Biomedical Signal Processing and Control 68 (2021) 102612

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Automated detection of sigmatism using deep learning applied to multichannel speech signal

Michal Krecichwost [a], Natalia Mocko [b,c], Pawel Badura [a,*]

[a] Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland
[b] Faculty of Humanities, Institute of Linguistics, University of Silesia, Sejmu Śląskiego 1, 40-001 Katowice, Poland
[c] Good Speech Zone Natalia Moćko, Strefa Dobrej Mowy Natalia Moćko, ul. Rataja 8/17, 40-837 Katowice, Poland

ARTICLE INFO

Keywords:
Multichannel acoustic signal
Speech analysis
Sigmatism
Deep learning
Convolutional neural network
Computer-aided speech diagnosis

ABSTRACT

This paper presents a system for the analysis of acoustic data for the computer-aided diagnosis and therapy of sigmatism in children. The analysis is focused on the detection and recognition of selected articulation disorders in sibilant sounds. The system relies on the dedicated data acquisition device recording the speech signal using 15 microphones spatially arranged around the speaker's mouth. The collected speech corpus contains 923 samples of the /s/ and /ʃ/ consonants from 98 five- and six-year-old children with either normative or pathological pronunciation features. Each recording is supplemented with a detailed speech therapy annotation. A dedicated multibranch convolutional neural network architecture was designed for the speech sample classification. The filter bank energy feature maps are extracted from each channel along with their two derivatives in the time domain. The feature maps are aggregated along different dimensions to constitute a four-dimensional data structure called acoustic volume, being the input data for the deep network. We proposed three ways to aggregate the multichannel data into the acoustic volume and two techniques for the data augmentation to enlarge the available dataset and avoid overfitting. Classification experiments involving different data subsets have proven the system's ability to detect the analyzed pronunciation disorders with reasonable accuracy. The framework with speech data organized spatially in five channels provides the most efficient classification.

1. Introduction

Dental sounds (sibilants) are considered to be one of the most difficult to articulate in Polish. The phonological system contains 12 sibilants in three series (/, z, ts, dz/, /ʃ, ʒ, tʃ, dʒ/, /ɕ, ʑ, tɕ, dʑ/), which, in relation to other European languages, is a quite large number (e.g., in English there are six sibilants: /s, z, ʃ, ʒ, tʃ, dʒ/). The normative pronunciation of dentalized sounds is a great challenge for preschool children. Difficulties appearing at the peripheral level are considered to be sigmatism (lisp) [1].

A child with sigmatism may deform one, two sounds, a series (four sounds), or all three series of sibilants. A non-normative pronunciation may be one of the following: elution (missing the sound, lack of articulation), substitution (replacing the sound with another one), or deformation (actual sigmatism). The latter disorder arises from the distortion of the sound, which is caused by a change in the place or type of its articulation [2].

Many types of normative and abnormal implementation of sibilants are described in the speech therapy literature [3]. In this study, four types were selected, due to the general agreement on their commonness: normative, interdental, addental, and dental articulation. In general, the main feature differentiating individual pronunciation types is the apex position in the oral cavity [4,5].

1.1. State of the art

Tools that allow to objectify the diagnostic process and redefine certain speech therapy concepts are studied and developed in various science centers [4]. The electromagnetic articulography [6,7] is a method for tracking the motion of articulators (lips, tongue, mandible, and soft palate) using an alternating magnetic field. Some works involve the multichannel audio signal recording [8,9] being a part of the research aimed at assessing the distribution of the acoustic field in pathological articulation. Multimodal systems are also being built to support the rehabilitation of people with motor speech disorders (such as aphasia) [10,11]. The multimodal speech capture system (MSCS)

* Corresponding author.
E-mail addresses: michal.krecichwost@polsl.pl (M. Krecichwost), natalia.mocko@us.edu.pl (N. Mocko), pawel.badura@polsl.pl (P. Badura).

https://doi.org/10.1016/j.bspc.2021.102612
Received 28 July 2020; Received in revised form 18 November 2020; Accepted 19 March 2021
Available online 20 April 2021

enables the recording of an acoustic signal, articulators' image, and tongue movements.

Automatic speech recognition (ASR) systems usually use single-microphone data. However, they also involve signals from multiple microphones. They do so to amplify the signal and reduce the reverberation effect and background noise [12]. Signal processing in multichannel ASR systems mostly consists of three stages: localization of the sound source, beamforming, and filtration. The resulting signal is aggregated to a single channel and amplified. It is then subjected to conventional processing, recognition, and classification methods [13]. The spatial speech signal is used for speech recognition. However, few works attempt to adapt spatial information to recognize pathological sounds. Current works use data from a single audio channel, limiting the diagnostic value of the data.

Studies using the acoustic analysis to detect the improper realization of sounds of different languages mainly focus on pronunciation errors while learning a foreign language [14,15]. Some works are dedicated to the detection and classification of non-normative pronunciation [16–18]. A small group of works is targeted at the non-normative articulation of Polish sounds [19–21]. Systems dedicated to Arabic, Chinese, or German can also be found [15,17,18]. However, the research mainly employs databases of recordings of adults simulating speech disorders. The literature shows that adult speech analysis methods often fail in children due to essential differences in spectral characteristics [22]. Studies on this subject mainly concern the phone substitution [16,19]. However, they do not focus on other types of pathology, such as deformation.

Standard features for speech recognition systems, e.g., Mel-frequency cepstral coefficients (MFCC) or human factor cepstral coefficients (HFCC), are most commonly used to describe dentalized sounds either [23,24]. Other characteristics to be mentioned include fricative formants and their levels [25,26], spectral moments, the spectral flatness measure (SFM) [27], the center of gravity (CoG), the spectrum shape, the sound duration [28,29], and the linear predictive coding (LPC) coefficients [30,31]. The research focuses on a binary classification of the speech sample: normative vs. pathological pronunciation [32,33]. There is a lack of studies on the definition of a specific type of disorder, which is important for the speech diagnosis.

Deep learning techniques are willingly employed for speech signal processing. The raw, single-channel acoustic signal is used in most cases so far [34–36]. This limits the possibility of spatial interpretation of input data that cannot be reflected only in the time domain. Among others, the recurrent neural networks (RNN) are used to analyze time signals. For natural language processing, the long short-term memory networks (LSTM) are used most often [37,38]. LSTMs aim to identify long-term time patterns in the input data. They are popular in hybrid systems for emotion recognition [39].

Some recent works involve also two-dimensional representations of the signal, most often, a spectrogram [40–43]. The advantage of the 2D input data can be found in more comprehensive information about the signal features (e.g., time-frequency relationships). Khamparia et al. [41] used both CNN and tensor deep stacking network (TDSN) to classify environmental sounds based on their resized 180×180 spectrograms. However, their results are difficult to relate to the speech therapy domain due to a completely different acoustic scope. Woloshuk et al. [40] investigated the use of convolutional neural networks (CNN) supplied by multichannel speech signals for the classification of different articulations of a sound /s/. Three types of 2D input data were analyzed: the spectrogram, the filter bank energy (FBE) feature map, and the MFCCs. The FBE-driven model yielded the highest classification efficiency. Shahrebabaki et al. [44] made similar observations as a result of a comparison of common acoustic features: MFCC, LPC, FBE, or line spectral frequencies (LSF) in the classification of selected phones. The FBEs produced results with higher stability and accuracy than the other features.

1.2. Aims and scope

The general aim of this study was to design an acoustic data analysis system for computer-aided diagnosis and therapy of sigmatism in children. The goal was to create and train a tool able to detect and classify selected disorders in the articulation of sibilant sounds. For the speech signal recording, we used a dedicated multichannel device described in [45]. The device acquires the acoustic signal from 15 microphones spatially arranged around the speaker's face. We collected a database of speech samples from 98 children aged five or six with comprehensive diagnostic annotations provided by speech therapy experts. The analysis involved 923 samples of the /s/ and /ʃ/ sounds extracted from the multichannel recordings. We divided the study into three experiments with different sets of articulation disorders.

For the speech sample classification, we designed a dedicated multibranch convolutional neural network. Based on the previous research [40], we involved the filter bank energies and their two derivatives in the time domain as base features. With multiple sources of individual acoustic signals, we proposed three feature map aggregation techniques, each yielding a four-dimensional time-frequency-space acoustic volume to supply the CNN. The multibranch part of our network automatically extracts features from different domains of the speech signal representation. To enlarge the database and avoid overfitting, we proposed two dedicated data augmentation techniques performed on both raw and transformed signals. The network parameters and settings were justified using a detailed analysis of its performance in various experiments.

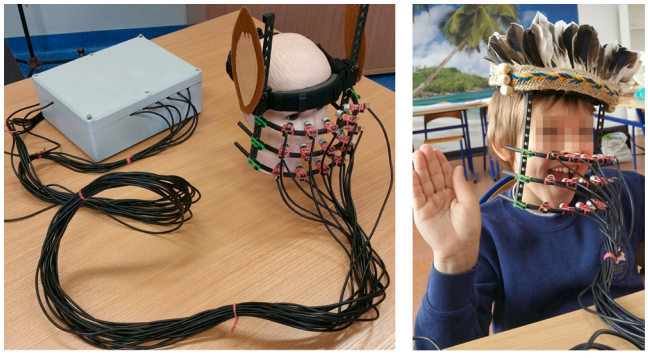
1.3. Paper structure

Section 2 presents the materials and methods: the data acquisition device, the speech corpus, and detailed specification of the methodology in terms of data preparation and the structure of the original CNN. The experiments and results are described in Section 3 and discussed in Section 4. Section 5 concludes the paper.

2. Materials and methods

2.1. Multichannel speech acquisition setup

The speech data were collected by using a dedicated acoustic mask for multichannel spatial speech recording (Fig. 1), described in detail in [45]. We acquire the speech signal using 15 Panasonic WM-61a microphones located on three rigid arches (five each) mounted to two connectors. The proposed assembly method enables adjusting the device to the individual head size and shape. Modifications in microphone positioning related to the speaker's mouth are also possible. However, in this study we established some general settings consulted with the speech therapy experts (Fig. 2): (1) microphone #8 located centrally in front of



(a) (b)

Fig. 1. The multichannel speech signal acquisition device shown with the data processing unit (a) and during the speech therapy examination (b).

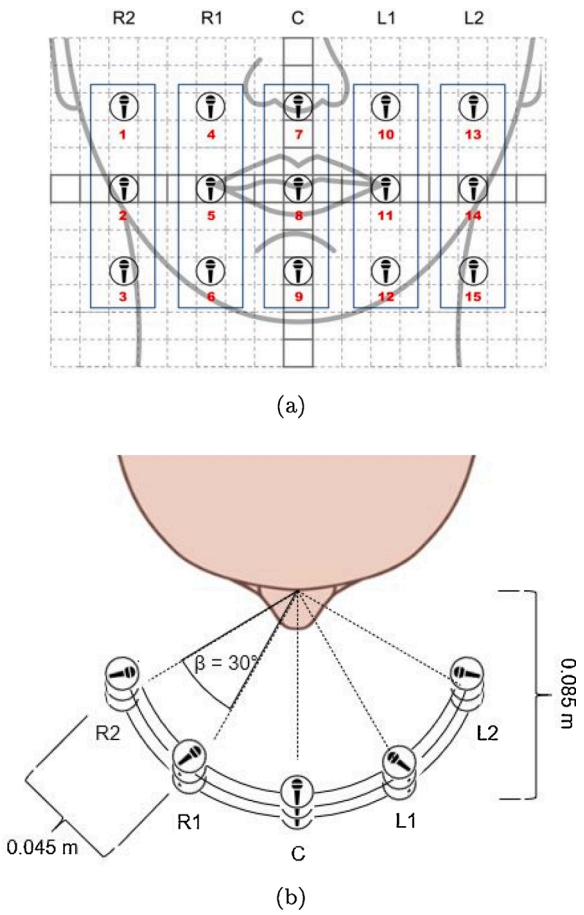


Fig. 2. Arrangement of microphones used in the study in a front (a) and top (b) view. Red font denotes the microphone number, black label refers to the uniform linear array (ULA) ID.

the mouth, (2) a distance between adjacent microphones in a row or column of ca. 4.5 cm, (3) a distance between a sensor and subject's mouth of ca. 8.5 cm. The mounting straps are equipped with removable sponges to secure wearing comfort and to adjust to the head size. Moreover, some decorative accessories, e.g., headdress or rabbit ears can be attached to make the mask friendly for the examined five- or six-year-old children (Fig. 1b).

Each microphone acquires a signal with a 44.1 kHz sampling frequency and passes it to the data processing unit. Then, the signals are amplified using the MAX9812 unit, A/D converted at 16 bits with the DaqBoard/3000USB Series data acquisition board and transmitted at 1 MHz to the computer [45].

2.2. Speech corpus

A database for computer-aided speech diagnosis and therapy was gathered from 98 children aged five or six [45]. The database consists of 923 multichannel recordings of 16 Polish words containing sibilants /s/ and /ʃ/ (Table 1). The words were illustrated in individual pictures, and the child's task was to name the picture. For each recording, a diagnostic description was prepared by the speech therapy expert. The assessment was made to determine the normative or pathological pronunciation of sibilants. Moreover, abnormal phone articulations were analyzed and annotated for the pathology type. Six classes of pronunciation type were detected:

- s_{norm} – normative /s/,
- s_{add} – addental /s/,

Table 1
Words used during the speech therapy examination and data acquisition.

Phoneme /s		Phoneme /ʃ	
Original word	English translation	Original word	English translation
samolot	*plane*	szafa	*wardrobe*
serce	*heart*	sznur	*cord*
strażak	*firefighter*	szufelka	*shovel*
pasek	*belt*	kalosze	*wellingtons*
parasol	*umbrella*	koszyk	*basket*
lis	*fox*	książka	*book*
pies	*dog*	lekarz	*doctor*
		nóż	*knife*
		wąż	*snake*

- s_{int} – interdental /s/,
- $\int_{norm}$ – normative /ʃ/,
- $\int_{int}$ – interdental /ʃ/,
- $\int_{den}$ – dental /ʃ/.

Distribution of speech samples through different classes is presented in Table 2. The normative speaker was defined as a person with the correct pronunciation of a given sibilant, a proper phonemic hearing (distinguishing speech sounds), without anatomic disorders of articulators or malocclusion, and with a full set of primary teeth. The participants did not feature respiratory tract infection. We had written consent from the child's parents or legal guardians and verbal agreement from the child to participate in the study.

2.3. Methodology

The acquired multichannel speech signal is processed according to the workflow presented in Fig. 3 and described in the following sections.

2.3.1. Data preprocessing

Before describing the individual signal processing workflow, it has to be noted that with 15 spatially distributed microphones, we have several possibilities to analyze the speech signal. In this study, we tested three frameworks with a different number of signals involved:

- a one-channel setup (1-CH) – the sole central microphone signal (Fig. 2a, mic #8),
- a 15-channel setup (15-CH) – all signals processed and analyzed independently until the CNN classification,
- a five-channel setup (5-CH) – five signals aggregated within five vertically oriented uniform linear arrays (ULA). The arrays are shown and denoted as R2, R1, C, L1, L2 in Fig. 2.

For the latter framework, an initial preprocessing step is required to aggregate the ULA signals (optional dashed box in Fig. 3). The procedure is described in detail in [45]. It employs the time delay of arrival (TDOA) algorithm [46] to estimate the incidence wave angle by using the generalized cross-correlation with phase transform (GCC-PHAT) [47]. With the obtained angle, the aggregated ULA signal can be determined through the delay-and-sum beamforming (DAS) [48]. As a result, five ULA signals are passed to the following blocks.

Irrespective of the framework (1-CH, 5-CH, or 15-CH), each single-channel signal under consideration was first subjected to a pre-emphasis filtering to enhance the high-frequency band essential for the sibilant analysis. We used the pre-emphasis filter with a coefficient

Table 2
Distribution of recordings through classes in the study.

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

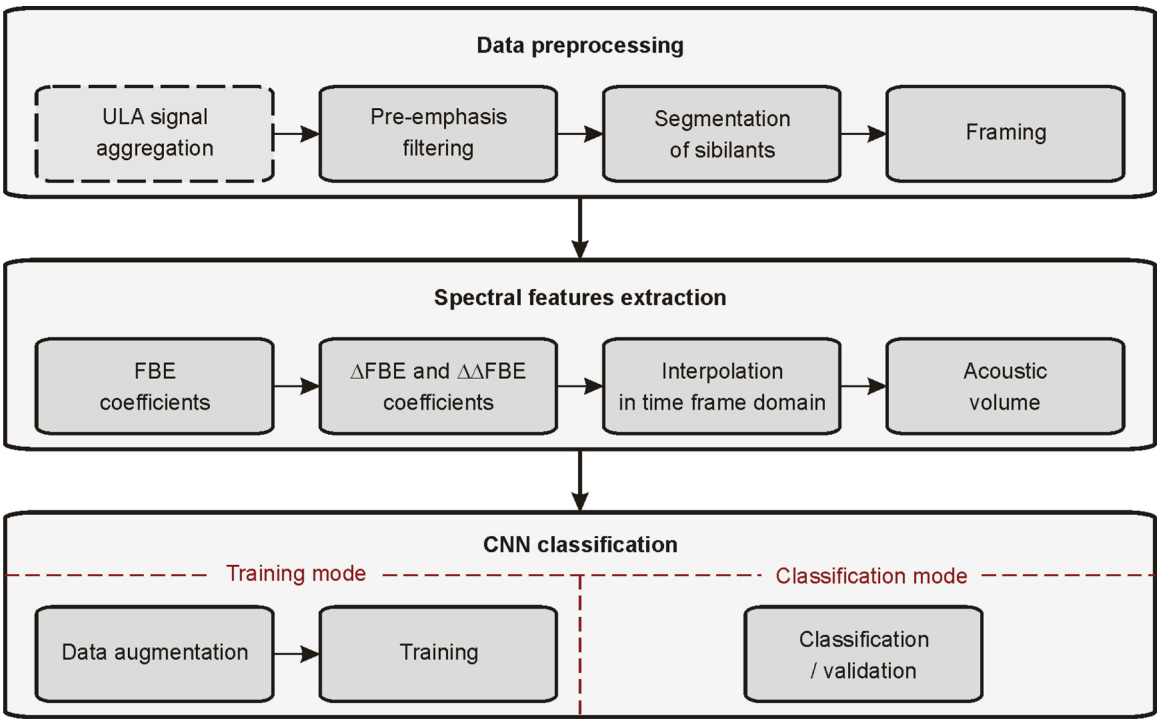


Fig. 3. Overall workflow of the multichannel speech sample classification.

$\alpha = 0.93$, yielding a ca. 6 dB gain within the higher frequency band. Then, all recorded signals and words were manually segmented to extract the sibilants. The process was performed independently by three experts in speech therapy and acoustic analysis. The final phoneme boundaries were determined by averaging the corresponding time-stamps provided by all three experts. The average segment duration was 172 ± 65 ms (a 39-820 ms range). Finally, the signals were divided into frames of a 15 ms duration with a Hamming window and a 10 ms overlap.

2.3.2. *Spectral features extraction*

In our previous works, we investigated multiple handcrafted feature vectors, including the MFCCs, RMS, fricative formants, or spectral moments [33,45,49]. However, we involved some basic classification CNNs for binary classification of different pronunciations based on spectral feature maps: spectrograms, FBEs, and MFCCs [40]. The FBE-related feature map produced the most efficient classification in [40]. Thus, in our current research, we focused on feeding the CNN classifier with the FBE features.

Procedures for preparation of data are presented in the middle block in Fig. 3. The basic FBE measures are determined based on responses of 64 passband filters with band centers spread linearly in a 1–22 kHz range. The FBE band covers relatively high acoustic frequencies due to the noise character of sibilant sounds. As a result, an FBE feature map of a $64 \times N_f$ size is obtained for each segment consisting of N_f frames.

Next, we employed two types of dynamic FBE features: the Delta and Delta–Delta coefficients [50,51]. The ΔFBE coefficients can be treated as a partial derivative of the FBE feature map over time. The cth Delta coefficient $\Delta\mathrm{FBE}_{f,c}$ within the fth frame is given by:

$$\Delta\mathrm{FBE}_{f,c} = \frac{\sum_{k=1}^{W} k(\mathrm{FBE}_{f+k,c} - \mathrm{FBE}_{f-k,c})}{2\sum_{k=1}^{W} k^2}, \tag{1}$$

where W is a half processing window width, $FBE_{f,c}$ is the cth FBE coefficient in the fth frame. The ΔFBE idea is shown in Fig. 4.

The ΔΔFBE coefficients are determined accordingly, as a partial derivative of the ΔFBE feature map within the time frame domain:

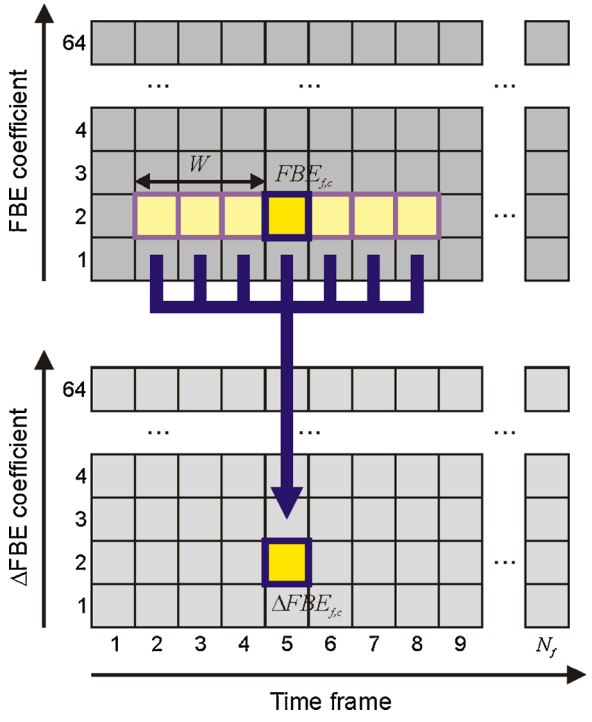


Fig. 4. Illustration of the determination of Delta coefficients. Delta–Delta coefficients are computed from Delta coefficients according to the same scheme.

$$\Delta\Delta\mathrm{FBE}_{f,c} = \frac{\sum_{k=1}^{W} k(\Delta\mathrm{FBE}_{f+k,c} - \Delta\mathrm{FBE}_{f-k,c})}{2\sum_{k=1}^{W} k^2}. \tag{2}$$

In either case, we set the half window width W to 3. In boundary cases, missing samples from outside the feature map are replaced by the nearest border frame's coefficients.

The sibilant segment duration differs between speakers and realizations. Thus, in general, feature maps do not have the same size in the time frame domain. Since the next processing stages require a fixed size of the input data, each feature map has to be padded or resampled to reach the assumed size of 64 × 64 coefficients. We have chosen the horizontal bilinear interpolation method from several approaches under consideration (zero-padding, last-frame padding, various interpolations).

With a set of three 64 × 64 FBE feature maps per channel, the next step was to combine the data and prepare an input structure for the classification. We called this structure an acoustic volume. It is prepared according to the following principles, taking into account the number of channels:

1 For the individual channel (signal), all three feature maps are concatenated along the third dimension. That might be referred to as a three-channel chromatic image color model, like RGB. So, the FBE, ΔFBE, and ΔΔFBE maps constitute the red, green, and blue channels of the chromatic image, respectively (Fig. 5).

   However, before the concatenation, the FBE map is transformed logarithmically, and each map is normalized to a 0–1 range. For the 1-CH setup, the acoustic volume is completed at that point.
2 In the 5-CH and 15-CH frameworks, there are multiple acoustic images produced by individual channels. The images are concatenated along the other dimension to prepare a 3D volume of color images (Fig. 6). In a 5-CH setup, the order of channels is R2-R1-C-L1-L2 (Fig. 6a), whereas in 15-CH the channels are ordered according to their numbers (Fig. 6b).

In fact, the acoustic volume is a four-dimensional collection of coefficients and has a size of 64 × 64 × CH × 3, with CH = 1, 5, or 15. Subsequent dimensions refer to the frequency, time, microphone/channel, and feature type, respectively.

2.3.3. Classification using multibranch convolutional neural network

All classification tasks addressed in this study were performed using a multibranch CNN model of a structure presented in Fig. 7a. The input acoustic volume is passed to three branches of the feature extraction (convolutional) part of the network. Each branch consists of five processing blocks of a general structure shown in Fig. 7b. Each block includes a 3D convolutional layer, a mini-batch normalization layer, a ReLU layer, and a max-pooling layer. Different branches are responsible for analyzing the volume and extracting features in different domains (dimensions or projections): in the time, frequency, and mixed time-frequency domain. Thus, processing blocks in branches differ in the kernel shapes for both convolutional and pooling layer. Kernel shapes and sizes in convolutional layers are illustrated in Fig. 8. Note that the kernels are, in fact, four-dimensional, with the fourth dimension corresponding to the RGB channels in the first processing block layer, and the 3D feature map number in the other ones. In each branch, subsequent convolutional layers contain 8, 8, 16, 16, and 32 kernels. Accordingly, pooling kernel shapes are adjusted to provide feature maps of the same size at the end of each branch (1 × 2 ×1, 2 × 2 ×1, 2 × 1 ×1 for the time, time-frequency, and frequency branches, respectively).

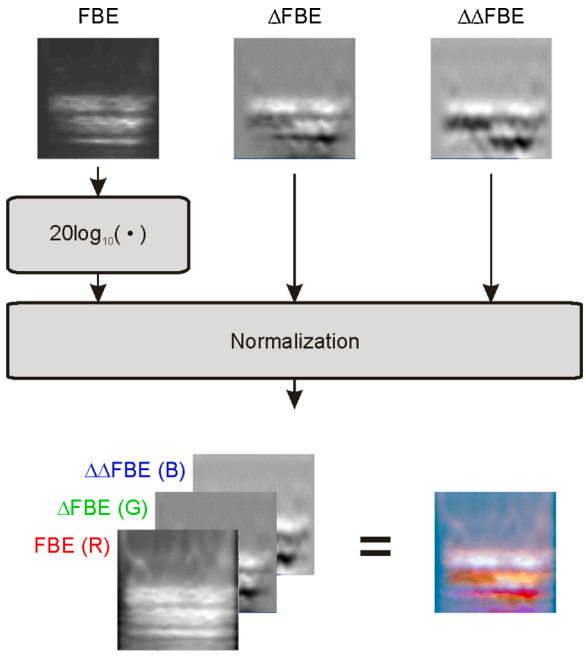


Fig. 5. Illustration of the acoustic image for a single channel.

Outputs of three branches are concatenated and passed to the classification part of the network. It consists of a total of three fully connected layers (Fig. 7c). First two have 1024 neurons each, being followed by a ReLU layer and a dropout layer with p = 0.2. The third fully connected layer has M neurons, where M is the number of classes considered in the experiment. Finally, the outputs of the latter layer are subjected to the softmax normalization and classification.

We trained the network using a stochastic gradient descent optimizer with momentum (SGDM) with a learning rate of 0.001 and a momentum of 0.95. The batch size was set to 128 and the maximum number of epochs to 25. The training process was terminated when the validation loss did not decrease in the recent five epochs. The distribution of classes in the training data was balanced through class weighting by means of a weighted cross-entropy loss function.

During training, we employed data augmentation techniques to increase the amount of training data and avoid overfitting. We did not involve several data augmentation techniques known from the image analysis approaches (e.g., rotation or reflection) since they barely address the phenomena of the spectral feature map proposed here. Instead, two dedicated techniques were employed:

1 Augmentation through segment partition in the time domain: we assumed that the articulation could be partitioned into three phases (the beginning, middle, and end of the phone) and we roughly estimated the partition in the time domain as 3:4:3 (Fig. 9).

   Thus, we divided each acoustic volume from the training set into three subvolumes and then resampled each of them through bilinear interpolation in the time domain to the desired input data size of 64 × 64 × CH × 3.
2 Augmentation through pre-emphasis: the other technique did not transform the input acoustic volume itself, yet it affected its production at the very early stage of preprocessing in the time domain – pre-emphasis filtering. Namely, we prepared augmented acoustic volumes corresponding to the volumes from the training set with different values of the pre-emphasis coefficient α of 0.90, 0.91, and 0.92, affecting the FBEs and their derivatives.

As a result, combinations of both techniques yielded a 16× increase in the training data size.

The training and validation were performed using a Deep Learning Toolbox (version 12.1) of a Matlab software (9.6.0.1099231, R2019a).

3. Experiments and results

With the speech corpus described in Section 2.2 and the CNN architecture from Section 2.3, we defined three pronunciation classification experiments:

1 Analysis of a phoneme /s/ (experiment E1) with samples from M = 3 classes: s_{norm}, s_{add}, and s_{int}.

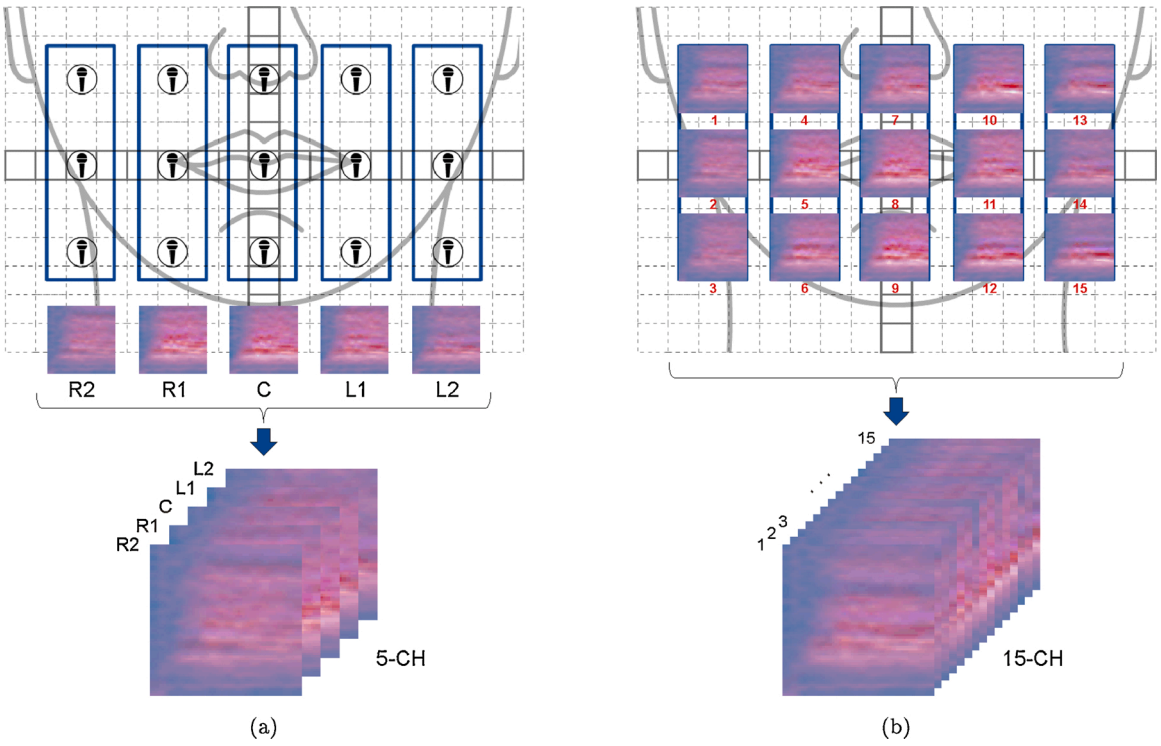


Fig. 6. Illustration of the acoustic volume in the five- (a) and fifteen-channel (b) frameworks.

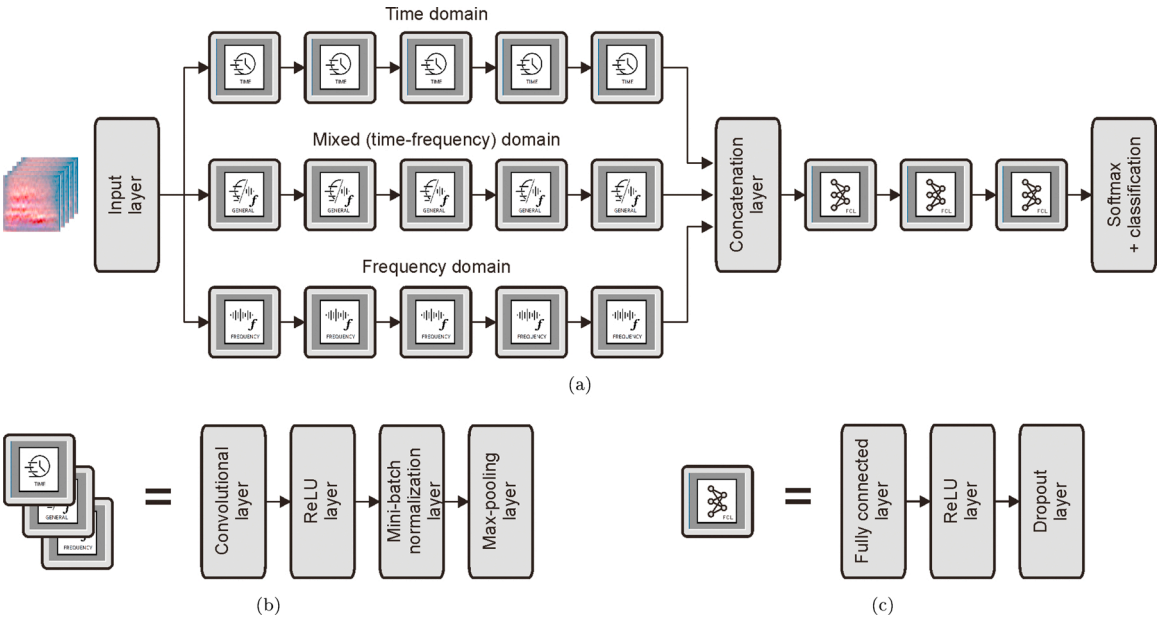


Fig. 7. General architecture of the CNN used for classification (a). Structures of the feature-extraction processing blocks and the fully connected layer blocks are shown in (b) and (c), respectively.

2 Analysis of a phoneme /ʃ/ (experiment E2) with samples from $M = 3$ classes: $ʃ_{norm}$, $ʃ_{int}$, and $ʃ_{den}$.
3 Analysis of both phonemes /s/ and /ʃ/ (experiment E3) with samples from $M = 6$ classes: s_{norm}, s_{add}, s_{int}, $ʃ_{norm}$, $ʃ_{int}$, and $ʃ_{den}$.

In each experiment, the dataset was randomly partitioned into the training, validation, and testing subsets (70%:15%:15%) in a speaker-wise mode, i.e., speech samples of a particular speaker could be present in one subset only. The speaker-wise partition makes the validation more reliable since potentially similar speaker-related speech samples cannot be used for both training and testing. We repeated each experiment E1–E3 10 times, each with a new random distribution of samples into the training, validation, and test subsets.

The classification efficiency metrics employed for the multiclass assessment in a single run were as follows:

(3) Sensitivity (true positive rate):

$$\mathrm{TPR} = \frac{1}{M}\sum_{i=1}^{M}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}. \tag{3}$$

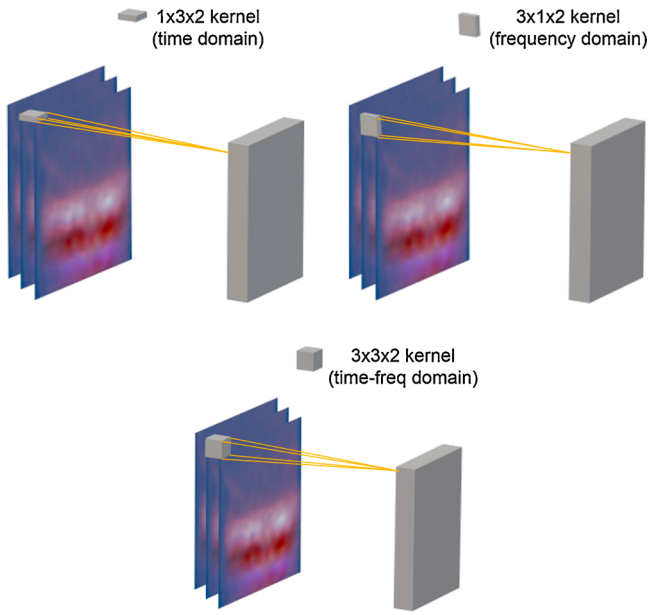


Fig. 8. Illustration of kernels used for convolution in different branches of the convolutional part of the CNN.

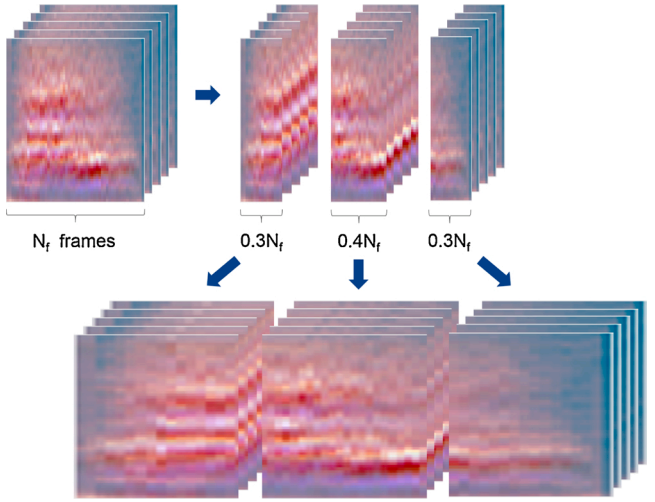


Fig. 9. Illustration of data augmentation through segment partition in the time domain.

(4) Specificity (true negative rate):

$$\mathrm{TNR} = \frac{1}{M}\sum_{i=1}^{M}\frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}. \tag{4}$$

(5) Accuracy:

$$\mathrm{ACC} = \sum_{i=1}^{M}\frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i}. \tag{5}$$

TP_i, TN_i, FP_i, FN_i in (3)–(5) denote the true positive, true negative, false positive, and false negative classifications within the ith of M classes, respectively. Metrics for the experiments are averaged through single runs.

In each experiment, we evaluated different settings of our method in terms of the number of channels involved (1-CH vs. 5-CH vs. 15-CH) with a different size and shape of an acoustic volume. Moreover, we compared the classification accuracies to those obtained by applying different filter banks at the feature extraction stage. The difference is in the frequency scale (Mel vs. linear) and range (depending on the experiment and spectral characteristics of sibilants). Finally, we compared the results to those yielded by the reference method proposed in [52] for speech command recognition from spectrograms. The six-layer deep reference network was re-trained using transfer learning with the recordings from our training subset. The results are presented and compared in the following sections.

3.1. Experiment E1 – analysis of phoneme /s/

Table 3 presents the results obtained using three frameworks of our method and a reference network [52]. The results are compared in terms of the accuracy *ACC* in Fig. 10 for the proposed filter bank (linear scale, 1-22 kHz range) and two other banks: (1) within the Mel-scale 0.3–8 kHz band typical for speech analysis [53], and (2) within the linear-scale 4–12 kHz band covering the meaningful part of the /s/ sound spectrum [27]. The model involving a five-channel acoustic volume outperforms the other approaches.

3.2. Experiment E2 – analysis of phoneme /ʃ/

Results for the /ʃ/ sound classification are presented in a similar manner in Table 4 and Fig. 11. Here, we used a filter bank with a linear 2-6 kHz band dedicated to the /ʃ/ consonant [54]. The 5-CH framework still offers way better accuracy than the other models and a reference method, despite a general improvement in their performance.

3.3. Experiment E3 – analysis of phonemes /s/ and /ʃ/

Efficiency metrics obtained during the six-class experiment E3[a] are gathered in Table 5. To deepen the analysis, we present the accumulated confusion matrix for the 5-CH framework in Table 6. A class-wise precision (positive predictive value) for the ith class is introduced in Table 6:

$$\mathrm{PPV}_i = \sum_{i=1}^{M}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}. \tag{6}$$

Clearly, the lowest sensitivity and precision values were obtained for two types of /ʃ/ mispronunciation: $ʃ_{int}$ and $ʃ_{den}$. Both can be explained from the acoustics point of view. Sound /ʃ/ with a dental sigmatism features very similar acoustic characteristics to the normative realization of the sound /s/ (note large number of misclassifications between s_{norm} and $ʃ_{den}$ classes in Table 6). On the other hand, the $ʃ_{int}$ samples are most often misclassified with the s_{int} sounds. The location of the tongue is very similar during sound generation in these two disorders. The obtained results suggest that the speech therapy specification assumed rather phonological interpretation (study of the relationship between

Table 3
Summary of the results obtained in the experiment E1. Presentation format: median (25th–75th percentile).

	TPR	*TNR*	*ACC*
1-CH setup	**0.40** (0.37–0.42)	**0.71** (0.68–0.74)	**0.48** (0.42–0.53)
5-CH setup	**0.76** (0.73–0.79)	**0.89** (0.87–0.90)	**0.78** (0.77–0.82)
15-CH setup	**0.36** (0.31–0.40)	**0.69** (0.66–0.72)	**0.43** (0.38–0.48)
Reference method [52]	**0.41** (0.41–0.43)	**0.72** (0.71–0.73)	**0.47** (0.46–0.52)

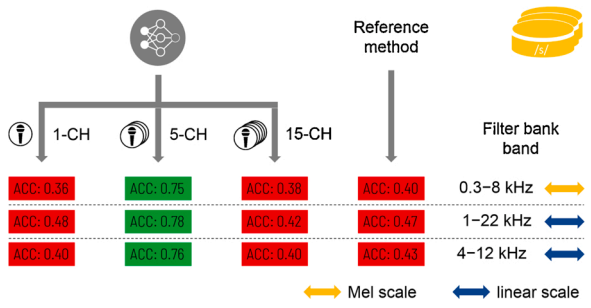


Fig. 10. Summary of the accuracy results obtained in the experiment E1 for different filter banks.

Table 4
Summary of the results obtained in the experiment E2. Presentation format: median (25th–75th percentile).

	TPR	*TNR*	*ACC*
1-CH setup	**0.54** (0.51–0.59)	**0.80** (0.77–0.82)	**0.65** (0.60–0.68)
5-CH setup	**0.72** (0.70–0.76)	**0.87** (0.85–0.88)	**0.75** (0.72–0.78)
15-CH setup	**0.53** (0.51–0.58)	**0.80** (0.78–0.81)	**0.63** (0.61–0.67)
Reference method [52]	**0.56** (0.52–0.60)	**0.80** (0.79–0.82)	**0.62** (0.60–0.68)

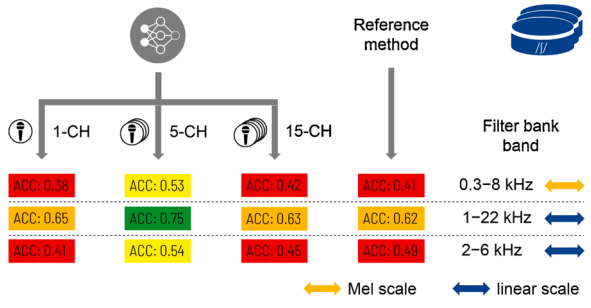


Fig. 11. Summary of the accuracy results obtained in the experiment E2 for different filter banks.

Table 5
Summary of the results obtained in the experiment E3[a]. Presentation format: median (25th–75th percentile).

	TPR	*TNR*	*ACC*
1-CH setup	**0.36** (0.32–0.38)	**0.89** (0.88–0.89)	**0.44** (0.41–0.45)
5-CH setup	**0.56** (0.53–0.57)	**0.92** (0.92–0.92)	**0.61** (0.60–0.62)
15-CH setup	**0.30** (0.29–0.32)	**0.87** (0.87–0.88)	**0.39** (0.38–0.41)
Reference method [52]	**0.34** (0.33–0.36)	**0.88** (0.88–0.89)	**0.42** (0.39–0.44)

sounds) than phonetic interpretation (study of physical properties of sound). Therefore, we took the above observations into account and combined two pairs of classes to perform the other classification experiment E3[b] over four classes: $s_{norm}/\int_{den}$, s_{add}, $s_{int}/\int_{int}$, and $\int_{norm}$ (only for the 5-CH model). Tables 7 and 8show the classification results and the accumulated confusion matrix, respectively. Fig. 12 presents the results of both experiments E3[a] and E3[b] with the filter banks in a linear and Mel scale up to 22 kHz.

Table 6
Confusion matrix obtained for the 5-CH framework in the experiment E3[a]. Matrix cells are accumulated over 10 experiments. Output classes in rows, target classes in columns. Class-wise recall (*TPR*) and precision (*PPV*) values are given to the right and at the bottom of the matrix, respectively.

	s_{norm}	s_{add}	s_{int}	$\int_{norm}$	$\int_{int}$	$\int_{den}$	*TPR*
s_{norm}	**414**	47	59	19	1	120	0.63
s_{add}	69	**159**	8	8	2	34	0.57
s_{int}	36	5	**374**	13	57	34	0.72
$\int_{norm}$	10	5	21	**551**	46	62	0.79
$\int_{int}$	0	1	95	50	**78**	13	0.33
$\int_{den}$	88	12	43	92	24	**101**	0.28
PPV	0.67	0.69	0.62	0.75	0.38	0.28	

Table 7
Summary of the results obtained in the experiment E3[b]. Presentation format: median (25th–75th percentile).

	TPR	*TNR*	*ACC*
5-CH setup	**0.75** (0.72–0.76)	**0.91** (0.91–0.92)	**0.75** (0.74–0.78)

Table 8
Confusion matrix obtained for the 5-CH framework in the experiment E3[b]. Matrix cells are accumulated over 10 experiments. Output classes in rows, target classes in columns. Class-wise recall (*TPR*) and precision (*PPV*) values are given to the right and at the bottom of the matrix, respectively.

	$s_{norm}/\int_{den}$	s_{add}	$s_{int}/\int_{int}$	$\int_{norm}$	*TPR*
$s_{norm}/\int_{den}$	**757**	78	125	59	0.74
s_{add}	79	**178**	13	9	0.64
$s_{int}/\int_{int}$	85	11	**636**	41	0.82
$\int_{norm}$	63	7	93	**529**	0.76
PPV	0.77	0.65	0.73	0.83	

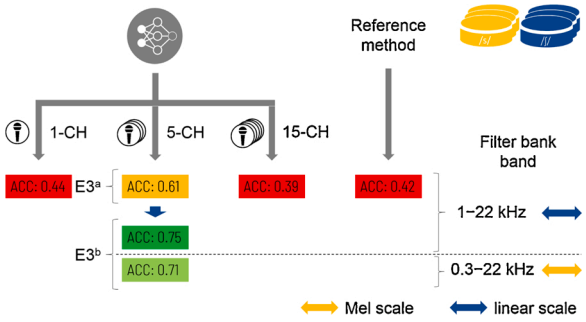


Fig. 12. Summary of the accuracy results obtained in the experiment E3 for different filter banks.

3.4. Optimization of selected hyperparameters and settings

According to the obtained results, we found the 5-CH model most efficient. To justify the CNN parameter settings and properties, we subjected this model to a deeper analysis involving experiments E1, E2, and E3[b]. We verified four parameters:

1 Kernel size in the convolutional layers (in the time-frequency projection, refer to the bottom kernel in Fig. 8). Values under investigation: 3 × 3, 5 × 5, 7 × 7. Kernel sizes in the time- and frequency domain follow the general size (e.g., 5 × 5 induces 1 × 5 and 5 × 1, respectively).
2 Model width (number of kernels in five subsequent convolutional layers). Values under investigation: 8-8-16-16-32, 16-16-32-32-64, 32-32-64-64-128.

3 Model depth (number of convolutional layers). Values under investigation: 4, 5, 6, 7.
4 Training optimizer. Options under investigation: SGDM (refer to Section 2.3), ADAM (adaptive model estimation; learning rate 0.00003, gradient decay factor 0.99, max. 50 epochs, 10-epoch validations patience).

Fig. 13 presents mean accuracies obtained in the experiments E1, E2, and E3^b (again, each repeated ten times) in different setups of a selected parameter. When analyzing an individual parameter, the others were set to their default values given in Section 2.3 and marked with a bold blue font in Fig. 13.

4. Discussion

The research presented in this paper is an attempt to address the computer-aided speech diagnosis and therapy by using advanced artificial intelligence tools. We developed the system in terms of both dedicated measurement equipment design and multichannel data processing workflow. All stages of the study were thoroughly consulted by the speech therapists, who are also its primary beneficiaries. The hardware was described and carefully tested in our previous work [45], proving its capability to support the speech therapist. Here, we performed several experiments on the acquired data with expert annotations to enable detection and classification of pronunciation pathology in sigmatism. The system is prepared to assist preschool children's speech therapy since the diagnosis and intervention are the most profitable in this age.

The idea to employ the convolutional neural network at the data analysis stage was investigated in multiple classification tasks involving common Polish words, containing two sibilant sounds: /s/ and /ʃ/. A relatively large number of recorded speakers, as well as various in-word surroundings of the sibilants (vowels or consonants), secure the reliability of the analysis. We designed a CNN framework analyzing a multidimensional input data structure instead of relying directly on time signals, which is quite common in the deep learning speech analysis approaches. The data acquisition device records the speech signal in 15 channels from various spatial locations in the speaker's mouth area. Thus, the amount and character of data enable the employment of deep learning techniques over multidimensional data representations. We decided to use the time-frequency feature maps obtained through filter bank energies and their derivatives. Such a choice resulting from our previous research was confirmed in this work.

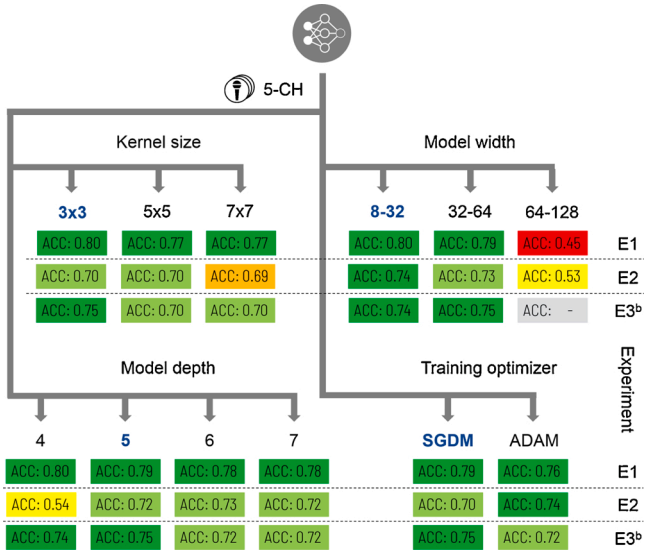


Fig. 13. Summary of the analysis of the 5-CH model parameters and settings.

The speech sample is represented by a four-dimensional acoustic volume. It consists of three 2D time-frequency feature maps extracted from a single microphone concatenated along the third dimension. In multichannel frameworks, the above structures from individual microphones (acoustic images) are assembled into a 4D acoustic volume. The analysis of such data is undoubtedly challenging and becomes applicable with the progressive development of multidimensional CNN models and procedures. Thus, we designed a multibranch CNN supplied by the acoustic volumes.

One of the main goals of the study was to investigate the contribution brought by the spatial character of speech data. Therefore, we defined three frameworks depending on the number of recording channels and data interpretation. The 1-CH model is, in fact, a conventional single-microphone system with a sensor placed directly in front of the speaker's mouth. Two other models take advantage of all 15 signals. The 15-CH framework produces the acoustic volume from 15 acoustic images, whereas the 5-CH involves the aggregation of data into five horizontally distributed uniform linear arrays. It turned out that the latter step is very profitable in handling the spatially distributed signals. The 5-CH outperforms both 1-CH and 15-CH by a significant margin in all classification experiments. The advantage of a 5-CH over a 15-CH model brings a series of conclusions and possible explanations. The ULA signal aggregation based on beamforming reduces noise from directions different from the speech source. The acoustic images in both models are arranged along one dimension. The 15-CH arrangement does not reflect the actual 2D spatial relationships between the sensors or does so to a limited extent (five neighboring vertical triplets) and may introduce unwanted interference. Finally, the ratio of the amount of data per sample and the database size becomes too large and may lead to overfitting. The 5-CH framework handles the above concerns best and is the most efficient. The horizontal energy distribution related to the lateral airflow during pathological pronunciation seems to bring distinctive information to the system, which was reported in some earlier works [45,55]. The comparison to the reference method for the speech command recognition from spectrograms also shows the advantage of our 5-CH model.

The 5-CH model's performance is comparable in the classification experiments with the accuracy of ca. 0.75. The other models' efficiency depends on the sibilant involved: they recognize the /ʃ/ realizations better than the /s/ ones. However, in all models and experiments, the linear 1–22 kHz band provides way better efficiency than any other frequency band under consideration. The conventional Mel-scale 300 Hz–8 kHz band generally fails, since it misses a lot of high-frequency spectrum particularly valuable in dentalized (noise) consonants. The bands dedicated to the particular sibilants in experiments E1 and E2 are not efficient either. Therefore, searching for distinctive features in higher frequencies seems to be profitable in dentalized sounds (sampling frequency of 16 kHz is too low). Moreover, the shifted frequency range reduces the vowel coarticulation effect that can be observed in the lower parts of the spectrum (below 2 kHz). The comparison between the Mel and linear scales in the experiment E3^b shows the latter's advantage, probably due to better resolution in the higher subbands.

Since the CNN proposed in this paper is an original contribution, we performed several experiments over its parameters and procedures. The analysis justifies the architecture choices for the 5-CH model. Apart from four parameters presented in Section 3.4, we also verified different settings of the CNN, e.g., the size of the resampled acoustic image, the number of training samples constituting a mini-batch, or the influence of some auxiliary layers. Final settings are the optimal ones according to the corresponding analysis.

The experiments prove that the multiclass analysis of sibilant realization is a challenging task. The speech database can be considered sufficient compared to similar approaches, despite the generally high requirements for deep learning. In this study, we attempted to augment the amount of data using dedicated procedures in both the time and

frequency domain. To secure the validation reliability, we took care to avoid training and testing using the same speaker's speech samples. With the data acquisition device, its flexibility, and redevelopment capabilities (e.g., by adding cameras tracking articulation organs during pronunciation), we can think of collecting a larger database or analyzing other sibilants in the future research.

Another conclusion was drawn based on the experiment E3[a] on the possible mismatch between phonological and phonetic interpretations of pronunciation. It suggests paying special attention to the preparation of speech corpus annotations performed by speech therapy experts. The diagnosis based on a broad set of observable articulation features is preferred in contemporary speech therapy practice. Not all observations relatively clear in the study of the behavior of the articulatory organ, and relationships between sounds are unambiguously reflected in acoustics. E.g., normative /s/ and dental /ʃ/ may be hard to distinguish based on the acoustic features only; putting the sound into the in-word context could increase the recognition capability. That brings into consideration the employment of speech recognition tools as a possible direction for the development of presented computer-aided speech diagnosis and therapy system. Thus far, we find the obtained classification abilities based on deep multichannel acoustic features promising for the assumed purposes.

5. Conclusion

A novel deep learning framework for computer-aided speech diagnosis and therapy in children was presented and assessed in this paper. The multichannel recording of the speech signal in arranged spatial locations provides valuable data for the pronunciation analysis. We proposed a 4D time-frequency-space acoustic volume representation to enable detection and classification of different pronunciation disorders in sigmatism. With the data already acquired and the acquisition device development possibilities, the system is ready to address multiple issues in the speech therapy domain.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

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

References

[1] D. Pluta-Wojciechowska, Peripheral Dyslalia. Diagnosis and Speech Therapy of Selected Forms of Disorders, (PL) Dyslalia obwodowa. Diagnoza i terapia logopedyczna wybranych form zaburzeń, Wydawnictwo Ergo-sum, 2019.
[2] A. Sołtys-Chmielowicz, Articulation Disorders. Theory and Practice. (PL) Zaburzenia artykulacji. Teoria i praktyka, Oficyna Wydawnicza Impuls, 2016.
[3] P. Marshalla, Frontal Lisp, Lateral Lisp: Articulation and Oral Motor Procedures for Diagnosis and Treatment Paperback, Marshalla Speech and Language, 2007.
[4] I. Anjos, N. Marques, M. Grilo, I. Guimar aes, J. Magalh aes, S. Cavaco, Sibilant consonants classification with deep neural networks, in: P. Moura Oliveira, P. Novais, L.P. Reis (Eds.), Progress in Artificial Intelligence, Springer International Publishing, Cham, 2019, pp. 435–447.
[5] J. Honova, P. Jindra, J. Pesak, Analysis of articulation of fricative praealveolar sibilant "S" in control population, Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech. Repub. 147 (2003) 239–242.
[6] W. Katz, S. Mehta, M. Wood, J. Wang, Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation, J. Acoust. Soc. Am. 141 (2017) 57–63.
[7] C. Kroos, Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500), J. Phonet. 40 (2012) 453–465.
[8] A. Lorenc, D. Król, K. Klessa, An acoustic camera approach to studying nasality in speech: the case of polish nasalized vowels, J. Acoust. Soc. Am. 144 (2018) 3603–3617.
[9] D. Król, A. Lorenc, R. Święciński, 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, 2015, pp. 5147–5151.
[10] N. Sebkhi, D. Desai, M. Islam, J. Lu, K. Wilson, M. Ghovanloo, Multimodal speech capture system for speech rehabilitation and learning, IEEE Trans. Biomed. Eng. 64 (2017) 2639–2649.
[11] L. Mik, A. Lorenc, D. Król, R. Wielgat, R. Święciński, 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. 66 (2018) 257–266.
[12] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, 1st ed., Springer, 2001.
[13] T.N. Sainath, R.J. Weiss, K.W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, C. Kim, Multichannel signal processing with deep neural networks for automatic speech recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 25 (2017) 965–979.
[14] W. Hu, Y. Qian, F.K. Soong, Y. Wang, Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers, Speech Commun. 67 (2015) 154–166.
[15] S. Wei, G. Hu, Y. Hu, R.-H. Wang, A new method for mispronunciation detection using support vector machine based on pronunciation space models, Speech Commun. 51 (2009) 896–905.
[16] Z.A. Benselam, M. Guerti, M. Bencherif, Arabic speech pathology therapy computer aided system, J. Comput. Sci. 3 (2007) 685–692.
[17] C. Valentini-Botinhao, S. Degenkolb-Weyers, A. Maier, E. Nöth, U. Eysholdt, T. Bocklet, Automatic Detection of Sigmatism in Children, International Journal of Child-Computer Interaction, 2012, pp. 1–4.
[18] A.F. Seddik, M.E. Adawy, A.I. Shahin, A computer-aided speech disorders correction system for Arabic language, International Conference on Advances in Biomedical Engineering, ICABME'13 (2013) 18–21.
[19] R. Wielgat, T. Zieliński, L. Holda, D. Król, T. Woźniak, S. Grabias, HFCC Based Pathological Speech Recognition, Advances in Quantitative Laryngology, Voice and Speech Researches, 2006.
[20] M. Zygis, M. Jaskuła, D. Pape, L.L. Koenig, Production of Polish sibilants in the process of language acquisition, Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia (2019) 1510–1514.
[21] V. Bukmaier, J. Harrington, The articulatory and acoustic characteristics of polish sibilants and their consequences for diachronic change, J. Int. Phonet. Assoc. 46 (2016) 311–329.
[22] S. Mostafa Mirhassani, T. Hua-Nong, Fuzzy-based discriminative feature representation for children's speech recognition, Digit. Signal Process. 31 (2014) 102–114.
[23] R. Wielgat, T. Zieliński, T. Woźniak, S. Grabias, D. Król, Automatic recognition of pathological phoneme production, Folia Phoniatr Logop 60 (2008) 323–331.
[24] J. Grzybowska, M. Klaczynski, Computer-assisted HFCC-based learning system for people with speech sound disorders, Annual Pacific Voice Conference, PVC'14 (2014) 1–5.
[25] Z. Miodońska, M. Kręcichwost, A. Szymańska, Computer-aided evaluation of sibilants in preschool children sigmatism diagnosis, in: Information Technologies in Medicine, ITIB'16, Springer International Publishing, Cham, 2016, pp. 367–376.
[26] W. Jassem, The formant patterns of fricative consonants, speech transmission laboratory, Q. Prog. Status Rep. 3 (1962) 6–15.
[27] J. Klesla, Acoustic Analysis of Polish Voiceless Fricatives Implemented by Deaf Children, (PL) Analiza akustyczna polskich spółgłosek tracych bezdźwięcznych realizowanych przez dzieci niesłyszące, Audiofonologia Problemy teorii i praktyki, 2004, pp. 107–118.
[28] M. Gordon, P.K.S. Barthmaier, A cross-linguistic acoustic study of voiceless fricatives, J. Int. Phonet. Assoc. (2002) 141–174.
[29] M. Zygis, S. Hamann, Perceptual and acoustic cues of polish coronal fricatives, International Congress of Phonetic Sciences, ICPhS'03 (2003) 395–398.
[30] S. Veena, N.S. Wankhede, M.S. Shah, Study of vocal tract shape estimation techniques for children, Proc. Comput. Sci. 79 (2016) 270–277.
[31] D. van Bergem, L. Pols, F. van Beinum, Perceptual normalization of the vowels of a man and a child in various contexts, Speech Commun. 7 (1988) 1–20.
[32] Z. Miodońska, M.D. Bugdol, M. Kr¸ecichwost, Dynamic time warping in phoneme modeling for fast pronunciation error detection, Comput. Biol. Med. 69 (2016) 277–285.
[33] M. Krecichwost, Z. Miodonska, P. Badura, J. Trzaskalik, N. Mocko, Multi-channel acoustic analysis of phoneme /s/mispronunciation for lateral sigmatism detection, Biocybern. Biomed. Eng. (2018).
[34] T. Purohit, A. Agrawal, V. Ramasubramanian, Acoustic scene classification using deep CNN on raw-waveform Technical Report, Technical Report, International Institute of Information Technology – Bangalore, India (2018).
[35] J. Lee, T. Kim, J. Park, J. Nam, Raw Waveform-Based Audio Classification Using Sample-Level CNN Architectures, 2017. ArXiv abs/1712.00866.
[36] W. Dai, C. Dai, S. Qu, J. Li, S. Das, Very deep convolutional neural networks for raw waveforms, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'17 (2017) 421–425.
[37] J. Chien, A. Misbullah, Deep long short-term memory networks for speech recognition, International Symposium on Chinese Spoken Language Processing, ISCSLP'16 (2016) 1–5.
[38] A. Graves, A. Mohamed, G.E. Hinton, Speech Recognition with Deep Recurrent Neural Networks, Computing Research Repository, 2013 abs/1303.5778.
[39] S. Basu, J. Chakraborty, M. Aftabuddin, Emotion recognition from speech using convolutional neural network with recurrent neural network architecture,

International Conference on Communication and Electronics Systems, ICCES'17 (2017) 333–336.
[40] A. Woloshuk, M. Kręcichwost, Z. Miodońska, D. Korona, P. Badura, Convolutional neural networks for computer aided diagnosis of interdental and rustling sigmatism, in: Information Technology in Biomedicine, ITIB'19, Springer International Publishing, Cham, 2019, pp. 179–186.
[41] A. Khamparia, D. Gupta, N.G. Nguyen, A. Khanna, B. Pandey, P. Tiwari, Sound classification using convolutional neural network and tensor deep stacking network, IEEE Access 7 (2019) 7717–7727.
[42] S. Ganapathy, V. Peddinti, 3-D CNN models for far-field multi-channel speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'18, 2018, pp. 5499–5503.
[43] S. Akhtar, F. Hussain, F.R. Raja, M. Ehatisham-ul haq, N.K. Baloch, F. Ishmanov, Y. B. Zikria, Improving mispronunciation detection of Arabic words for non-native learners using deep convolutional neural network features, Electronics 9 (2020).
[44] A. Sabzi Shahrebabaki, A.S. Imran, N. Olfati, T. Svendsen, A comparative study of deep learning techniques on frame-level speech data classification, Circuits, Syst., Signal Process. 38 (2019) 3501–3520.
[45] M. Krecichwost, Z. Miodonska, J. Trzaskalik, P. Badura, Multichannel speech acquisition and analysis for computer-aided sigmatism diagnosis in children, IEEE Access 8 (2020) 98647–98658.
[46] N.M. Kwok, J. Buchholz, G. Fang, J. Gal, Sound source localization: microphone array design and evolutionary estimation, IEEE International Conference on Industrial Technology, ICIT'05 (2005) 281–286.
[47] M. Imran, A. Hussain, N.M. Qazi, M. Sadiq, A methodology for sound source localization and tracking: development of 3d microphone array for near-field and far-field applications, 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (2016) 586–591.
[48] M. Omologo, M. Matassoni, P. Svaizer, Speech Recognition with Microphone Arrays, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 331–353.
[49] A. Woloshuk, M. Kręcichwost, Z. Miodońska, P. Badura, J. Trzaskalik, E. Pietka, CAD of Sigmatism Using Neural Networks. Information Technology in Biomedicine, ITIB'18, Springer International Publishing, Cham, 2019, pp. 260–271.
[50] M.A. Hossan, S. Memon, M.A. Gregory, A novel approach for MFCC feature extraction, International Conference on Signal Processing and Communication Systems, ICSPCS'10 (2010) 1–5.
[51] P.P. Das, S.M. Allayear, R. Amin, Z. Rahman, Bangladeshi dialect recognition using Mel frequency cepstral coefficient, delta, delta–delta and Gaussian mixture model, International Conference on Advanced Computational Intelligence, ICACI'16 (2016) 359–364.
[52] Matlab, Speech Command Recognition Using Deep Learning, https://www.mathworks.com/help/deeplearning/examples/deep-learning-speech-recognition.html, accessed: 2020-07-10.
[53] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.
[54] P.F. Reidy, Spectral dynamics of sibilant fricatives are contrastive and language specific, J. Acoust. Soc. Am. 140 (2016) 2518–2529.
[55] A. Lorenc, Normative Pronunciation of Polish Nasal Vowels and Lateral Consonants, (PL) Wymowa normatywna polskich samogłosek nosowych i spółgłoski bocznej, Elipsa, 2016.