



Visual features of the oral region in Polish sibilants produced by children with various sibilance patterns

Agata Sage¹, Zuzanna Miodońska¹, Michał Kręcichwost¹, Ewa Kwaśniok², Paweł Badura¹

¹Faculty of Biomedical Engineering, Silesian University of Technology, Poland

²Medical Facility 'Therapy' – Center for Health and Human Development, Poland

agata.sage@polsl.pl

Abstract

This paper analyzes the 2D and 3D visual features of Polish sibilants /s/ and /ʂ/ articulated by 189 children aged 4;11-8;0. Our goal was to (1) analyze differences in shape image features of delineated articulators (lips, tongue, and mouth—lips and space between them) in selected sibilants and (2) investigate a relationship between visual parameters and sibilance pattern (the character of a gap between teeth that allows the articulation of sibilant sounds). We employed 2D and 3D shape image features and linear mixed-effect models. The results proved significant differences between sibilants in 2D and 3D visual features of lips and mouth. Different sibilance patterns considered in the study also significantly impacted visual parameters distribution. Our findings indicate the potential of video data in computer-aided speech diagnosis.

Index Terms: sibilants, child speech, speech disorders, image features

1. Introduction

Sibilants appear as one of the last groups of sounds in the articulation development process [1]. The tongue muscle coordination and the aerotactile acuity of the tongue required to articulate a sibilant are complex and need time to improve [2, 3]. One of the children's most common speech disorders is sibilant-related sigmatism (lisp). The percentage of people with sigmatism in the population varies between studies reported in the literature. However, in various parts of the world, including Poland, this problem is described as frequent [4, 5, 6].

The number of computer methods for articulation and speech analysis is increasing. Most works and open-access databases consider the pronunciation of adult speakers [7, 8], since obtaining child speech data is difficult. Studies addressing child pronunciation constitute a smaller but still growing group [9, 10, 11]. Although some of the literature solutions claim to be accurate and provide high-quality data, they frequently interfere with the oral cavity of the speaker (e.g., electromagnetic articulography [12] or electropalatography [13]). Recording devices that require such direct contact with the speaker constitute a barrier in the case of unrestrained child speech analysis. Thus, researchers began to employ contactless methods, such as recording acoustic signals (single or multichannel). Sibilants are often investigated regarding spectral properties [14], which are also related to frication noise [10, 15].

In addition to acoustic analysis, speech therapists often use visual investigation of the speaker's articulators. The positions of articulators during pronunciation largely shape the spectral envelope of the acoustic signal. Dysfunctions in speech organ movement are often associated with articulation problems. We expect that visual data should reflect speech organ positioning.

Therefore, research on using video or images for speech diagnosis and therapy can provide essential information for computer-based articulation analysis. However, the applicability of visual data of the oral region in this matter has been investigated to a small extent thus far. The works reported analysis of various visual data, including articulatory ultrasound images [16], magnetic resonance imaging [17], but also camera videos [18, 19]. Bilkova et al. [18] used camera data for speech therapy purposes with the analysis of articulators' appearance. Except for Bilkova's study, the area of image processing for the diagnosis of articulation, especially sibilant-related, needs to be explored. Sage et al. [20] proposed a hybridization of acoustic and 2D image features based on texture and shape to analyze the place of articulation in sounds /s/ and /ʂ/, where tongue's shape features proved to differentiate dental and interdental patterns in speakers significantly. Following these findings, we decided to conduct further experiments on the usefulness of visual data in non-contact articulation analysis.

This study addresses the issue of sibilance—a process of creating an additional gap (in addition to an airflow gap between active and passive articulators) that allows the production of sibilant sounds [21]—in sibilants articulation. The normal sibilance involves a close-up of upper and lower incisors, but researchers report different distortions. We include two of them in our analysis: (1) diastemic pattern of sibilance, where the additional gap is caused by diastema of the upper incisors, (2) vertically distorted sibilance that occurs when the gap is formed incorrectly due to the upper and lower incisors overlap excessively or are too far apart [22, 23].

In this paper, we report the analysis of 2D and 3D visual features of the oral region in Polish sibilants /s/ and /ʂ/ produced by 189 children aged 4;11-8;0. Our investigation was performed based on the data collected from speakers with three sibilance patterns: normal, diastemic, and distorted vertically. We considered two research objectives: (1) the analysis of differences in shape image features of articulators in selected sibilants and (2) the investigation of the possible impact of sibilance patterns on visual parameters. We employed a set of features proposed in some previous works [20] and added 3D image parameters built in time as a stack of following frames covering the articulation. The statistical inference was performed using linear mixed-effect models.

2. Materials

The data analyzed in this study was part of a multimodal audio-visual database of Polish child speech. Our team, composed of speech therapy experts and biomedical engineers, collected material in six kindergartens and school facilities. It was collected using a multimodal acquisition device that enables au-

Table 1: *Speech corpus used in the study. Words and logatomes containing /s/ in lines 1–9, while containing /ʃ/ in lines 10–22. Syl.—syllable count, Prec.—preceding phoneme, Pos.—word position (1—initial, 2—medial, 3—final).*

No.	Word	IPA	Translation	Stress	Syl.	Prec.	Pos.	No.	Word	IPA	Translation	Stress	Syl.	Prec.	Pos.
1	pies	/pʲɛs/	dog	+	1	e	3	12	szafa	/ʃafa/	wardrobe	+	2	-	1
2	strażak	/ˈstrazak/	firefighter	-	1	-	3	13	szufelka	/ʃuˈfelka/	dustpan	-	3	-	1
3	samolot	/saˈmɔlot/	airplane	-	3	-	1	14	nóż	/nuʃ/	knife	+	1	u	3
4	sałata	/saˈwata/	lettuce	-	3	-	1	15	wąż	/vɔwɔʃ/	snake	+	1	ŵ	3
5	parasol	/paˈrasɔl/	umbrella	-	3	a	2	16	książka	/kɕɔwɔʃka/	book	+	2	ŵ	2
6	las	/las/	forest	+	1	a	3	17	lekarz	/ˈlekaʃ/	physician	-	2	a	3
7	ciastka	/ˈtɕastka/	cookies	+	2	a	2	18	sznurek	/ʃnurek/	cord	+	2	-	1
8	sadzawka	/saˈdzafka/	pond	-	3	-	1	19	kucharz	/ˈkuxaʃ/	cook	-	2	a	3
9	sa	/sa/	—	+	1	-	1	20	szalik	/ʃalik/	scarf	+	2	-	1
10	koszyk	/ˈkɔʃɨk/	basket	-	2	o	2	21	kasza	/ˈkaʃa/	groats	-	2	a	2
11	kalosze	/kaˈlɔʃɛ/	rain boots	-	3	o	2	22	sza	/ʃa/	—	+	1	-	1

Table 2: *Summary of the dataset used in this study.*

Sibilance pattern	/s/			/ʃ/		
	Speakers		Words	Speakers		Words
	F	M		F	M	
Normal	66	53	861	78	59	1419
Diastemic	6	4	73	7	4	121
Vertically distorted	20	25	404	14	25	379
Sum			1338			1909

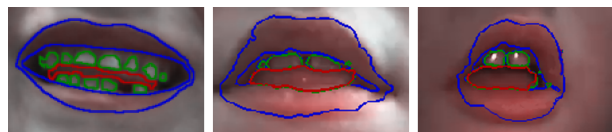


Figure 1: *Sample results of the segmentation method reported in [25]. The lips are marked in blue, the teeth in green, and the tongue in red.*

dio data recording with fifteen spatially distributed microphones and capturing a video stream [24]. The research involves only video recordings captured at 30 fps. The project was approved by the local Bioethical Committee.

The recording session for each speaker had three stages: (1) speech recording while naming pictures visible on the screen, (2) recording of a set of speech therapy exercises, and (3) a speech therapy examination performed by the speech and language pathologist (SLP) according to the dedicated diagnostic protocol for sigmatism-related articulation assessment. The language material involved over 50 words and phrases with all 12 Polish sibilants. However, in this study, we focused on two phonemes: /s/ and /ʃ/ (Table 1).

The part of the dataset used in this study included 189 children aged 4;11–8;0 with normal, diastemic, and vertically-distorted sibilance patterns. There were 100 female speakers (aged 6;4±0;9) and 89 male speakers (6;5±0;8). In some children, sibilance patterns differed between phonemes. Table 2 presents the dataset statistics.

3. Methods

3.1. Data preprocessing

The workflow began with extracting sibilants from words spoken by the children. An expert indicated the beginning and end of phonemes using spectrograms and waveforms of time-synchronized acoustic data. Then, we applied two-stage image segmentation to delineate lips, tongue, teeth, and mouth (defined in this study as the area of lips and space between them) in camera frames [25]. The method applies object detection using YOLOv6 to crop images to the mouth area bounding box, followed by segmentation with the DeepLabv3+ model (Fig. 1). We reviewed the results and rejected low-quality delineations to improve the credibility of further steps.

3.2. Visual features extraction

Although the delineations included mouth, lips, tongue, and teeth, we focused on the first three. We proposed a set of shape image features of two categories: 2D and 3D. The 2D approach considered the shape parameters of a given articulator in each frame capturing the sibilant articulation. Our 3D approach involved the volumes structured as stacks of subsequent frames. In this approach, the third dimension was related to the time of articulation, and our goal was to reflect changes in the articulators’ movement.

Eight 2D image features were calculated separately for the lips, mouth, and tongue [26, 27]: surface area (A^{2D}), perimeter (P^{2D}), sphericity (S^{2D}), spherical disproportion (SD^{2D}), major and minor axis lengths (Ax_{major}^{2D} , Ax_{minor}^{2D}), elongation (E^{2D}), and maximum Feret diameter (D_{Feret}^{2D}). The 3D set comprised 16 features extracted from each 3D volume individually [26, 27]: voxel volume (V^{3D}), surface area (A^{3D}), surface area to volume ratio (SVR^{3D}), sphericity (S^{3D}), spherical disproportion (SD^{3D}), compactness #1 (C_1^{3D}), compactness #2 (C_2^{3D}), major, minor, and least axis lengths (Ax_{major}^{3D} , Ax_{minor}^{3D} , Ax_{least}^{3D}), elongation (E^{3D}), maximum Feret diameter (D_{Feret}^{3D}), and maximum diameter in the XY, XZ, YZ planes (D_{XY}^{3D} , D_{XZ}^{3D} , D_{YZ}^{3D}). Thus, we determined 24 2D parameters for each frame and 48 3D features for each sibilant.

3.3. Statistical analysis

We used linear mixed-effect (LME) models to perform statistical analysis. Each sibilant was limited to the middle 50% of its frames, i.e., we rejected 25% from the beginning and end of the phoneme segment. We determined the models for each feature described in the previous section separately and performed several experiments with various model structures, including the following **fixed effects**:

- Phoneme: /s/, /ʃ/,
- Sibilance: normal, diastemic, vertically distorted.

Table 3: The summary of models with at least one statistically significant fixed effect for lips and mouth features (responses). Fixed effects are denoted as Ph:sz for phoneme /ʃ/, S:dias for diastemic sibilance pattern, and S:ver for vertically distorted sibilance. The intercept in our study is sibilant /s/ and normal sibilance. Significant findings are bolded.

		p	Est.	Std. Er.	T			p	Est.	Std. Er.	T
Lips											
E_{2D}	Int	<0.001	0.53	0.01	42.05	Ax_{major}^{3D}	Int	<0.001	3.51	0.01	241.34
	Ph:sz	<0.001	0.05	0.01	3.72		Ph:sz	0.20	-0.02	0.01	-1.30
	S:dias	0.40	-0.02	0.03	-0.84		S:dias	0.07	-0.08	0.04	-1.83
	S:ver	0.42	-0.01	0.01	-0.80		S:ver	0.04	-0.02	0.01	-2.03
E_{3D}	Int	<0.001	0.57	0.01	47.47	Ax_{minor}^{2D}	Int	<0.001	2.81	0.03	108.22
	Ph:sz	<0.001	0.04	0.01	3.04		Ph:sz	0.03	0.06	0.03	2.22
	S:dias	0.48	-0.02	0.03	-0.71		S:dias	0.01	-0.15	0.06	-2.49
	S:ver	0.97	0.00	0.01	-0.04		S:ver	0.02	-0.05	0.02	-2.44
F_{3D}	Int	<0.001	0.28	0.02	12.99	Ax_{minor}^{3D}	Int	<0.001	2.95	0.03	117.17
	Ph:sz	0.92	0.00	0.03	-0.10		Ph:sz	0.04	0.05	0.02	2.15
	S:dias	0.35	0.02	0.02	0.93		S:dias	0.02	-0.13	0.06	-2.28
	S:ver	0.60	0.00	0.01	0.53		S:ver	0.40	-0.02	0.02	-0.84
Ax_{major}^{2D}	Int	<0.001	3.44	0.01	237.73	V_{3D}	Int	<0.001	7.11	0.09	75.08
	Ph:sz	0.04	-0.03	0.01	-2.21		Ph:sz	0.55	0.07	0.11	0.61
	S:dias	0.05	-0.08	0.04	-1.95		S:dias	0.02	-0.28	0.12	-2.28
	S:ver	<0.001	-0.03	0.01	-3.33		S:ver	0.77	0.01	0.04	0.30
Mouth											
E_{2D}	Int	<0.001	0.50	0.01	40.82	Ax_{minor}^{3D}	Int	<0.001	2.83	0.02	128.12
	Ph:sz	<0.001	0.06	0.01	4.67		Ph:sz	<0.001	0.07	0.02	3.37
	S:dias	0.57	-0.02	0.03	-0.57		S:dias	0.03	-0.12	0.05	-2.23
	S:ver	0.76	0.00	0.01	-0.31		S:ver	0.94	0.00	0.02	-0.08
E_{3D}	Int	<0.001	0.52	0.01	46.13	P_{2D}	Int	<0.001	4.25	0.01	286.14
	Ph:sz	<0.001	0.06	0.01	4.89		Ph:sz	0.26	-0.01	0.01	-1.16
	S:dias	0.66	-0.01	0.03	-0.44		S:dias	0.01	-0.11	0.04	-2.50
	S:ver	0.55	0.01	0.01	0.59		S:ver	<0.001	-0.04	0.01	-4.78
Ax_{major}^{2D}	Int	<0.001	3.40	0.01	248.06	A_{2D}	Int	<0.001	6.43	0.03	199.93
	Ph:sz	<0.001	-0.04	0.01	-3.60		Ph:sz	0.31	0.03	0.03	1.03
	S:dias	0.06	-0.08	0.04	-1.92		S:dias	0.01	-0.24	0.09	-2.70
	S:ver	<0.001	-0.03	0.01	-3.17		S:ver	<0.001	-0.09	0.02	-3.97
Ax_{major}^{3D}	Int	<0.001	3.48	0.01	253.23	A_{3D}	Int	<0.001	6.94	0.05	137.89
	Ph:sz	<0.001	-0.04	0.01	-3.50		Ph:sz	0.86	0.01	0.06	0.18
	S:dias	0.07	-0.08	0.04	-1.85		S:dias	0.01	-0.21	0.08	-2.76
	S:ver	0.11	-0.01	0.01	-1.59		S:ver	0.22	-0.03	0.02	-1.23
Ax_{minor}^{2D}	Int	<0.001	2.72	0.02	116.29	V_{3D}	Int	<0.001	7.57	0.09	79.80
	Ph:sz	<0.001	0.07	0.02	3.09		Ph:sz	0.91	0.01	0.12	0.12
	S:dias	0.01	-0.14	0.06	-2.47		S:dias	0.01	-0.28	0.11	-2.55
	S:ver	0.06	-0.04	0.02	-1.91		S:ver	0.27	-0.04	0.04	-1.10

The random structure involved:

- Speaker as random intercept,
- by-speaker random slopes for Phoneme, WordPosition, Stress, PrecedingPhoneme, and syllablesCount,
- Word as random intercept,
- by-word random slopes for word Sibilance and Speaker.

We rejected models that did not meet the assumption of the normality of the residuals. The final model was selected based on the log-likelihood comparison and included Phoneme and Sibilance as fixed effects and the following random structure: (1 + Phoneme + WordPosition + Stress + SyllableCount + PrecedingPhoneme | Speaker) + (1 + Sibilance | Word). We performed the statistical analysis in R (version 4.2.4) with $p = 0.05$.

4. Results & discussion

As a result of LME modeling, we obtained a statistically significant effect for either Phoneme: [/s/, /ʃ/] or Sibilance: [normal, diastemic, vertically distorted] in seven features of the

lips and ten of the mouth (Table 3). We also analyzed features describing tongue shape, but LME models returned only statistically insignificant results.

We considered phonemes that differ in the place of articulation (dental in /s/ and retroflex /ʃ/) and vary in mouth positioning during articulation. We found differences between their realization, mainly in lips and mouth elongation and minor axis lengths (2D and 3D). The higher E^{3D} values in /ʃ/ for lips and mouth suggest the more rounded shape of the speech organs during articulation compared to Polish /s/(Fig. 2). A similar relation was observed in Ax_{minor}^{2D} . Its values were smaller for /s/ than /ʃ/, although major axis lengths were greater in the first. This also indicates that the articulation of the retroflexes is associated with a protruded, circular position of the lips, consistent with linguistic literature on this matter [28, 29].

The different sibilance patterns considered in this study also significantly impacted visual features' distribution (Fig. 3). We found them in both the lips and the mouth and both 2D and 3D extraction approaches. The lengths of the minor and major axes of the lips (Ax_{minor}^{2D} , Ax_{major}^{2D}) and the mouth perimeter (P^{2D}) significantly differentiated all three sibilance patterns (Fig. 3). The medians of all features are highest in the nor-

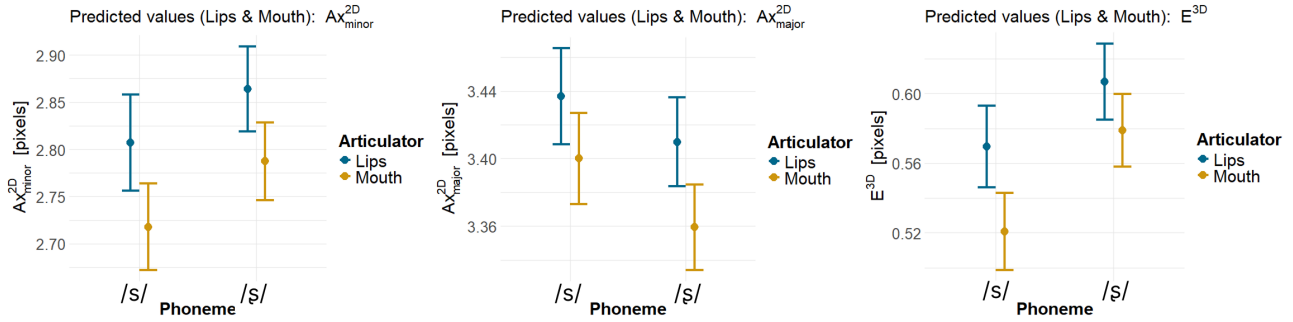


Figure 2: Values predicted by LME model for selected shape features of lips and mouth, each grouped by phonemes and articulators.

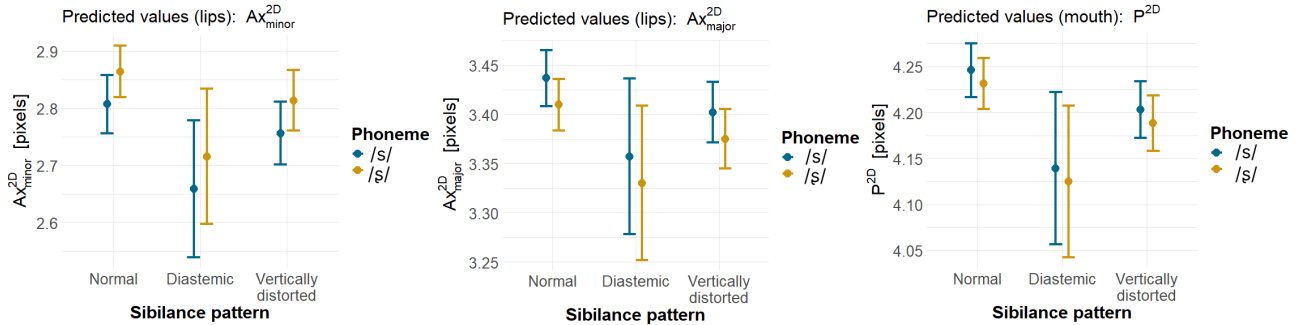


Figure 3: Values predicted by LME model for selected shape features of lips and mouth, each grouped by fixed effects.

mal sibillance gap realization and lowest in the diastema. This corresponds to creating an airflow gap in diastema, so all these parameters are expected to be lower than in normal and vertically distorted sibillance patterns. In turn, in a vertically distorted manner, the upper and lower incisors overlap excessively or are too far apart. As we had no information on the type of distortion and the range of results was between normal and diastemic, we assumed that excessive overlap of the incisors prevailed in children. However, the results suggest that the area between the lips was still greater than in the case of a diastema.

Our analysis also included shape features of the tongue, as it plays a key role in articulation. However, we only observed insignificant results in all parameters, so we presented results limited to the mouth and lips. Of all articulators subjected to segmentation, the tongue was the most difficult to delineate accurately. Lips were almost always visible in our video frames, whereas the tongue was present less frequently. Our experiments suggest a limited usability of the tongue features in the sibillance analysis.

We investigated an interaction `Phoneme * Sibillance` in LME models to examine whether differences in feature distribution between gap creation patterns are phoneme-dependent. Retroflex phonemes (e.g., /ʂ/) appear late in children [30], and they are often substituted with sounds with a dental place of articulation [31]. It is defined as a substitution where the realization of retroflex is identical to the realization of the dental phoneme (e.g., /s/). However, only two visual features were statistically significant in the following configurations (both for mouth) — Ax_{major}^{3D} Phoneme: /ʂ/ and Sibillance: diastemic ($p = 0.04$, $Est. = 0.03$, $Std.Err. = 0.02$, $T = 2.03$) and SD^{3D} Phoneme: /ʂ/ and Sibillance: vertically distorted ($p = 0.05$, $Est. = 0.01$, $Std.Err. = 0.01$, $T = 2.00$). However, as we found no associations between the results and

phenomena in sibillance production, we focused on the individual fixed effects in this paper.

Our study has some limitations. We analyzed only two out of twelve Polish sibilants using restrained language material. The availability of Polish words that contain given sounds in uniform acoustic neighborhoods and are present in the active vocabulary of preschool children is limited. We also consider a more comprehensive set of visual features in the future as our segmentation method may extract other articulators. However, to our knowledge, the group of 189 children (of 201 recorded by our team) constitutes the largest population of child speakers included in an acoustic analysis of the articulation of Polish sibilants so far.

5. Conclusions

Our study showed significant visual differences between phonemes /s/ and /ʂ/ produced by Polish children with normal, diastemic, and vertically distorted sibillance patterns. An investigation based on data from 189 speakers proved that both aspects could be distinguished by the shape of the lips and the mouth. The presented experiments demonstrate that visual data of the oral area contain diagnostic information about speech. However, speech production is a complex and multilayered process, so further development is necessary to encompass different aspects of articulation.

6. Acknowledgements

This work was supported partially by the National Science Centre, Poland, research project No. 2023/51/D/HS2/02577: "Longitudinal investigation of sibilant articulation development in children: a statistical modeling approach based on instrumental evidence and data mining methods", and partially by the Pol-

ish Ministry of Science, Poland, statutory financial support No. 07/010/BK_25/1047.

7. References

- [1] L. L. Koenig, C. H. Shadle, J. L. Preston, and C. R. Mooshammer, "Toward improved spectral measures of /s/: Results from adolescents," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 4, pp. 1175–1189, 2013.
- [2] N. Francis, M. Jamie, and B. Gick, "Aerotactile acuity as a predictor of sibilant contrast," *Canadian Acoustics*, vol. 40, no. 3, pp. 22–23, Sep. 2012.
- [3] K. Iskarous, C. H. Shadle, and M. I. Proctor, "Articulatory-acoustic kinematics: The production of american english /s/," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 944–954, 02 2011. [Online]. Available: <https://doi.org/10.1121/1.3514537>
- [4] O. Amr Rey, P. Sánchez Delgado, R. Salvador Palmer, V. Paredes Gallardo, and R. M. Cibrián Ortiz de Anda, "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," *Psicología educativa*, vol. 28, no. 2, pp. 195–207, 2022.
- [5] J. A. T. Sarah Lockenvitz and J. Oxley, "The sociolinguistics of lispings: a review," *Clinical Linguistics & Phonetics*, vol. 34, no. 12, pp. 1169–1184, 2020, pMID: 32646249.
- [6] A. Trębacz, "Articulation disorders of the sigmatism type and the level of self-esteem of students completing the stage of early school education. own research conducted in the greater poland region," *Interdyscyplinarne Konteksty Pedagogiki Specjalnej*, no. 30, pp. 187–222, 2020. [Online]. Available: <https://pressto.amu.edu.pl/index.php/ikps/article/view/27063>
- [7] C. Deka, A. Shrivastava, A. Abraham, S. Nautiyal, and P. Chauhan, "Ai-based automated speech therapy tools for persons with speech sound disorder: a systematic literature review," *Speech, Language and Hearing*, pp. 1–22, 06 2024.
- [8] N. Gohider and O. A. Basir, "Recent advancements in automatic disordered speech recognition: A survey paper," *Natural Language Processing Journal*, vol. 9, p. 100110, 2024.
- [9] I. Anjos, M. Grilo, M. Ascensão, I. Guimarães, J. Magalhães, and S. Cavaco, "A model for sibilant distortion detection in children," 11 2018.
- [10] Z. Miodonska, P. Badura, and N. Mocko, "Noise-based acoustic features of Polish retroflex fricatives in children with normal pronunciation and speech disorder," *Journal of Phonetics*, vol. 92, p. 101149, 2022.
- [11] J. Li, M. Hasegawa-Johnson, and K. Karahalios, "Enhancing child vocalization classification with phonetically-tuned embeddings for assisting autism diagnosis," 09 2024, pp. 5163–5167.
- [12] M. Bourhis, P. Perrier, C. Savariaux, and T. Ito, "Quick speech motor correction in the absence of auditory feedback," *Frontiers in Human Neuroscience*, vol. 18, 2024.
- [13] A. Lee, M. Liker, Y. Fujiwara, I. Yamamoto, Y. Takei, and F. Gibbon, "Epg research and therapy: further developments," *Clinical linguistics & phonetics*, vol. 37, pp. 1–21, 06 2022.
- [14] K. S. Nataraj, P. Pandey, and H. Dasgupta, "Estimation of place of articulation of fricatives from spectral features," *International Journal of Speech Technology*, vol. 26, pp. 1–18, 12 2023.
- [15] Z. Miodonska, M. Krecichwost, E. Kwasniok, A. Sage, and P. Badura, "Frication noise features of Polish voiceless dental fricative and affricate produced by children with and without speech disorder," in *Interspeech 2024*, 2024, pp. 3125–3129.
- [16] S. Hamilton, S. Schwab-Farrell, R. Seward, J. Avant, T. Zhang, S. Li, K. Eary, T. D. Mast, M. Riley, and S. Boyce, "A qualitative analysis of clinician perspectives of ultrasound biofeedback for speech sound disorders," *American Journal of Speech-Language Pathology*, vol. 32, pp. 1–23, 03 2023.
- [17] M. Ruthven, A. M. Peplinski, D. M. Adams, A. P. King, and M. E. Miquel, "Real-time speech mri datasets with corresponding articulator ground-truth segmentations," *Scientific Data*, vol. 10, no. 860, 2023.
- [18] Z. Bilková, A. Novozámský, M. Bartoš, A. Domínez, Šimon Greško, B. Zitová, M. Paroubková, and J. Flusser, "Human computer interface based on tongue and lips movements and its application for speech therapy system," *Journal of Electronic Imaging*, vol. 32, no. 1, pp. 389–1–389–1, 2020.
- [19] A. Gaodida, H. Koppisetty, K. Potdar, and A. Biwalkar, "Aiding speech therapy using audio and video processing," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–5.
- [20] A. Sage, Z. Miodonska, M. Krecichwost, and P. Badura, "Hybridization of acoustic and visual features of Polish sibilants produced by children for computer speech diagnosis," *Sensors*, vol. 24, no. 16, 2024. [Online]. Available: <https://www.mdpi.com/1424-8220/24/16/5360>
- [21] E. Krajna, "Developmental phonetic norm – expectations and facts, (PL) rozwojowa norma fonetyczna – oczekiwania i fakty," *Logopedia*, vol. 1, no. 1, pp. 33–46, 2005.
- [22] T. Yoshinaga, K. Tada, K. Nozaki, and A. Iida, "A simplified model for the vocal tract of [s] with inclined incisors," in *Interspeech 2021*, 2021, pp. 3166–3170.
- [23] K. M. Leavy, G. J. Cisneros, and E. M. LeBlanc, "Malocclusion and its relationship to speech sound production: Redefining the effect of malocclusal traits on sound production," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 150, no. 1, pp. 116–123, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0889540616001980>
- [24] M. Krecichwost, A. Sage, Z. Miodonska, and P. Badura, "4d multimodal speaker model for remote speech diagnosis," *IEEE Access*, vol. 10, pp. 93 187–93 202, 2022.
- [25] A. Sage and P. Badura, "Detection and segmentation of mouth region in stereo stream using yolov6 and deeplab v3+ models for computer-aided speech diagnosis in children," *Applied Sciences*, vol. 14, no. 16, 2024. [Online]. Available: <https://www.mdpi.com/2076-3417/14/16/7146>
- [26] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 10 2017. [Online]. Available: <https://doi.org/10.1158/0008-5472.CAN-17-0339>
- [27] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, and E. Neri, "A deep look into radiomics," *La Radiologia Medica*, vol. 126, pp. 1296–1311, 2021.
- [28] S. Hamann, "Retroflex fricatives in slavic languages," *Journal of the International Phonetic Association*, vol. 34, no. 1, pp. 53–67, 2004.
- [29] A. Lorenc, M. Żygis, Łukasz Mik, D. Pape, and M. Sósokuthy, "Articulatory and acoustic variation in polish palatalised retroflexes compared with plain ones," *Journal of Phonetics*, vol. 96, p. 101181, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0095447022000560>
- [30] M. Żygis, D. Pape, M. Jaskała, and L. L. Koenig, "Do children better understand adults or themselves? an acoustic and perceptual study of the complex sibilant system of polish," *Journal of Phonetics*, vol. 100, p. 101227, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0095447023000165>
- [31] D. Emiluta-Rozya and D. Lipiec, "Articulation disorders - causes, symptomatology, classifications, (PL) Zaburzenia artykulacji – przyczyny, symptomatologia, klasyfikacje," in *Preschool and early school speech therapy*, A. Domagała and U. Mirecka, Eds. Wydawnictwo Harmonia, 2021.