɔza/	goat	1	178	żarówka	/ʐaˈrufka/	bulb	1	171	**zi**arno	/ˈʑarnɔ/	grain	2a	198
zegar	/ˈzɛgar/	clock	1	199	**rz**eka	/ˈʐɛka/	river	1	174	ba**zie**	/ˈbaʑɛ/	catkins	2a	199
zabawki	/zaˈbafci/	toys	1	190	je**ż**e	/ˈjɛʐɛ/	hedgehogs	1	178	**zi**a	/ˈʑa/	—	2b	196
ma**z**aki	/maˈzaci/	markers	1	167	ró**ż**a	/ˈruʐa/	rose	1	171					
za	/ˈza/	—	2b	196	strażak	/ˈstraʐak/	firefighter	1	186					
					ży**r**afa	/ʐiˈrafa/	giraffe	1	193					
					ża**b**a	/ˈʐaba/	frog	1	189					
					wa**rz**ywa	/vaˈʐiva/	vegetables	1	186					
					rza	/ˈʐa/	—	2b	193					
/t͡s/					/t͡ʂ/					/t͡ɕ/				
cebula	/t͡seˈbula/	onion	1	197	**cz**apka	/ˈt͡ʂapka/	cap	1	196	**ci**astka	/ˈt͡ɕastka/	cookies	1	146
owo**c**e	/ɔˈvɔt͡se/	fruits	1	196	ka**cz**ka	/ˈkat͡ʂka/	duck	1	188	bo**ci**an	/ˈbɔt͡ɕan/	stork	1	190
widele**c**	/vʲiˈdɛlɛt͡s/	fork	1	197	biega**cz**	/ˈbʲɛgat͡ʂ/	runner	2a	199	łokieć	/ˈwɔcɛt͡ɕ/	elbow	2a	198
ta**c**a	/ˈtat͡sa/	tray	2a	193	**cz**a	/ˈt͡ʂa/	—	2b	193	**ci**a	/ˈt͡ɕa/	—	2b	197
paja**c**	/ˈpajat͡s/	clown	2a	197										
ca	/ˈt͡sa/	–	2b	199										
/d͡z/					/d͡ʐ/					/d͡ʑ/				
dzwonek	/ˈd͡zvɔnɛk/	bell	1	199	**dż**okej	/ˈd͡ʐɔkɛj/	jockey	2a	197	**dzi**adek	/ˈd͡ʑadɛk/	grandfather	1	191
sa**dz**awka	/saˈd͡zafka/	pond	2a	196	ra**dż**a	/ˈradʐa/	raja	2a	198	ło**dzie**	/ˈwɔd͡ʑɛ/	boats	2a	190
dza	/ˈd͡za/	—	2b	198	**dż**a	/ˈd͡ʐa/	—	2b	194	**dzi**a	/ˈd͡ʑa/	—	2b	196
Vowels														
u	—	—	2b	199										
i	—	—	2b	195										
a	—	—	2b	199										
iu	—	—	2b	195										
ia	—	—	2b	196										

Table 2. A set of words with highlighted sibilants. Column 'P' refers to the part of the examination: 1 – words presented on the screen and named by the child, 2a, 2b – words and logotomes, respectively, repeated by the child following the SLP. Words and logotomes are organized within sibilants. The bottom section contains vowels. Column "#" displays the number of word/logotome occurrences in the dataset. IPA stands for the International Phonetic Alphabet.

The MDAD comprises a 5 V-powered central unit and three recording arcs (Fig. 3c). Each arc uses five electret Panasonic WM-61a microphones with omnidirectional characteristics[36]. Fifteen audio signals are recorded at a 44.1 kHz sampling rate and synchronized in time in a semicylindrical 3 × 5 array with ca. 5 cm distances



(a) ver. 1, closed construction

(b) ver. 2, open construction

(c) inside view (ver. 2)

(d) sample dual-camera frame

Fig. 3 Multimodal data acquisition device: (**a**) closed construction, prototype from[35]; (**b**) open construction, recent version; (**c**) inside view to the measuring part; red numbers present the microphone (audio channel) numbers, "LC" and "RC" indicate the left and right camera, respectively; (**d**) sample dual-camera view; the picture comes from horizontal concatenation of the left and right camera frame. The individuals in the pictures or their legal guardians consented to publishing their image in the manuscript.

Audio		**Video**	
Number of audio channels	15	Number of cameras	2
Audio sampling rate	44.1 kHz	Video frame rate	30 fps
Microphone type	Panasonic WM-61A (electret)	Camera type	Arducam 8MP 1080P Auto Focus
Microphone bandpass	20 Hz–16 kHz	Camera resolution	640 × 480 VGA
Sound pressure level	120 dB		
Signal-to-noise ratio	62 dB		
Sensitivity (1 kHz, 94 dB SPL):	-35 ± 4 dB		

Table 3. Technical parameters of the multimodal data acquisition device.

between adjacent microphones. A pair of Arducam 8MP 1080P Auto Focus cameras[37] installed between two bottom arcs are used to produce a stereovision stream. Both capture an unobstructed view of the articulators during speech production from a distance of ca. 15 cm at 30 frames per second with a 480 × 640 resolution each (Fig. 3d). We added LED lighting to improve the quality of image data and illuminate the speaker's oral area. The main technical parameters are given in Table 3. The software for data recording was developed in Matlab[38].

The construction elements were mostly printed in 3D, and the structure resembling a bicycle helmet was adapted to the preschool children's characteristics and limitations. The MDAD was made more subject-friendly with additional elements that were visually attractive to the child, e.g., artificial rabbit ears or a plume (Fig. 3b).

Before each recording session, the device was safely and comfortably placed on the speaker's head and eventually repositioned by the operator in its mobile part to adjust the distance from the sound source to the sensors. We prepared a dedicated adjustment interface to secure as much repeatable interspeaker and intraspeaker data acquisition as possible. Despite sufficient mobility, the MDAD is mechanically stable regarding the sound source and the scene during measurements. We used two versions of the MDAD to record the data: ver. 1 (closed construction, Fig. 3a) in first 53 speakers, and ver. 2 (open construction, Fig. 3b) in the remaining 148.

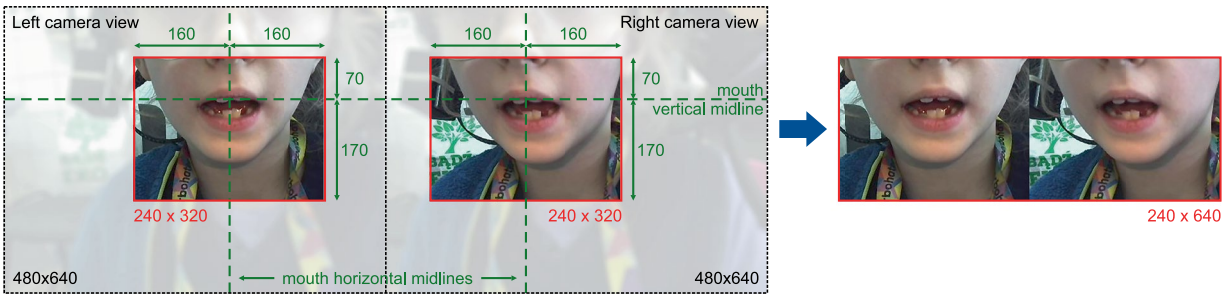


Fig. 4 Illustration of the video cropping procedure. The individuals in the pictures or their legal guardians consented to publishing their image in the manuscript.

Data preprocessing. The AV data shared in our dataset was prepared through a sequence of preprocessing operations. First, the 15-channel audio recording and dual-camera video stream were synchronized in time with both video frames concatenated horizontally (Fig. 3d). Then, we manually segmented audio data into segments: words, logotomes, and phones (employing inventory of 37 phonemes[39]). The segmentation was prepared in Audacity[40] based on the time series and spectrogram representations of the central microphone signal. No normalization or other audio data processing took place. Based on the audio segmentation results, we trimmed the AV stream in time in each segment to adjust the number of video frames and audio samples based on both sampling rates.

Finally, we applied additional cropping to the video frames to limit the scene size for improved anonymity (Fig. 4). The procedure involved the determination of the speaker- and recording-wise mouth vertical and horizontal midlines and cropping with fixed margins to show the limited oral area. The midlines were obtained for each frame using a YOLO v6 object detector trained to recognize the lips region of interest (ROI)[41,42]. Then, we determined midline valid for a particular participant and examination/recording part as a median across all related frames. Finally, all frames were cropped with fixed limits shown in Fig. 4. Therefore, the output frame size is 240×640.

We published the AV data of each segment in two forms. First, the uncompressed 15-channel audio stream was stored in a WAV format. Second, we prepared a video stream in an MP4 format, using an H.264/MPEG-4 AVC encoder[43] with a high-quality constant rate factor (CRF) of 18. The latter representation is treated as video data, but we added a synchronized audio track for easier database browsing. The single-channel audio comes from the central microphone (channel #8) compressed using the advanced audio coding (AAC) standard with a 192k bitrate.

Speech examination questionnaire. Each participant's articulation was examined in detail under a dedicated diagnostic questionnaire prepared by our SLT team. The description addressed the child's speech production (especially sibilants), and anatomical and physiological issues (regarding the tongue frenulum, upper lip, palate, and teeth; e.g., swallowing, breathing, and tongue mobility). The experts also assessed the production of individual sibilants. The questionnaire consisted of 196 fields (including general data on the examination and descriptive elements), all filled in Polish. For this study, we combined the original items and obtained a concise subset of 95 fields, mostly of categorical type (Table 4). Note that the definitions of articulatory features are provided with the dataset in the PDF file (see section **Dataset files**).

The goal was to prepare two independent diagnoses by two SLPs. One was done by the expert attending the examination simultaneously with the recording session (see section **Speech and language therapy examination with a data recording session**). The other was prepared by the second expert the other day without data recording. We collected double diagnoses in 181 out of 201 participants and single diagnoses in the remaining 20 cases. The dataset contains 185 diagnoses from expert E1 and 197 from expert E2.

Data Records

Database structure. The PAVSig dataset is available under the following DOI[44]: https://doi.org/10.7910/DVN/IHZRGB. The structure of the main folder of the PAVSig repository is shown in Fig. 5a. There is a separate folder with the audio, video, and speech diagnosis data from each participant, named `00XXX` (`XXX` stands for the anonymized three-digit ID of a participant; the extreme folder names are `00030` and `00237`), five CSV files with respective dataset specifications, and a PDF file presenting the diagnosis dictionary.

All CSV files with dataset summaries use semicolon as a delimiter and are encoded using a UTF-8 standard. Two types of special characters must be imported carefully: Polish letters with diacritics and IPA (International Phonetic Alphabet) symbols. Although the dataset could bring the most valuable contribution to Polish speech research, we also paid attention to presenting all resources in English.

Participant folder. A complete participant folder contains two audio data subfolders, two video data subfolders, and a CSV file with a double-expert speech diagnosis (Fig. 5b). In the case of a missing recording of one of two parts, there are only single audio and video subfolders. The subfolder naming rule is: `00XXX-R-audio` or `00XXX-R-video`, where `R` = `1` or `2` refers to the first or second part of the recording session. The participant diagnosis from one or two SLPs is stored in a CSV file named `00XXX-diag.csv`.

Group	#	Description
General fields	3	Participant and expert IDs, possible respiratory infection.
Anatomical and functional assessment of articulators	20	Examination outcome of tongue frenulum, superior and inferior labial frenulum, occlusion, palate, breathing, swallowing, and tongue posture. Description of articulatory movements in five vowels and six tongue exercises.
Assessment of articulation of specific sounds	6 × 12	Six articulatory features determined for each sibilant: active articulator, place of articulation, voicing, palatality, nasality, manner of articulation.
Total	95	

Table 4. General specification of the speech examination questionnaire items.

Audio data. Each audio data folder contains a complete set of WAV files with audio segments extracted from the recording of the corresponding part of the session (Fig. 5c). The segments contain either words or phones within words. Each WAV file stores an uncompressed 15-channel audio stream recorded at 16 bits and 44.1 kHz in a setup described in the **Methods** section (the order of channels in the WAV file corresponds to the arrangement shown in Fig. 3c).

The file nomenclature protocol is as follows:

- The file name for a word or logotome is `<word>.wav` (`word` is a Polish word with removed diacritics and spaces).
- The file name for a phoneme `p` within a word `parentWord` is `<parentWord_p>.wav`. There are some special cases here:

– To stay within Latin alphabet, sibilants are written as they sound in Polish. That makes `p` to be one-, two, or even three-letter patterns (e.g., see `ci` denoting /tɕ/ in Fig. 5c or `zi`, `drz` instead of /ʑ/, /dʑ/, respectively, in Fig. 5d). For more details see Table 6, field **sibilant**.

– If a word contains more than one phoneme of a certain type, a counter value follows `p` in the second and possibly next occurrence (see `zaba_a.wav` and `zaba_a2.wav` in Fig. 5c).

– Some speakers produced the same word twice. In such cases, the second occurrence is indicated by adding `2` after the word name, e.g., `owoce2.wav` (note that the second "o" here is stored in a file named `owoce2_o2.wav`).

– Finally, there are some words produced in a different form, e.g., "dzwon" instead of "dzwonek" or "siatkówka" instead of "siatka". In such cases, `word` and `parentWord` in the filename is always a correct form consistent with Table 2, although all existing phonemes are stored (e.g., "siatkówka" produces a word segment `siatka.wav` and the following phoneme segments:

```
siatka_si.wav,
siatka_a.wav,
siatka_t.wav,
siatka_k.wav,
siatka_u.wav,
siatka_f.wav,
siatka_k2.wav,
siatka_a2.wav).
```

The word change is indicated in the dataset through a mechanism described in the **Data Validation** section.

Moreover, to provide data that is easily applicable to standard speech analysis software (Praat[45]), we have added a separate text file to each word containing segmentation and annotation data in a TextGrid format[46]. The file naming rule is as follows: annotations of the word/logotome in the audio file `<word>.wav` are included in the text file `<word>.txt`. In this case, the phoneme labels are transcribed using IPA.

Video data. The video data folder has the same number of MP4 files as the audio folder with the same nomenclature rules and identical file names (Fig. 5d). Individual file stores a single audiovisual segment (word or phoneme) of the recording: a dual-camera view cropped to a 240 × 640 size, as described in the **Methods** section, with a single-channel audio signal from the central microphone #8.

Participant diagnosis. The participant diagnosis file `00XXX-diag.csv` includes six columns: three with the diagnosis in Polish and three with the English translation. In either triplet, the first column contains the questionnaire field name (e.g., the name of articulation or phonetic cue), and the next two store the responses from SLPs E1 and E2. If one of the experts did not examine the child, the corresponding column is empty. Note that the content of the participant diagnosis CSV file is a portion of data extracted from the complete diagnosis dataset described in the **Diagnosis summary** section.

Dataset files. *Participant summary.* The participant dataset `participantSummary.csv` gathers the anonymized data of the children participating in the study. The dataset fields (columns) are specified in Table 5. The **articulation** field is an attempt to assess each participant's articulation with a single, simplified label. For each SLT assessment, we took the three most significant features per sibilant (place and manner of articulation,

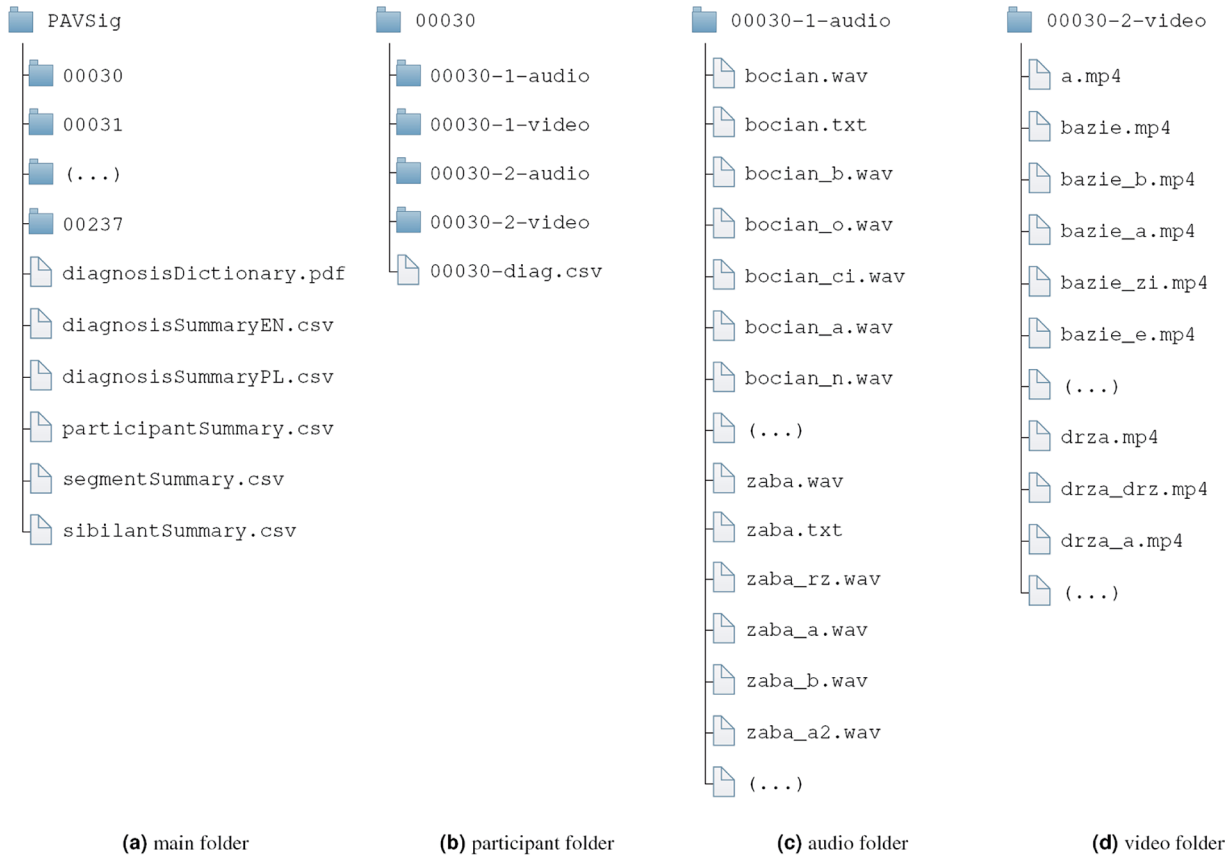


Fig. 5 Illustration of the data repository structure at different levels of the folder tree: (**a**) main folder, (**b**) participant folder, (**c**) audio folder, (**d**) video folder.

Variable	Description	Format, options
participant	Participant ID.	Integer; ranges from 30 to 237.
sex	Participant sex.	A single-character code: `F` – female, `M` – male.
age	Participant age at the time of the examination.	A char string "`Yy Mm`", where `Y`, `M` – years and months, respectively.
unit	Randomized ID of the preschool unit.	Integer; ranges from 1 to 6.
deviceVer	Recording device (MDAD) version ID.	Integer: 1 – ver. 1, closed construction (Fig. 3a), 2 – ver. 2, open construction (Fig. 3b).
participantFolderName	Participant folder name in the repository (Fig. 5).	A five-character code `00XXX`, where `XXX` is a three-digit representation of the **participant** field.
recording_1 recording_2	Presence/completeness of the recording of part 1 or 2 of the examination	Numerical: 1 (complete), 0.5 (incomplete—refers to missing audio data in channels 1–5 from the top recording arc), or 0 (missing).
nSegments_1 nSegments_2	Total number of AV segments in the recording of part 1 or 2 of the examination.	Integer or empty if there is no recording.
nWords_1 nWords_2	Total number of AV word segments in the recording of part 1 or 2 of the examination.	Integer or empty if there is no recording.
nPhones_1nPhones_2	Total number of AV phoneme segments in the recording of part 1 or 2 of the examination.	Integer or empty if there is no recording.
articulation	Initial overall classification of the participant's articulation of sibilants.	A char string: `typical` or `atypical`.

Table 5. Data dictionary for the participant dataset.

voicing). A single SLT diagnosis yielded the `typical` label if all features were assessed as typical, and `atypical` otherwise. Overall, the participant's articulation was classified as `typical` if none of SLT assessments detected distorted pronunciation. However, this field should be treated with caution, as articulation diagnosis is complex, and some alterations from the target norm are natural in the progress of articulation development. We recommend that dataset users revise the individual features and select the categorization scheme based on their research needs.

Variable	Description	Format, options
word	Parent word containing the sibilant.	Char; Polish word or logotome with removed diacritics. The only exception from this rule is the logotome "drza", which is originally "dża" (/d͡ʒa/) in Polish, but would be undistinguishable without diacritic from "dza" (/d͡za/).
sibilant	Sibilant Latin code.	A one-, two- or three-character code for the sibilant: s – /s/ (s), z – /z/ (z), c – /t͡s/ (c), dz – /d͡z/ (dz), si – /ɕ/ (ś), zi – /ʑ/ (ź), ci – /t͡ɕ/ (ć), dzi – /d͡ʑ/ (dź), sz – /ʂ/ (sz), rz – /ʐ/ (ż), cz – /t͡ʂ/ (cz), drz – /d͡ʐ/ (dż).
stress	Placement of stress on the syllable containing the sibilant.	Boolean: 1 – stress, 0 – no stress.
wordPosition	Sibilant position within the word.	Integer: 1 – initial, 2 – medial, 3 – final position.
precedingPhone	Phoneme before the sibilant.	Char; "–" if the sibilant opens the word.
followingPhone	Phoneme after the sibilant.	Char; "–" if the sibilant closes the word.
syllableCount	Number of word syllables.	Integer.
wordEN	English translation of the word.	Char.
wordPL	Polish form of the word with diacritics.	Char, UTF-8 encoding.
wordIPA	IPA transcription of the word.	Char, UTF-8 encoding.
sibilantIPA	IPA transcription of the sibilant.	Char, UTF-8 encoding.

Table 6. Data dictionary for the sibilant dataset.

Sibilant summary. The sibilant dataset `sibilantSummary.csv` specifies all sibilants in the speech corpus. The data dictionary is given in Table 6 (also refer to Table 2).

Segment summary. The `segmentSummary.csv` describes all AV segments available in the dataset. That concerns words, logotomes, and phonemes. Each entry in the segment summary has a corresponding WAV file in the audio subfolder and an MP4 file in the video subfolder. Table 7 presents the segment data dictionary.

Diagnosis summary. Two files store the complete set of SLT annotations: `diagnosisSummaryPL.csv` with the original diagnoses in Polish and `diagnosisSummaryEN.csv` with the English translation. In either case, 95 SLT examination questionnaire fields, e.g., the participant or expert ID or the name of articulation cue) are organized in columns with a field name in the first row. The entries (rows) are sorted in ascending order of participant ID and then expert ID. The participant diagnosis placed in the participant folder comes from extracting the appropriate subarrays (the field names and participant-related rows) from Polish and English datasets, transposing them to the 95×3 tables, and concatenating horizontally.

The `diagnosisDictionary.pdf` file contains the definitions of articulation-related fields of the speech examination questionnaire.

Technical Validation

Equipment validation. A thorough technical validation of the audio data acquisition component of the MDAD was performed and presented in detail in[34]. Below, we briefly report the experiments and results. Since then, we have redesigned the MDAD regarding its construction and usability. However, we have used the same microphones and their arrangement for audio recording and hardware for processing.

1. In the first experiment, we tested the MDAD using synthetic signals in accordance with the Polish standard PN-EN ISO 3746[47] specifying acoustic measurements of sound level (*SL*) in conditions close to the free field. The experiments were performed in a special acoustically adapted room where the noise rating (*NR*) was acceptable for recording studios in the NR 25–30 range. Based on measuring the *SL* and a signal-to-noise ratio (*SNR*), we found out that all 15 microphones record the signal in the same way in different tone frequencies between 1 and 8 kHz. Depending on the tone, the mean *SNR* was between 61.3 and 65.7 dB, safely acceptable for medium-class recording equipment[48,49].
2. In the second experiment, we verified the MDAD's ability to detect abnormal air outflow during articulation[50,51]. For this purpose, a human speaker simulated various air blows (central, left, and right outflow, each repeated three times) in ten attempts. The energy distributions indicated the appropriate reactions of the sensors to the directional acoustic stimuli.

The dual-camera video recording system has been added to the MDAD and shown originally in[35]. In that study, we presented the ability to support repeatable interspeaker and intraspeaker data acquisition by adjusting the mask position on a subject's head through a dedicated visualization interface. We superimposed reference lines on the camera images to help the operator reliably place the MDAD on the speaker's head. We use them to align possible stereovision viewpoints with the characteristic points of the face, e.g., the philtrum. We also estimated the extrinsic and intrinsic geometric parameters of the stereo system for eventual calibration purposes by finding the geometrical relationship between two cameras by observing the same point[52,53]. We used a chessboard template with known dimensions and geometry to calibrate individual cameras separately. Then,

Variable	Description	Format, options
participant	Participant ID.	Integer; ranges from 30 to 237.
participantFolderName	Participant folder name in the repository (Fig. 5).	A five-character code `00XXX`, where `XXX` is a three-digit representation of the **participant** field.
recordingType	Number of recording (examination) part the segment comes from.	Integer, 1 or 2.
segment	Segment name/code.	Char. Polish word, logotome, or phoneme with removed diactitics. Words and sibilants encoded according to the **word** and **sibilant** format in Table 6. Other special cases of phonemes: `bi` – /bʲ/, `ki` – /ć/, `l` – /w/, `pi` – /pʲ/.
segType	Type of segment.	A single-character code: `w` – word/logotome, `p` – phoneme.
parentWord	Parent word containing the phoneme.	Char; Polish word or logotome with removed diacritics. Empty if the segment is a word or logotome.
dataValidity	Segment data validity level	Numerical, range 0–1. Dictionary (see **Data Validation** section): 1.0 – correct segment, 0.9 – minor issues, e.g., slightly different word version or declination, 0.5 – major issues, e.g., a speech error changing sibilant environment.
videoFramesNr	Number of video frames within the segment.	Integer.
videoSegmentFileName	Name of the segment video file (MP4).	Char.
videoSegmentPath	Relative path to the segment video file (MP4) within the repository.	Char.
audioSamplesNr	Number of audio samples within the segment.	Integer.
audioSegmentFileName	Name of the segment audio file (WAV).	Char.
audioSegmentPath	Relative path to the segment audio file (WAV) within the repository.	Char.
textGridSegmentFileName	Name of the segment TextGrid file. Valid in words only.	Char.
textGridSegmentPath	Relative path to the segment TextGrid file within the repository. Valid in words only.	Char.

Table 7. Data dictionary for the segment dataset.

we determined translation and rotation matrices between the cameras and yielded a mean calibration error of 0.39 pixels[54].

Data completeness remarks. During PAVSig collection, we faced some issues with the data completeness, leading to the exclusion of participants or lack of data, some of which have already been mentioned in the paper. This section provides an explanation of these categories.

Unsuitable participants or missing recording session. Of 208 participants under consideration, four were excluded due to the following reasons:

- Two speakers were of non-Polish origin (Ukrainian) and thus non-native Polish speakers.
- One child's recordings and examination were unreliable and disrupted by the presence of drains in their ears.
- One child was examined by an SLP but was later unavailable to participate in the examination with the recording session. Thus, they were excluded from the study with nothing more than a single diagnostic questionnaire.

Data acquisition failures. In rare cases, we experienced some technical problems with the data acquisition. That concerned both audio (unacceptable noise, missing samples) and video (missing frames, synchronization issues). That led to the exclusion of three participants, producing the final research sample size of 201. Moreover, there are four speakers with one part of the recording unavailable beacause of that (three in part 1 and one in part 2).

Incomplete audio data. In 21 speakers, especially in the early stage of the study involving the first version of the device, there was an issue with a part of the multichannel audio stream. Due to technical reasons, the signal from microphones #1–5 (top recording arc) was damaged and unavailable. The remaining ten channels (#6–15, including the central channel #8) are complete. We indicated these cases in the **recording_1, recording_2** fields in the participant summary (Table 5).

Missing diagnoses. Due to organizational reasons, we were not always able to perform the second SLT examination. The total number of missing diagnoses is 20, but each child has at least one questionnaire (see section **Speech examination questionnaire**).

Data validation. We performed an extensive review and validation of the recorded and processed data. The validation covered a manual assessment of ca. 13k segments containing words and logotomes after applying all interventions to the research sample and participant-related constituents, as described in the **Data completeness remarks** section.

The expected number of word/logotome segments was 13,668: 201 participants $\times$ (38 words in part 1 + 30 words and logotomes in part 2). With three speakers with missing part 1 and one with missing part 2 recordings, this number was limited by 144 ($3 \times 38 + 1 \times 30$) to 13,524. A portion of segments (718, 3.5 per speaker) was missing due to several reasons:

Segment	#	Data validity level			Segment	#	Data validity level			Segment	#	Data validity level		
		1.0	0.9	0.5			1.0	0.9	0.5			1.0	0.9	0.5
Words (part 1)					**Words (part 2)**					**Sibilants**				
bocian	190	190	0	0	bazie	199	189	5	5	c	1,179	1,163	11	5
cebula	197	196	1	0	biegacz	199	193	5	1	ci	732	705	23	4
ciastka	146	133	13	0	dzokej	197	191	5	1	cz	784	761	18	5
czapka	196	196	0	0	kasza	198	192	5	1	drz	589	570	15	4
dziadek	191	191	0	0	lodzie	190	178	5	7	dz	592	538	49	5
dzwonek	199	157	42	0	lokiec	198	190	6	2	dzi	577	558	10	9
jeze	178	123	1	54	pajac	197	193	3	1	rz	1,639	1,562	9	68
kaczka	188	185	0	3	paz	199	191	6	2	s	1,642	1,563	39	40
kalosze	190	189	0	1	radza	198	192	5	1	si	951	926	16	9
koszyk	195	191	0	4	sadzawka	196	189	3	4	sz	2,364	2,311	41	12
koza	178	178	0	0	taca	193	188	3	2	z	934	922	10	2
ksiazka	190	190	0	0	w pasie	198	189	5	4	zi	593	568	16	9
kucharz	192	190	2	0	ziarno	198	190	6	2	**Other phonemes**				
las	180	179	0	1	**Sibilant logotomes (part 2)**					a	13,643	13,325	239	79
lekarz	142	140	2	0	ca	199	194	4	1	b	965	951	9	5
mazaki	167	167	0	0	cia	197	192	4	1	bi	199	193	5	1
noz	196	195	0	1	cza	193	187	5	1	d	388	388	0	0
owoce	196	195	0	1	drza	194	187	5	2	e	3,795	3,639	55	101
parasol	196	177	19	0	dza	198	193	4	1	f	1,066	1,047	12	7
pies	182	160	0	22	dzia	196	189	5	2	g	398	385	12	1
roza	171	171	0	0	rza	193	185	5	3	h	192	190	2	0
rzeka	174	174	0	0	sa	198	194	3	1	i	1,128	1,119	7	2
salata	161	160	1	0	sia	197	191	4	2	j	572	507	9	56
samolot	196	196	0	0	sza	194	188	4	2	k	4,646	4,565	50	31
siatka	167	165	1	1	za	196	192	3	1	ki	198	190	6	2
strazak	186	175	0	11	zia	196	189	5	2	l	1,635	1,598	34	3
szafa	191	188	3	0	**Vowels (part 2)**					l’	934	912	13	9
szalik	193	180	12	1	a	199	196	3	0	m	366	364	2	0
sznurek	142	131	11	0	i	195	192	3	0	n	931	864	63	4
szufelka	145	144	1	0	ia	196	193	3	0	o	3,103	2,997	86	20
warzywa	186	186	0	0	iu	195	193	2	0	p	993	948	37	8
waz	195	194	1	0	u	199	196	3	0	pi	182	160	0	22
widelec	197	197	0	0						r	1,655	1,593	48	14
zaba	189	186	3	0						t	1,049	1,018	18	13
zabawki	190	190	0	0						u	1,411	1,390	18	3
zarowka	171	171	0	0						w	1,345	1,301	43	1
zegar	199	192	7	0						y	581	576	1	4
zyrafa	193	193	0	0										
Total	6,935	6,715	120	100	Total	5,895	5,716	127	52	Total	53,951	52,367	1,026	558
		96,8%	1,7%	1,4%			97,0%	2,2%	0,9%			97,1%	1,9%	1,0%
					Total	12,830	12,431	247	152	**Total (dataset)**	**66,781**	64,798	1,273	710
					(words&logotomes)		96,9%	1,9%	1,2%			97,0%	1,9%	1,1%

Table 8. Segments distribution in the dataset. The left section contains words presented on the screen (part 1 of the examination), the central section covers words and logotomes repeated by the child following the SLP (part 2), and the right section lists phonemes (sibilants shown separately). Column “#” displays the total number of segment occurences in the dataset. Additional three columns show the distribution of data validity levels in each case. The bottom middle section combines the word/logotome statistics. The bottom right section summarizes the entire collection of segments.

- the children did not produce them at all, mostly in part 1, based on naming the pictures shown on the screen;
- the children used a synonym to name the picture (e.g., “mazaki”—"pisaki”, “lekarz”—"pan doktor”, or “sznurek”—"lina”);
- speech was severely disturbed by noise or other sounds.

On the other hand, in 24 cases, the child produced the same word or logotome twice. Hence, the total number of word/logotome segments we share is 12,830 (13, 524 − 718 + 24).

To make the data use easy and reliable, we performed an extended validation of the available set of words and logotomes, primarily for the sibilant analysis. We assigned three different data validity levels (DVL) to each segment:

- DVL = 1.0 – the segment is considered correct and suitable for the analysis.
- DVL = 0.9 – the segment presents a slightly different word version or declination than required. However, the word change does not affect the sibilant and its environment, so it is suitable for the analysis. Examples: "dzwonek"—"dzwon", "żaba"—"żabka", "ciastka"—"ciasto", "parasol"—"parasolka". Note that the speech error may affect the number of word syllables or stress.
- DVL = 0.5 – the segment presents a significantly different word version or declination that affects the sibilant environment. Examples: "pies"—"piesek", "jeże"—"jeż", "strażak"—"straż", "kaczka"—"kaczuszka". All words and logotomes modified in any other way also fall into this category.

Table 8 presents detailed statistics of our dataset regarding the total number of specific segments—words, logotomes, and phonemes—also with the DVL distributions. Note that the phoneme DVL is always inherited after its parent word or logotome. There are ca. 2% and 1% segments with minor and major issues, respectively, so 97% of segments are considered correct. The DVLs are stored in the segment dataset under the **dataValidity** field (Table 7).

There is a total of 66,781 segments in PAVSig (12,830 words and logotomes, 53,951 phonemes). The number of sibilant occurrences varies between 593 in "zi" (3.0 per speaker) and 2,364 in "sz" (11.8 per speaker). The total number of sibilants is 12,576 (62.6 per speaker).

Usage Notes

The dataset is available under the Data Use Agreement (DUA) with data access requirements given in the repository[44].

Code availability

No custom code was used to generate or process the data described in the manuscript.

Received: 1 July 2024; Accepted: 29 August 2025;
Published online: 02 October 2025

References

1. Demenko, G., Wagner, A. & Cylwik, N. The use of speech technology in foreign language pronunciation training. *Archives of Acoustics* **35**, 309–330 (2010).
2. Minczakiewicz, E. 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* **1**, 149–169 (2017).
3. Gacka, E. & Kaźmierczak, M. Speech screening examinations as an example of activity in the field of speech-language therapy. *Logopaedica Lodziensia* **1**, 31—42, https://doi.org/10.18778/2544-7238.01.04 (2017).
4. Smit, A. B. Phonologic error distributions in the Iowa-Nebraska articulation norms project. *Journal of Speech, Language, and Hearing Research* **36**, 533–547, https://doi.org/10.1044/jshr.3603.533 (1993).
5. Lockenvitz, S., Tetnowski, J. A. & Oxley, J. The sociolinguistics of lisping: a review. *Clinical Linguistics & Phonetics* **34**, 1169–1184, https://doi.org/10.1080/02699206.2020.1788167 (2020).
6. Amr Rey, O., Sánchez Delgado, P., Salvador Palmer, R., Ortiz De Anda, M. C. & Paredes Gallardo, V. Exploratory study on the prevalence of speech sound disorders in a group of Valencian school students belonging to 3rd grade of infant school and 1st grade of primary school. *Psicologia educativa* **28**, 195–207, https://doi.org/10.5093/PSED2022A1 (2022).
7. Grigorova, E., Ristovska, G. & Jordanova, N. P. Prevalence of phonological articulation disorders in preschool children in the city of Skopje. *PRILOZI* **41**, 31–37, https://doi.org/10.2478/prilozi-2020-0043 (2020).
8. Van Borsel, J., Van Rentergem, S. & Verhaeghe, L. The prevalence of lisping in young adults. *Journal of Communication Disorders* **40**, 493–502, https://doi.org/10.1016/j.jcomdis.2006.12.001 (2007).
9. Harrison, J. E., Weber, S., Jakob, R. & Chute, C. G. ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making***21**, https://doi.org/10.1186/s12911-021-01534-6 (2021).
10. Benselam, Z., Guerti, M. & Bencherif, M. Arabic Speech Pathology Therapy Computer-Aided System. *Journal of Computer Science* **3**, 685–692 (2007).
11. Valentini-Botinhao, C. *et al*. Automatic detection of sigmatism in children. In *Proc. WOCCI 2012 - Workshop on Child, Computer and Interaction*, 1–4 (Portland, USA, 2012).
12. Anjos, I. *et al*. A model for sibilant distortion detection in children. In *DMIP '18* (2018).
13. Żygis, M. & Padgett, J. A perceptual study of polish fricatives, and its implications for historical sound change. *Journal of Phonetics* **38**, 207–226, https://doi.org/10.1016/j.wocn.2009.10.003 (2010).
14. Bickford, A. C. & Floyd, R.*Articulatory Phonetics: Tools for Analyzing the World's Languages* (SIL International Dallas, 2006).
15. Zsiga, E. C.*The Sounds of Language: An Introduction to Phonetics and Phonology* (Wiley-Blackwell, 2012).
16. Miodonska, Z., Badura, P. & Mocko, N. Noise-based acoustic features of Polish retroflex fricatives in children with normal pronunciation and speech disorder. *Journal of Phonetics* **92**, 101149, https://doi.org/10.1016/j.wocn.2022.101149 (2022).
17. Krecichwost, M., Mocko, N. & Badura, P. Automated detection of sigmatism using deep learning applied to multichannel speech signal. *Biomedical Signal Processing and Control* **68**, 102612, https://doi.org/10.1016/j.bspc.2021.102612 (2021).
18. Katz, W., Mehta, S., Wood, M. & Wang, J. Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation. *The Journal of the Acoustical Society of America* **141**, 57–63, https://doi.org/10.1121/1.4973907 (2017).
19. Kroos, C. Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). *Journal of Phonetics* **40**, 453–465, https://doi.org/10.1016/j.wocn.2012.03.002 (2012).
20. Wood, S., Wishart, J., Hardcastle, W., Cleland, J. & Timmins, C. 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. *Developmental Neurorehabilitation* **12**, 66–75, https://doi.org/10.1080/17518420902738193 (2009).

21. Cleland, J., Timmins, C., Wood, S., Hardcastle, W. & Wishart, J. Electropalatographic therapy for children and young people with Down's syndrome. *Clinical Linguistics & Phonetics* **23**, 926–939, https://doi.org/10.3109/02699200903061776 (2009).
22. Bilibajkić, R., Vojnović, M. & Šarić, Z. Detection of lateral sigmatism using support vector machine. *Speech and Language 2019* 322–328 (2019).
23. Król, D., Lorenc, A. & Święciński, R. Detecting Laterality and Nasality in Speech with the use of a Multi-channel Recorder. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'15, 5147–5151, https://doi.org/10.1109/ICASSP.2015.7178952 (2015).
24. Lorenc, A., Król, D. & Klessa, K. An acoustic camera approach to studying nasality in speech: The case of polish nasalized vowels. *The Journal of the Acoustical Society of America* **144**, 3603–3617, https://doi.org/10.1121/1.5084038 (2018).
25. Wei, S., Hu, G., Hu, Y. & Wang, R.-H. A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication* **51**, 896–905, https://doi.org/10.1016/j.specom.2009.03.004 (2009).
26. Bílková, Z. *et al.* Automatic evaluation of speech therapy exercises based on image data. In *Lecture Notes in Computer Science*, 397–404, https://doi.org/10.1007/978-3-030-27202-9_36 (Springer International Publishing, 2019).
27. Bilkova, Z. *et al.* ASSISLT: Computer-aided speech therapy tool. In *2022 30th European Signal Processing Conference (EUSIPCO)*, 598–602, https://doi.org/10.23919/eusipco55093.2022.9909627 (IEEE, 2022).
28. Sage, A. *et al.* Deep learning approach to automated segmentation of tongue in camera images for computer-aided speech diagnosis. In *Advances in Intelligent Systems and Computing*, 41–51, https://doi.org/10.1007/978-3-030-49666-1_4 (Springer International Publishing, 2020).
29. Chotikkakamthorn, K., Ritthipravat, P., Kusakunniran, W., Tuakta, P. & Benjapornlert, P. A lightweight deep learning approach to mouth segmentation in color images. *Applied Computing and Informatics*https://doi.org/10.1108/aci-08-2022-0225 (2022).
30. Miled, M., Messaoud, M. A. B. & Bouzid, A. Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications* **82**, 551–571, https://doi.org/10.1007/s11042-022-13321-0 (2022).
31. Milewski, S. Phoneme frequency in preschool children's spoken texts, (PL) Frekwencja fonemów w tekstach mówionych dzieci w wieku przedszkolnym. *Logopedia* **24**, 67–83 (1997).
32. Milewski, S. & Binkuńska, E. The phonostatistic and phonotactic structure of selected Polish difficult word sequences (tongue twisters), (PL) Struktura fonostatystyczno-fonotaktyczna wybranych polskich słownych ciagów trudnych (lingwołamek). *Prace Jezykoznawcze* **26**, 143–160, https://doi.org/10.31648/pj.10593 (2024).
33. Krajna, E. & Bryndal, M. 100-Word Articulation Test: Auditory Analysis of Recordings and Attempt at Test Normalization, (PL) 100-wyrazowy test artykulacyjny. Analiza słuchowa nagrań i próba normalizacji testu. *Audiofonologia* **XIV**, 137–174 (1999).
34. Krecichwost, M., Miodonska, Z., Trzaskalik, J. & Badura, P. Multichannel speech acquisition and analysis for computer-aided sigmatism diagnosis in children. *IEEE Access* **8**, 98647–98658, https://doi.org/10.1109/ACCESS.2020.2996413 (2020).
35. Krecichwost, M., Sage, A., Miodonska, Z. & Badura, P. 4D multimodal speaker model for remote speech diagnosis. *IEEE Access* **10**, 93187–93202, https://doi.org/10.1109/ACCESS.2022.3203572 (2022).
36. Panasonic. Omnidirectional Back Electret Condenser Microphone Cartridge, Series: WM-61A, WM-61B. https://www.alldatasheet.com/datasheet-pdf/pdf/528408/PANASONIC/WM-61A.html [accessed: 17-May-2024].
37. ArduCam. Arducam 8MP 1080P Auto Focus USB Camera Module with Microphone [accessed 20-March-2023].
38. The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). https://www.mathworks.com (2022).
39. Jassem, W. Polish. *Journal of the International Phonetic Association* **33**, 103–107, https://doi.org/10.1017/S0025100303001191 (2003).
40. Audacity Team. Audacity (R): Free Audio Editor and Recorder. https://www.audacityteam.org/ (2023).
41. Sage, A. & Badura, P. Detection and segmentation of mouth region in stereo stream using yolov6 and deeplab v3+ models for computer-aided speech diagnosis in children. *Applied Sciences***14**, https://doi.org/10.3390/app14167146 (2024).
42. Li, C. *et al.* YOLOv6: A single-stage object detection framework for industrial applications. *arXiv:2209.02976,* https://doi.org/10.48550/ARXIV.2209.02976 (2022).
43. International Telecommunication Union Telecommunication Standardization Sector (ITU-T). H.264 : Advanced video coding for generic audiovisual services (2003).
44. Krecichwost, M. *et al.* PAVSig: Polish multichannel Audio-Visual child speech dataset with double-expert Sigmatism diagnosis, https://doi.org/10.7910/DVN/IHZRGB (2025).
45. Boersma, P. & Weenink, D. Praat: doing phonetics by computer. https://www.fon.hum.uva.nl/praat/. [accessed: 3-April-2025]
46. Boersma, P. & Weenink, D. TextGrid file formats. https://www.fon.hum.uva.nl/praat/manual/TextGrid_file_formats.html [accessed: 3-April-2025].
47. Polish Committee for Standardization (PCS). 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 płaszczyznaą odbijającą dźwięk. Standard, PN-EN ISO 3746:2011, PCS (2017).
48. Rosen, S. & Howell, P.*Signals and systems for speech and hearing* (Brill, 2011).
49. Baken, R. & Orlikoff, R.*Clinical Measurement of Speech and Voice*. Speech Science (Singular Thomson Learning, 2000).
50. Klatt, D. H., Stevens, K. N. & Mead, J. Studies of articulatory activity and airflow during speech. *Annals of the New York Academy of Sciences* **155**, 42–55, https://doi.org/10.1111/j.1749-6632.1968.tb56748.x (1968).
51. Lorenc, A., Żygis, M., Łukasz, M., Pape, D. & Sóskuthy, M. Articulatory and acoustic variation in Polish palatalised retroflexes compared with plain ones. *Journal of Phonetics* **96**, 101181, https://doi.org/10.1016/j.wocn.2022.101181 (2023).
52. Zhang, Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 1330–1334, https://doi.org/10.1109/34.888718 (2000).
53. Heikkila, J. & Silven, O. A four-step camera calibration procedure with implicit image correction. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1106–1112, https://doi.org/10.1109/CVPR.1997.609468 (1997).
54. Bier, A. & Luchowski, L. Error analysis of stereo calibration and reconstruction. In Gagalowicz, A. & Philips, W. (eds.) *Computer Vision/Computer Graphics CollaborationTechniques*, 230–241, https://doi.org/10.1007/978-3-642-01811-4_21 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009).

Acknowledgements

This work was supported by the National Science Centre, Poland, research project No. 2018/30/E/ST7/00525: "Hybrid System for Acquisition and Processing of Multimodal Signal in the Analysis of Sigmatism in Children", partially by the Foundation for Polish Science (FNP), partially by the Polish Ministry of Science, Poland, statutory financial support No. 07/010/BK_25/1047, and partially by the European Funds for Silesia 2021-2027 Program co-financed by the Just Transition Fund Project "Development of the Silesian Biomedical Engineering Potential in the Face of the Challenges of the Digital and Green Economy (BioMeDiG)" under Grant FESL.10.25-IZ.01-07G5/23. The authors want to thank Natalia Moćko, PhD, Marcin Biesok, and Wojciech Galiński for their valuable help in audio data analysis.

Author contributions

All authors are justifiably credited with authorship. The detailed contribution is as follows: M.K., Z.M., J.T., and P.B. conceived and designed the analysis, M.K. designed, produced, and validated the equipment, M.K., Z.M., A.S., J.T., E.K., and P.B. collected the data, M.K., Z.M., A.S., and P.B. contributed data or analysis tools, M.K. and P.B. designed and prepared the database, J.T. and E.K. provided SLT assessments, M.K., Z.M., A.S., and P.B. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.