Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Multi-channel acoustic analysis of phoneme /s/ mispronunciation for lateral sigmatism detection

Michal Krecichwost [a,*], Zuzanna Miodonska [a], Pawel Badura [a], Joanna Trzaskalik [b], Natalia Mocko [c]

[a] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
[b] Non-Resident Faculty of Jesuit University Ignatianum in Cracow, Krakow, Poland
[c] Silesia's Center of Hearing and Speech MEDINCUS, Katowice, Poland

ARTICLE INFO

Article history:
Received 31 October 2017
Received in revised form
8 August 2018
Accepted 21 November 2018
Available online 12 December 2018

Keywords:
Computer-aided speech diagnosis
Lateral sigmatism
Multi-channel speech acquisition
Automated acoustic analysis

ABSTRACT

The paper presents a method for computer-aided detection of lateral sigmatism. The aim of the study is to design an automated sigmatism diagnosis tool. For that purpose, a reference speech corpus has been collected. It contains 438 recordings of a phoneme /s/ surrounded by certain vowels with normative and simulated pathological pronunciation. The acoustic signal is recorded with an acoustic mask, which is a set of microphones organised in a semi-cylindrical surface around the subject's face. Frames containing /s/ phoneme are subjected to beamforming and feature extraction. Two different feature vectors containing, *e.g.*, Mel-frequency cepstral coefficients and fricative formants, are defined and evaluated in terms of binary classification involving support vector machines. A single-channel analysis is confronted with multi-channel processing. The experimental results show that the multi-channel speech signal processing supported by beamforming is able to increase the pathology detection capabilities in general.

1. Introduction

Sigmatism (lisping) is one of the most frequent speech pathologies in children [1–3]. It is related to misarticulation of sibilant phonemes (sibilants), which, *e.g.*, in Polish are: //, //, /AA/ [4]. Depending on the articulation pattern various types of sigmatism can be identified, e.g., interdental, lateral, nasal, strident, or palatal. In lateral lisp, a typical medial air flow is disturbed by closing the organs responsible for articulation [5]. Thus, lateral air flow occurs.

Studies on speech disorders are dominated by observation examinations, which makes their conclusions hardly objective. The speech diagnosis of sigmatism is mostly based on observation of the articulators' work. Nevertheless, it is not always possible to precisely observe and describe phenomena taking place in the oral cavity, especially in children. In such cases the diagnosis could be assisted with acoustic analysis techniques. A computer-assisted speech diagnosis tool could help in the diagnosis and therapy in several ways, e.g., indicating articulation specifics, which are difficult to observe, or creating home rehabilitation multimedia tools.

* Corresponding author.
E-mail address: michal.krecichwost@polsl.pl (M. Krecichwost).
https://doi.org/10.1016/j.bbe.2018.11.005

Acoustic studies concerning methods of pronunciation error detection are conducted in different countries. The methods are often based on popular speech analysis techniques. However, most of the proposed tools are designed for second-language learners [6–9]. According to our knowledge, only two propositions of speech analysis methods for sigmatism diagnosis assistance can be found in literature [10,11]. These projects focus on a general evaluation of the phoneme (normative/pathology), not on a detailed analysis of any sigmatism type. Solutions dedicated to lateral sigmatism patients can hardly be found. As this kind of lisp concerns a lateral air flow, acoustic signal analysis based on a single-channel acquisition may not provide sufficient amount of information [12,13]. The employment of a larger number of microphones could allow for using a wide spectrum of spatial signal processing methods, like beamforming [14]. Beamforming techniques rely on a space-time processing of signals acquired by microphone matrix and are used, i.a., in radar systems, seismology, and acoustics [15,16]. Signals from sensors spread over a matrix are subjected to a spectral analysis in terms of amplitude and phase, which determines their mutual correspondence. An example of a study on normative phonemes acoustic description based on a microphone matrix is described in [17].

The aim of this study is to design a diagnostic tool in order to objectify the lateral sigmatism diagnosis and support the therapeutic process. The main contribution of this study is the methodology for processing and classifying spatial speech signals containing phoneme /s/ recorded with a multi-channel registration device described in [18]. Feature vectors combining information recorded in side channels are extracted and employed at the classification stage. The side channel signals are obtained by means of digital beamforming techniques. Also, certain features are determined by aggregating signals acquired in different locations. This spatial analysis approach for /s/ laterality evaluation is the main novelty of the study.

The authors collected a reference database including mispronunciations simulated by adults under the supervision of a speech pathologist. The examinations were conducted over the phoneme /s/. This phoneme has been chosen as it appears early in speech evolution and is relatively easy to induce. Simulation of phoneme /s/ pronunciation with lateral lisping is not particularly difficult for the adult under the speech therapist control. Promising results obtained and described in this paper encourage to develop the data acquisition tool for children examination.

The paper is organised as follows. The methodology is introduced in Section 2 in terms of signal acquisition, processing, feature extraction, and classification. Section 3 presents the materials, experimental setup, and obtained results. The study is discussed in Section 4 in terms of its impact and future development perspectives. Section 5 concludes the paper.

2. Methods

The consecutive steps of the acoustic signal acquisition and analysis procedure are presented in Fig. 1.

2.1. *Signal acquisition*

The acquisition device (acoustic mask, Fig. 2) was designed and prototyped by authors of this paper and speech therapists specialized in sigmatism [18]. The device is adjustable to fit any head size and enables multi-channel, spatial, repeatable sound acquisition.

The arrangement of the increased number of microphones on a semi-cylindrical surface brings data from various locations. Each sensor is frontally oriented in relation to the source sound. Pathological lateral air flow occurrence can be reflected in the recordings, as changes of amplitude and lateral noise become possible to acquire [18].

2.2. *Signal analysis*

2.2.1. *Segmentation*

Manual segmentation is performed in order to extract regions of interest (sibilant sounds) from the recordings. The segmentation is conducted by a speech analysis expert and a speech therapist. The segment boundaries are marked on the spectrogram of the central microphone signal, and then applied to all other channels.

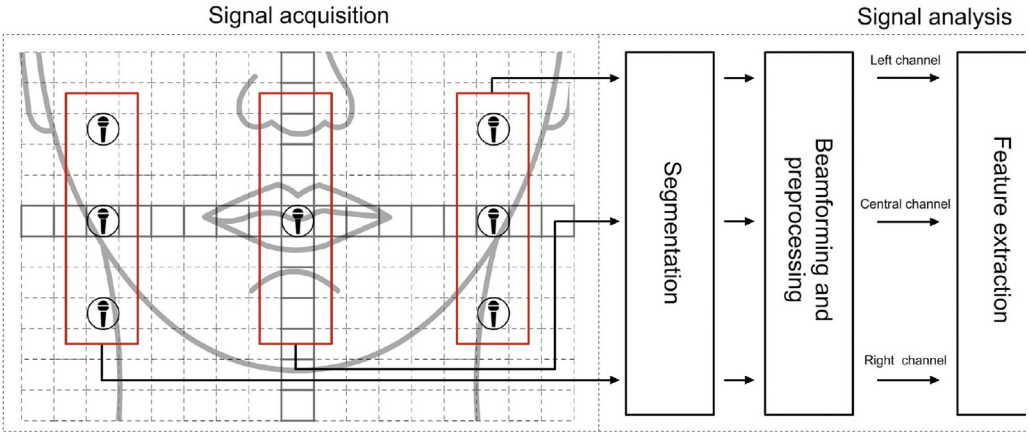


Fig. 1 – System workflow. Signals acquired by the measuring equipment (left) are subjected to acoustic analysis (right).

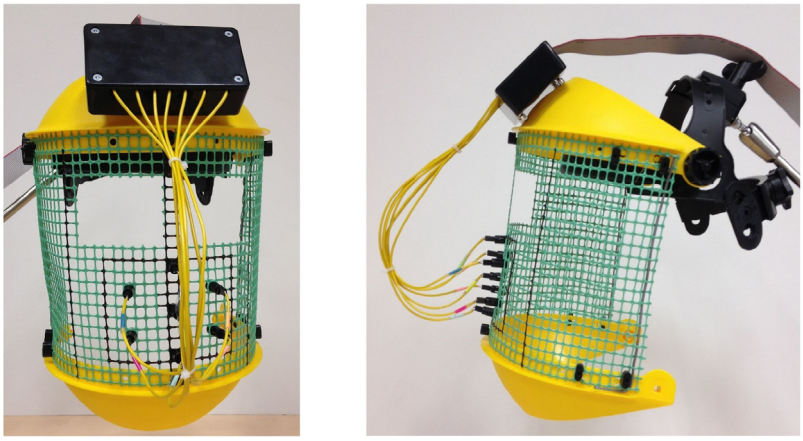


Fig. 2 – The acoustic mask – measuring device, with example configuration of adjustable microphone arrangement.

2.2.2. *Beamforming and preprocessing*

Beamforming is performed in order to aggregate data from different channels [16]. It considers the delays in different microphone sound registration. Thus, it also emphasizes the features of the side signals (left and right), e.g., noise resulting from the lateral air flow. In the presented device, microphones are placed on a semi-cylindrical surface. Therefore, they cannot be treated as one microphone array. However, all left/right microphones are collinear, so they can be analyzed as two separate uniform linear arrays (ULA) [19,20]. The central microphone is considered as an independent source of the signal. Overall, the beamforming stage aggregates all recorded data to 3 signals: central (from the central microphone), left, and right (based on partial data from lateral microphones).

The presented method employs Filter-and-Sum beamforming [21]. First, delays resulting from different angles of wave incidence on the microphone array are determined (Fig. 3). The angles are calculated using Direction-of-Arrival (DOA) estimation method [20,22], with bottom microphones used as references for linear matrices. Next, the channels are filtered with a high-pass finite impulse response (FIR) filter. The filter employed in this study is adjusted to the friction noise characteristic to the /s/ sound (high-pass, 181st-order filter with cut-off frequency 3 kHz). Finally, the signals are summed considering previously calculated delays.

The three signals obtained as a result of the beamforming stage (left, right, and central channel) are subsequently normalized to range $[-1, 1]$ using the maximal and minimal values over all channels. Then, the preemphasis filtering is performed and the Hamming window is employed to divide the signal into 25 ms frames with 10 ms overlap. Preemphasis and windowing are performed independently for the left, right, and central signal [23,24].

2.2.3. *Feature extraction*

The following features are extracted from each frame of the analyzed signal:

- 1st to 13th Mel-frequency cepstral coefficients (MFCC) [25–28],
- root mean square value (RMS),
- 1st to 3rd fricative formants (FF) and their levels (FFL) [29,30,28].

The MFCC are successfully used in speech recognition, as they reflect the ear's natural response for acoustic stimulation. The MFCC extraction procedure is presented in Fig. 4. Spectrum of the frame is mapped onto the Mel scale using a Mel filterbank. Then, the logarithms of the resulting data are calculated and discrete cosine transform (DCT) is performed, yielding the MFCC values.

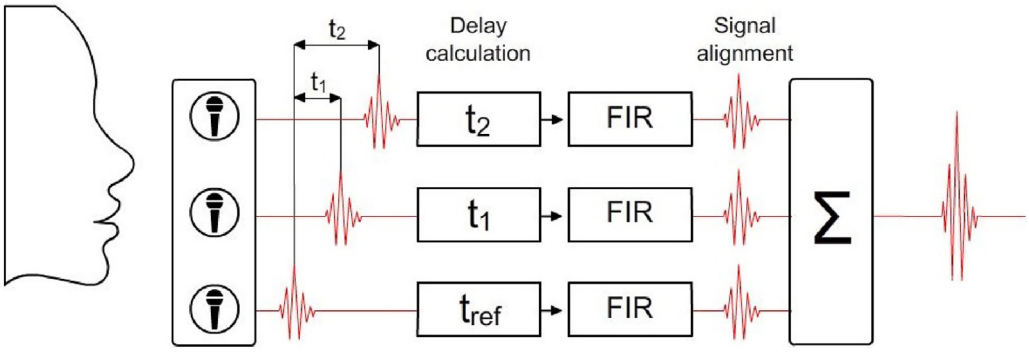


Fig. 3 – Filter-and-Sum beamforming workflow.

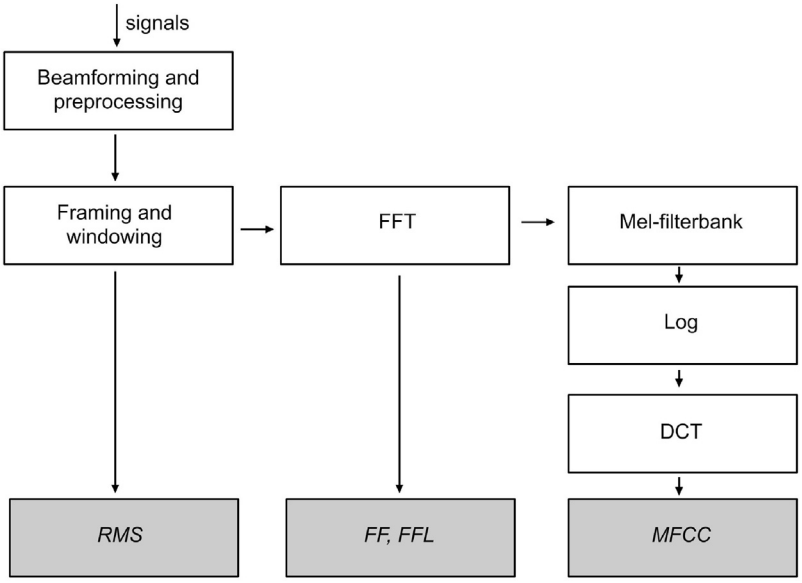


Fig. 4 – Flowchart of the feature extraction procedure.

The root mean square value is related to the energy of the signal registered in specific channels. In normative speech, RMS values for central channel are higher than for any other. During lateral air flow, side channels may produce higher values of this feature. RMS is calculated according to a formula:

$$RMS_n = \sqrt{\frac{1}{S}\sum_{i=1}^{S}|s_i|^2} \quad (1)$$

where n is the index of current frame, S is the count of samples in current frame and s_i is the current signal sample.

Fricative formants are used in classification of different sibilants as one of main distinctive features in this area [31,32]. FFs correspond to concentrations of energy in high-frequency bands of the analyzed signal. Formants are characterized by their levels and frequencies and are often employed for sibilants recognition. In this study, fricative formants are calculated according to the method previously described in [33]. First, the three highest peaks in the frame's spectrum above 3 kHz are found. Then, the obtained frequencies and levels are median-filtered within a segment to reduce the noise.

Feature values obtained for consecutive frames are aggregated in order to create a single feature set for each segment (Fig. 5). Mean values are calculated for MFCC. The third quantiles are calculated for other features. Feature extraction is performed for each channel (central, left, and right) independently.

Feature vectors for two different experiments conducted within the study are presented in Tab. 1. In the first

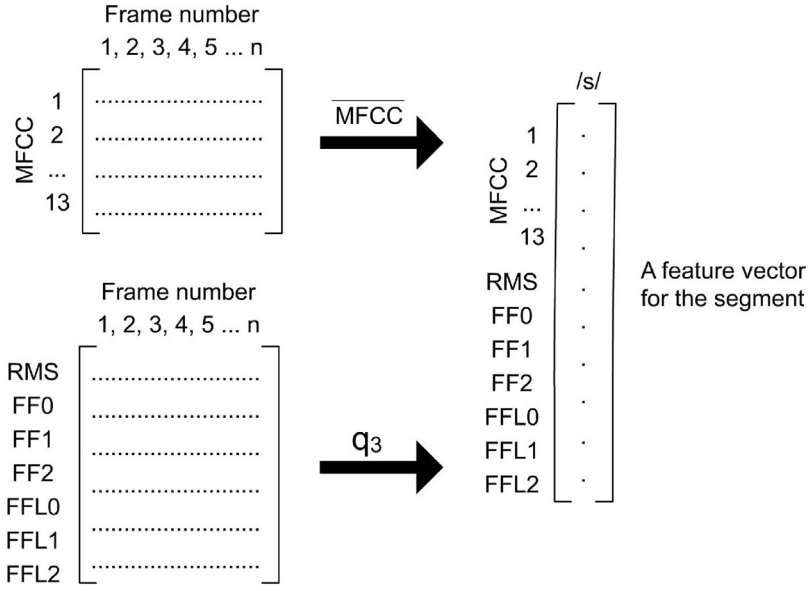


Fig. 5 – The method of features aggregation.

Table 1 – Feature vectors used in the experiments. Subscripts C, L, R denote central, left, and right channel, respectively. $q_3(\cdot)$ denotes the third quantile of a set.

	SC - single-channel	MC - multi-channel
FV1	$[\overline{MFCC}_{1,C}, \ldots, \overline{MFCC}_{13,C}]$	$[\overline{MFCC}_{1,C}, \ldots, \overline{MFCC}_{13,C}, \overline{MFCC}_{1,L}, \ldots, \overline{MFCC}_{13,L}, \overline{MFCC}_{1,R}, \ldots, \overline{MFCC}_{13,R}]$
FV2	$[\overline{MFCC}_{1,C}, \ldots, \overline{MFCC}_{13,C}, q_3(RMS_C), q_3(FF0_C), q_3(FF1_C), q_3(FF2_C), q_3(FFL0_C), q_3(FFL1_C), q_3(FFL2_C)]$	$[\overline{MFCC}_{1,C}, \ldots, \overline{MFCC}_{13,C}, \overline{MFCC}_{1,L}, \ldots, \overline{MFCC}_{13,L}, \overline{MFCC}_{1,R}, \ldots, \overline{MFCC}_{13,R}, q_3(RMS_C), q_3(RMS_L), q_3(RMS_R), q_3(FF0_C), q_3(FF1_C), q_3(FF2_C), q_3(FFL0_C), q_3(FFL1_C), q_3(FFL2_C), q_3(FFL0_L), q_3(FFL1_L), q_3(FFL2_L), q_3(FFL0_R), q_3(FFL1_R), q_3(FFL2_R)]$

experiment, the classification is conducted using average values of MFCC only. Therefore, the feature vector for single-channel (SC) case contains thirteen MFCC for the central channel, and the feature vector for multi-channel (MC) case contains MFCC values for all three channels (central, left, and right). In the second experiment the feature vectors are extended with aggregated RMS values for the central channel, RMS ratios for left and right channel, frequencies of fricative formants, and levels of fricative formants (for central channel in SC case, and for all the channels in MC case). Frequencies of fricative formants for the central channel are employed in both SC and MC analysis, as the formants' frequencies are similar for all the channels.

2.2.4. *Classification*

The binary classification (pathology/norm) is performed using support vector machine (SVM) with linear kernel function [34,35]. SVM divides the problem space into two subspaces by determining a hyperplane separating training samples. The classifier is supplied with features selected from the set described in previous section.

The normalized regularization coefficient C was initially set to 1.0 and then investigated by means of receiver operating characteristic (ROC) analysis. Sequential Minimal Optimization (SMO) was used to find the separating hyperplane [36]. The value for the Karush-Kuhn-Tucker (KKT) conditions for the SMO training method was set to 0.6, meaning 60% of the variables are allowed to violate the KKT conditions. Before training, the data was standardized by centering the data points at their mean and scaling them to have unit standard deviation.

3. Experiments and results

3.1. *Materials*

A speech corpus was recorded for the needs of the study. The speech material was designed by speech pathologists and consisted of pseudowords containing phoneme /s/ surrounded by vowels (ASA, ESE, ISI, OSO, USU, YSY). A similar vocalic environment before and after the phoneme reduces the coarticulation impact on the analyzed phoneme. The speakers were asked to pronounce the pseudowords in two ways: (1) properly, according to Polish phonetic norm, and (2) simulating lateral sigmatism with front occlusion caused by touching foreteeth with the tip of the tongue (Fig. 6). The speaker group consisted of speech therapists, experienced in sigmatism diagnosis and therapy, as well as adults with proper pronunciation who were instructed how to perform the simulation. In total, 438 pseudowords pronounced by 7 subjects (3 male and 4 female) were included in the speech corpus.

Audio recordings were acquired at the sampling rate of 44.1 kHz and the resolution of 16 bit. The registration was performed in a quiet room with no audible noise sources. At each time, the acoustic mask was stabilized on the subject's head with fastening straps to calibrate the central microphone with the subject's philtrum.

3.2. *Experiment setup*

The aim of the experiments was to compare the efficiency of pronunciation pathology detection for single-channel (SC) and multi-channel (MC) analysis. A total of $7 \times 2 \times 2 = 28$ separate classification experiments have been conducted according to the following system settings and parameters:

- six experiments for each pseudoword separately and one additional experiment over the entire database (6 + 1 cases);
- a single microphone track vs. multi-channel (2 cases);

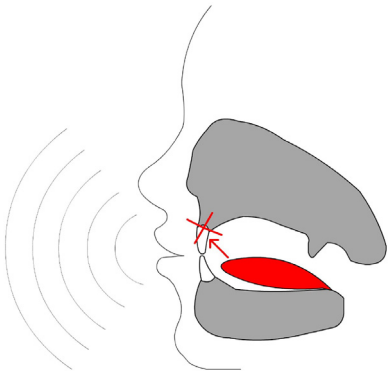


Fig. 6 – Position of the tongue during sigmatism simulation.

- two different feature vectors as defined in Section 2.2.3 (two cases).

In order to secure the evaluation reliability by means of training and testing sets independence, the experiments involved a leave-subject-out 7-fold cross validation scheme [37,38]. The database was divided into 7 groups, each containing recordings of a single subject. Each experiment yielded the number of true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) pathology detections. On their basis, classification accuracy measures were determined:

- sensitivity:

$$TPR = \frac{TP}{TP + FN}, \tag{2}$$

- specificity:

$$SPC = \frac{TN}{TN + FP}, \tag{3}$$

- accuracy:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}. \tag{4}$$

- precision:

$$PPV = \frac{TP}{TP + FP}, \tag{5}$$

- F1 score:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}. \tag{6}$$

Fig. 7 presents the sensitivity, specificity, accuracy, precision and F1 score values obtained over the entire audio dataset of 438 recordings using both feature vectors (*FV1*, top, and *FV2*, bottom) in a single microphone mode (SC) and with the Filter-and-Sum beamforming (MC), with the normalized SVM regularization coefficient $C = 0.11$ and 4th order polynomial kernel. A ROC analysis has been performed over the entire dataset ('all') with the normalized SVM regularization coefficient C used as variable parameter. Fig. 8 presents the ROC curves for different feature vectors and single- and multi-channel acquisition, each with three different SVM kernel functions, whereas Table 2 presents the corresponding area under ROC curve (AUC) values. The algorithm settings used in the experiment illustrated in Fig. 7 were established based on the ROC analysis.

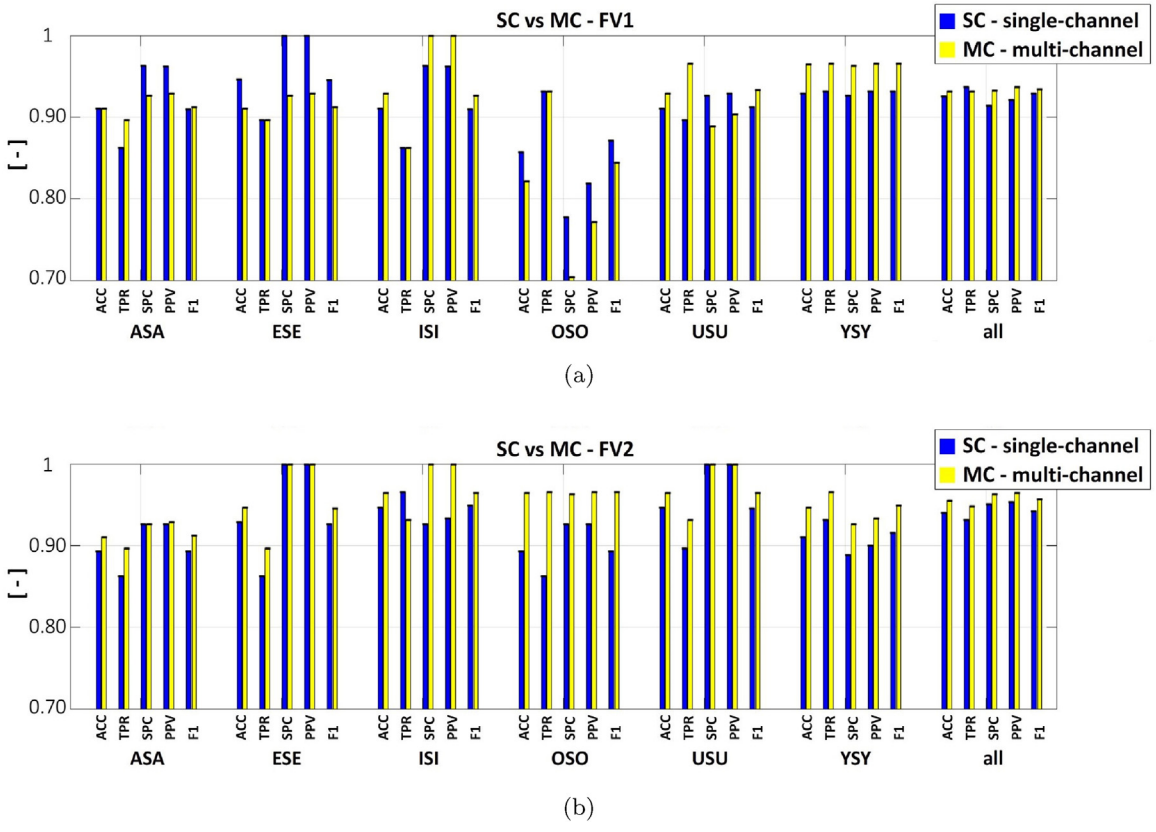
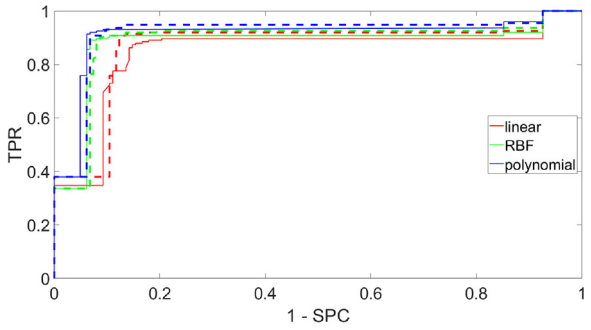


Fig. 7 – A single-channel vs. multi-channel classification efficiency in the leave-subject-out 7-fold cross validation experiment in various feature vectors: *FV1* (a) and *FV2* (b) with the normalized SVM regularization coefficient $C = 0.11$ and 4th order polynomial kernel. Paired bars reflect the corresponding single-channel vs. multi-channel experiment.



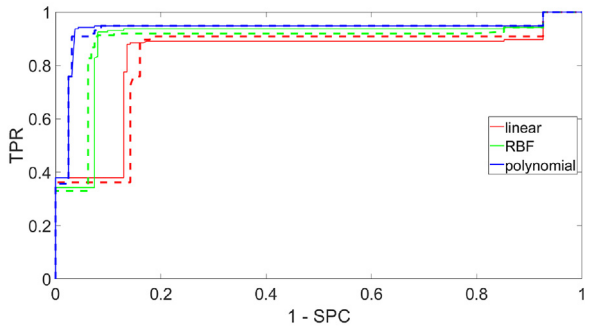


Fig. 8 – ROC curves for *FV1* (a) and *FV2* (b) over the single channel (dashed line) and multi-channel (solid line) signals in the leave-subject-out 7-fold cross validation experiment ('all' pseudowords). In each case three pairs of curves refer to three different SVM kernel functions. The normalized SVM regularization coefficient *C* was used as the ROC parameter.

Table 2 – The AUC values for different settings corresponding to Fig. 8 in the leave-subject-out 7-fold cross validation experiment.

Feature vector		*FV1*		*FV2*	
Single/multi-channel		*SC*	*MC*	*SC*	*MC*
SVM kernel	Linear	0.87	0.85	0.83	0.83
	RBF	0.89	0.88	0.89	0.90
	Polynomial	0.91	0.91	0.93	0.94

In order to additionally validate the system, another experiment was performed involving 7 speakers other than in the original database. They were recorded according to the acquisition protocol described in Section 3.1, yielding another set of pseudowords of two classes. The classifier was trained using the original database and validated using the new one. Results obtained over the entire dataset of pseudowords ('all') are presented in Fig. 9 with the corresponding AUC values gathered in Table 3.

4. Discussion

Speech pathology detection investigated in this paper can be summarized and discussed in several aspects. First of all, the employment of typically used features (MFCC) extracted from the central audio channel provides classification accuracy comparable to the results reported earlier in the literature [33,11,10]. In order to address the speaker independency and detection repeatability issues we employed the leave-subject-out 7-fold cross validation as well as additional testing involving independent speakers.

Digital beamforming used for spatial aggregation of audio signal is able to emphasize important indicators of pathological realizations of the sibilant /s/. The spatial feature values are related to noise patterns acquirable in side areas due to dismedial air flow. For example, inter-microphone energy ratios improved the efficiency of pathology detection [18]. The high-pass filtering used as a part of beamforming allows for selection of a narrow frequency band with patterns specific for a chosen sibilant [39].

The main goal of the study was to verify whether the use of additional audio signals acquired from multiple locations around the patient mouth can improve mispronunciation detection accuracy. Results obtained with the use of Filter-and-Sum beamforming and proposed feature vectors confirm the above assumption in general.

The feature vector extended by *RMS* and fricative formants (*FV2*) provides generally higher classification accuracy scores. Fig. 7b indicates the advantage of the multi-channel signal processing over the single-channel system. Fig. 8 and Table 2 prove that the extended vector (*FV2*) of features extracted from the multi-channel signal and passed to the SVM with polynomial kernel outperforms the other approaches. The additional experiment involving independent speakers (Fig. 9 and Table 3) support the above conclusions. Note, that proposed definition of features aggregated throughout each time segment (e.g. averaging the frame-wise MFCC or selection of the third quantile for the remaining features) fixes the feature space dimensionality and makes the analysis more flexible and independent of the segment duration.

5. Conclusion

The paper describes a method for processing of acoustic signals recorded by the multi-channel acquisition system. The aim was to reliably detect mispronunciation of a phoneme /s/ for computer-aided diagnosis of lateral sigmatism. The methodology involving spatial signal analysis, Filter-and-Sum beamforming, and proposed segment-wise feature vectors has proven its detection capabilities outperforming the single-channel-based analysis over a database of normative and simulated pathological /s/ realizations. The obtained results set some possibilities of the future research for the development of system for automatic speech diagnosis and therapy support: (a) employment of other sibilants in the detection process, (b) collection of a large reference database of recordings involving speech therapy patients, (c) further development of the spatial, multi-channel acquisition device,

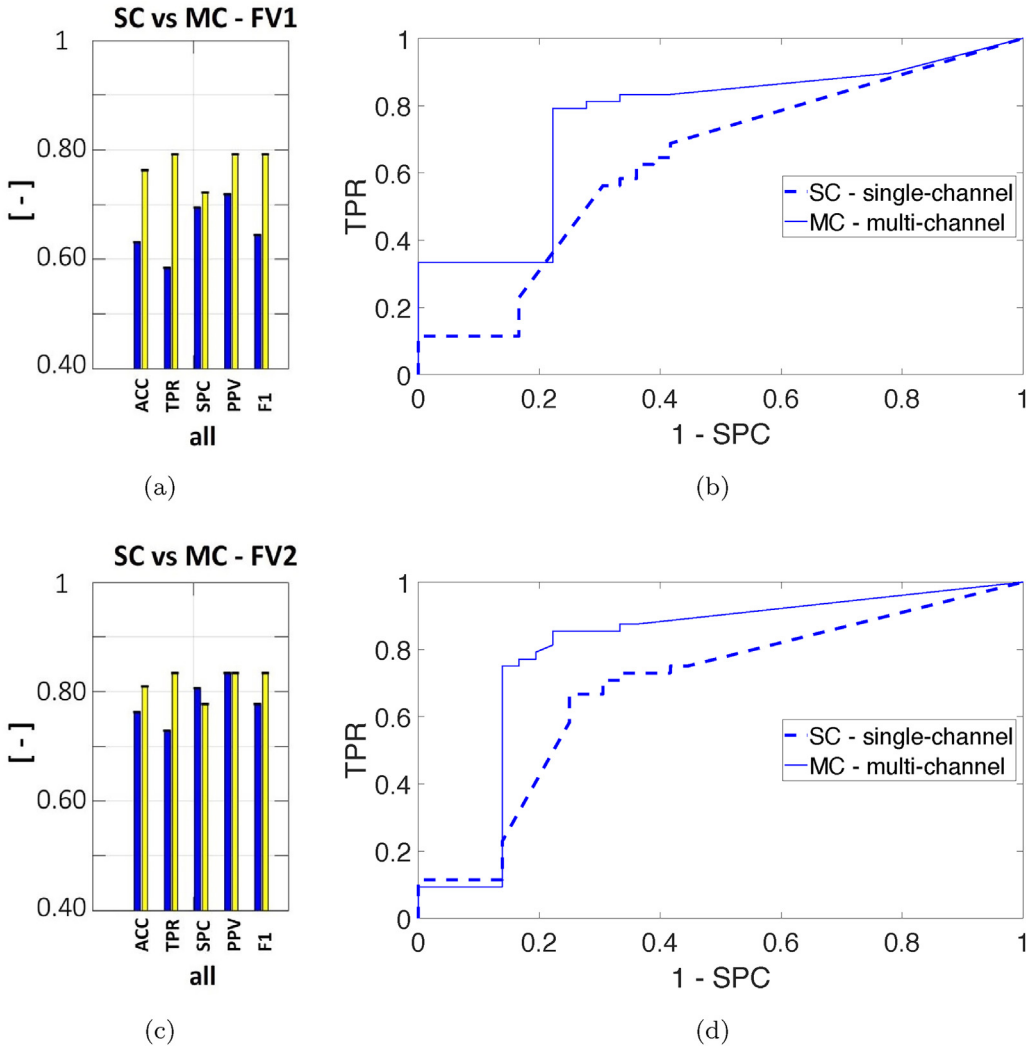


Fig. 9 – A single-channel (blue) vs. multi-channel (yellow) classification efficiency comparison ('all' pseudowords) in the validation experiment involving independent speakers in various feature vectors: *FV1* (a) and *FV2* (c) with the normalized SVM regularization coefficient *C* = 0.11 and 4th order polynomial kernel. Paired bars reflect the corresponding single-channel vs. multi-channel experiment. ROC curves for *FV1* (b) and *FV2* (d) plotted for the polynomial SVM kernel over the single channel (dashed line) and multi-channel (solid line) signals. The normalized SVM regularization coefficient *C* was used as the ROC parameter.

Table 3 – The AUC values for different settings corresponding to Fig. 9 in the validation experiment involving independent speakers.

Feature vector	FV1		FV2	
Single/multi-channel	SC	MC	SC	MC
Polynomial SVM kernel	0.63	0.76	0.69	0.80

(d) definition and analysis of new hybrid features merging acoustic and articulation aspects of speech generation.

Authors' Contribution

Michal Krecichwost: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review and Editing, Visualization. Zuzanna Miodonska: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft. Pawel Badura: Writing - Original Draft, Writing - Review and Editing, Validation, Supervision. Joanna Trzaskalik: Conceptualization, Resources. Natalia Mocko: Validation, Writing - Review and Editing.

Acknowledgements

This research was supported by the Polish Ministry of Science and Silesian University of Technology statutory financial support No. BK-209/RIB1/2018. The authors would like to thank Mr. Andre Woloshuk for his English language corrections.

REFERENCES

[1] Bilibajkic R, Saric Z, Jovicic ST, Punisic S, Subotic M. Automatic detection of stridence in speech using the auditory model. Comput Speech Lang 2016;36:122–35. http://dx.doi.org/10.1016/j.csl.2015.08.006

[2] Irwin JV. Distribution and production characteristics of /s/in the vocabulary and spontaneous speech of children. Speech Lang 1982;7:217–35. http://dx.doi.org/10.1016/B978-0-12-608607-2.50013-3

[3] Borsel JV, Rentergem SV, Verhaeghe L. The prevalence of lisping in young adults. J Commun Disorders 2007;40(6):493–502. http://dx.doi.org/10.1016/j.jcomdis.2006.12.001

[4] Lobacz P, Dobrzanska K. Acoustic description of sybilant phones in pronunciation of pre-school children, (PL) Opis akustyczny glosek sybilantnych w wymowie dzieci przedszkolnych. Audiofonologia 1999;15:7–26.

[5] Trzaskalik J. Sigmatismus lateralis in Polish logopedics literature. Theoretical considerations, (PL) Seplenienie boczne w polskiej literaturze logopedycznej. Rozważania teoretyczne. Forum Logopedyczne 2016;24:33–46.

[6] Hu W, Qian Y, Song FK, Wang Y. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Commun 2015;67:154–66. http://dx.doi.org/10.1016/j.specom.2014.12.008

[7] Su P-H, Wu C-H, Lee L-S. A recursive dialogue game for personalized computer-aided pronunciation training, Audio, Speech, and Language Processing. IEEE/ACM Trans 2015;23(1):127–41. http://dx.doi.org/10.1109/TASLP.2014.2375572

[8] Wang H, Qian X, Meng H. Phonological modeling of mispronunciation gradations in L2 English speech of L1 Chinese learners. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 7714–8. http://dx.doi.org/10.1109/ICASSP.2014.6855101

[9] 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 2009;51(10):896–905. http://dx.doi.org/10.1016/j.specom.2009.03.004. spoken Language Technology for Education

[10] Valentini-Botinhao C, Degenkolb-Weyers S, Maier A, Noeth E, Eysholdt U, Bocklet T. Automatic detection of sigmatism in children. Proc. WOCCI 2012 - Workshop on Child; 2012. pp. 1–4.

[11] Benselam Z, Guerti M, Bencherif M. Arabic speech pathology therapy computer aided system. J Comput Sci 2007;3(9):685–92.

[12] Skorek E. Faces of Speech Impediments. Warszawa: (PL) Oblicza wad wymowy; 2001.

[13] Ostapiuk B. Dyslalia. About speech quality testing in speech therapy, (PL) Dyslalia. O badaniu jakości wymowy w logopedii. Wydawnictwo Naukowe Uniwersytetu Szczecińskiego; 2013.

[14] Brandstein M, Ward D. Microphone Arrays. Signal Processing Techniques and Applications. Springer-Verlag Berlin Heidelberg; 2001.

[15] Benesty J, Sondhi M, Huang Y. Springer Handbook of Speech Processing. Springer; 2008.

[16] Benesty J, Chen J, Huang Y. Microphone Array Signal Processing, Springer Topics in Signal Processing. Springer Berlin Heidelberg; 2008.

[17] Krol D, Lorenc A, Swiecinski R. Detecting laterality and nasality in speech with the use of a multi-channel recorder. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5147–51. http://dx.doi.org/10.1109/ICASSP.2015.7178952

[18] Krecichwost M, Miodonska Z, Trzaskalik J, Pyttel J, Spinczyk D. Acoustic Mask for Air Flow Distribution Analysis in Speech Therapy. Cham: Springer International Publishing; 2016. p. 377–87. http://dx.doi.org/10.1007/978-3-319-39796-2_31

[19] Salvati D, Drioli C, Foresti GL. On the use of machine learning in microphone array beamforming for far-field sound source localization. 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). 2016. pp. 1–6. http://dx.doi.org/10.1109/MLSP.2016.7738899

[20] Pasha S, Ritz C. Informed source location and DOA estimation using acoustic room impulse response parameters. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). 2015. pp. 139–44. http://dx.doi.org/10.1109/ISSPIT.2015.7394316

[21] Argentieri S, Danes P, Soueres P. Prototyping filter-sum beamformers for sound source localization in mobile robotics. Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on; 2005. p. 3551–6. http://dx.doi.org/10.1109/ROBOT.2005.1570660

[22] Ren W-J, Hu D-H, Ding C-B. An improved method to sort and pair TDOA based on the correlation between TDOAs. Radar (Radar), 2011 IEEE CIE International Conference on, Vol. 2. 2011. pp. 1041–4. http://dx.doi.org/10.1109/CIE-Radar.2011.6159730

[23] Rabiner L, Schafer R. Theory and Applications of Digital Speech Processing. 1st Edition. Upper Saddle River, NJ, USA: Prentice Hall Press; 2010.

[24] Oppenheim AV. Digital signal processing. Englewood Cliffs, N.J: Prentice-Hall; 1975.

[25] Rabiner L, Juang B-H. Fundamentals of Speech Recognition. Upper Saddle River, NJ, USA: Prentice-Hall, Inc; 1993.

[26] Sahidullah M, Saha G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun 2012;54(4):543–65.

[27] Koolagudi SG, Rastogi D, Rao KS. Identification of language using mel-frequency cepstral coefficients (MFCC). Procedia Engineering 2012;38:3391–8. International Conference on Modelling, Optimization and Computing.

[28] Miodonska Z, Bugdol MD, Krecichwost M. Dynamic time warping in phoneme modeling for fast pronunciation error detection. Comput Biol Med. doi:10.1016/j.compbiomed.2015.12.004.

[29] Nowak PM. The role of vowel transitions and frication noise in the perception of Polish sibilants. J Phonetics 2006;34 (2):139–52. http://dx.doi.org/10.1016/j.wocn.2005.03.001

[30] Haley KL, Seelinger E, Mandulak KC, Zajac DJ. Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach. J Phonetics 2010;38(4):548–54. http://dx.doi.org/10.1016/j.wocn.2010.07.006

[31] Zygis M, Hamann S. Perceptual and acoustic cues of Polish coronal fricatives. Proc. 15th ICPhS. 2003. pp. 395–8.

[32] Gordon M, Barthmaier P, Sands K. A cross-linguistic acoustic study of voiceless fricatives. J Int Phonetic Assoc 2002;141-174.

[33] Miodonska Z, Krecichwost M, Szymanska A. Computer-Aided Evaluation of Sibilants in Preschool Children Sigmatism Diagnosis. Cham: Springer International Publishing; 2016. p. 367–76. http://dx.doi.org/10.1007/978-3-319-39796-2_30

[34] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2(2):121–67. http://dx.doi.org/10.1023/A:1009715923555
[35] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995. pp. 273–97.
[36] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep. April 1998. https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/.
[37] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surveys 2010;4:40–79. http://dx.doi.org/10.1214/09-SS054
[38] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2; 1995. p. 1137–43.
[39] Jassem W. The formant patterns of fricative consonants. STL-QPSR 1962;3(3):6–15.