Acoustic Field Distribution in Speech with the use of the Microphone Array

Daniel Król [a],*, Anita Lorenc

[a] Department of Computer Sciences, State Higher Vocational School, Tarnów, Poland
[b] Department of Speech Therapy and Applied Linguistics, Maria Curie-Skłodowska University, Sowińskiego 17, 20-040 Lublin
*Corresponding author: dankrol@gmail.com

Abstract

This article presents a 16-channel microphone-array recorder/processor that allows for a simultaneous and non-invasive detection of oral, oronasal and nasal segments in speech. Such devices and methods have not been used in the research on the articulation of sounds in the world's languages. In this paper analysis of Polish nasal vowel was presented. Adaptive beamforming method used for rendering three-dimensional acoustic fields of the recorded audio data has been shown.

Key words: microphone array, 3D acoustic field distribution, adaptive beam-forming, oral, oronasal and nasal articulations

Introduction

There are a plenty of instrumental techniques dedicated to examine nasality in speech. You can get familiar with them not only from phonetics and phonology publications [1], but also those from engineering and technical fields as well as clinical ones [2].

R.A. Krakow and M.K. Huffman [1] divide techniques of assessing nasality in speech for three groups depending on: 1) the source of velopharyngeal movements, 2) their characteristics and 3) the effects of such movements.

The examination on assessment of the source of velopharyngeal movements refer to muscles engaged in the control of work of soft palate. One of techniques is electromyography (EMG), in which electrical activity connected to muscle cramps is measured [3].

The investigation of characteristics of velopharyngeal movements is made through techniques of illustrating and tracing the outlines. There are a few techniques of illustrating the area of cramp and throat among which are: fiberoptic endoscopy [4], where during the investigation an endoscope with a camera registering the dynamics of work of structures observed is introduced into nasal cavity, radiography (in the tradition of Polish phonetic studies the most important publications are those of H. Koneczna and W. Zawadowski [5] as well as B. Wierzchowska [6]); currently the technique of radiography is used in CAT scans [7], nuclear magnetic resonance (MRI) [8] and ultrasonography [9]. The outlines of speech organs are created thanks to extraction of dynamic data on the basis of localization of previously marked points. Thanks to use of velotrace device one can monitor in mechanical way and get analog record of position of soft palate [10]. In the examination using X-ray microbeam [11] or electromagnetic articulography [12] special sensors are attached to the area of soft palate or tongue. Nasometer uses the technique of photodetection during the examination [13].

The last group of examination methods refers to the effects of velopharyngeal movements. They give aerodynamical effects as well as acoustic ones. Aerodynamical signals can be assessed thanks to masks that enable to estimate a medium value of the flow of air from the nose [14, 15]. Acoustic signals are registered via microphones. In TONAR-based [15] nasometers, acoustic signal is registered parallely through two miscrophones – oral and nasal [17]. During examination there is a special hoop inserted on the head of speaker, which divides the oral and nasal canals. There are also two separate microphones attached to it. Currently the most modern way of registering acoustic signal and analysis of its spacial area during speaking is examination with microphone matrix using beamforming technique [18, 19]. In the contrary to above mentioned ways of examination, this one is uninvasive teqchinque (e.g. endoscopes, myography, nasometers, Rothenberg mask, ultrasonography), not using direct contact with examined organ – soft palate (e.g. EMA, Velotrace, myography), not requiring huge financial resources (see MRI, X-ray microbeam) and is based on natural position of examined person during speech (see MRI).

The article presents the possibilities of using the technique of adaptive beamforming in the examination and analysis of oral, oronasal and nasal articulation.

Method

Analysis was carried out with the use of a 16-channel microphone-array recorder specially designed and built for the purposes of the experiment. An analysis of the spatial distribution of the acoustic fields in the recordings with the beamforming technique allowed for rendering three-dimensional acoustic fields in the articulation of the nasal vowels. The 16-channel circular microphone-array and recorder/processor (MARP-16) has been designed and built (Fig. 1).

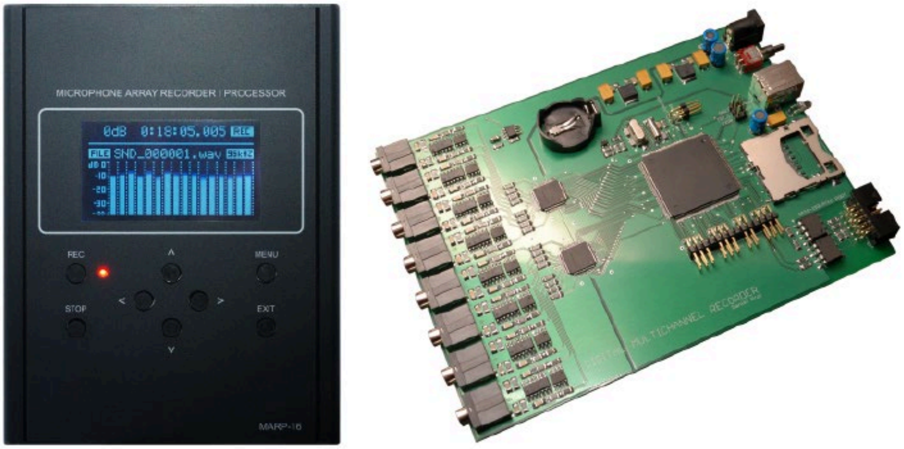
MICROPHONE ARRAY RECORDER | PROCESSOR
0dB 0:18:05.005 REC
FILE SND_000001.wav 96kHz
REC
MENU
STOP
EXIT
MARP-16
DIGITAL MULTICHANNEL RECORDER

Figure 1. Sixteen-channel microphone array recorder/processor MARP-16

In analog section was used low noise amplifiers and analog-to-digital converters with successive approximation register (SAR), dedicated to measuring equipment. The superiority of the SAR technique over the sigma-delta technology is presented in the literature [20, 21], [22]. The acquisition and pre-processing of the recorded audio data was realized by 32-bit floating point digital signal processor (DSP) with Cortex M4F core. After preprocessing stage the acquired audio data were stored on an SDHC/SDXC memory card in 16-channel WAV format. The multichannel recorder was controlled from the main computer by opto-isolated interface for minimize noise. Combining the 16-channel circular microphone array with the adaptive beamforming method used for rendering three-dimensional acoustic fields of the recorded audio data. Preliminary research results using fixed beamforming have been presented in [18, 19]. A basic problem of fixed beamforming is poor directivity factor in low frequency range and undesirable side lobes [24]. The novel rendering method using an adaptive beamforming based on LCMV (Linearly Constrained Minimum Variance) algorithm, proposed by Frost [23]. Figure 2 show a block diagram of the LCMV beamformer where the filter coefficients h0-hN-1 are adapted using a constrained version of the LMS algorithm (CLMS) [24].

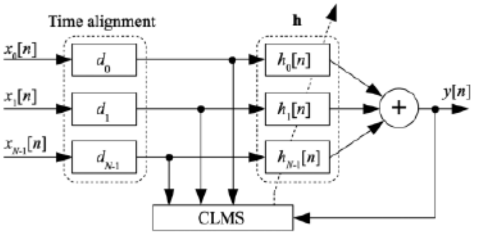

Figure 2. Block diagram of LCMV algorithm with separate time alignment block

The LCMV algorithm minimizes the noise power at the output with a constraint on the filter response in look direction. Beampattern comparing of fixed (delay-sum) and adaptive (LCMV) beamforming for a few selected frequencies have been shown in Figure 3. The geometry of microphone array were based on the size and shape of the Carsten's AG500 articulograph cube in which it was installed. The array was placed in the frontal wall of the cube, in front of the speaker's face (Fig. 4).

The circular microphone array scanned, with the use of the beamforming technique a 250×250 mm square plain with the resolution of 5mm. In result, a matrix with dimension 50×50 of the acoustic field distribution was obtained. The sampling frequency in the recorder connected to the microphone array was set to 96 kHz for increase the angular resolution of the beamsteering [25].

Speakers

One of the major assumptions of the experiment was a precise selection of speakers, who, in the opinion of a team of experts, used the careful style of the standard variety of contemporary Polish. To this end, it was made an attempt to extract the relevant criteria in diagnosing normative speech (phonetic and orthophonic criteria as well as those deriving from the theory and practice of normative linguistics, it is also necessary to take into account biological criteria: anatomical, functional and perceptual) [26]. The selection of the participants was therefore special: in accordance with established specific normative criteria mentioned above, 20 adult speakers of Polish (10 women and 10 men) were chosen out of a pre-selected group of 200 candidates.

Speech material

The realisation of variants of basic vowels of Polish language were examined in word-medial position, in the syllable of two-syllable words with accent – always in the same context. Assessed vowels were foregone with unvoiced plosive consonant [p]. Due to the possibility of realisation of Polish nasal vowels [ɔ̃] and [ɛ̃], a voiceless frica-

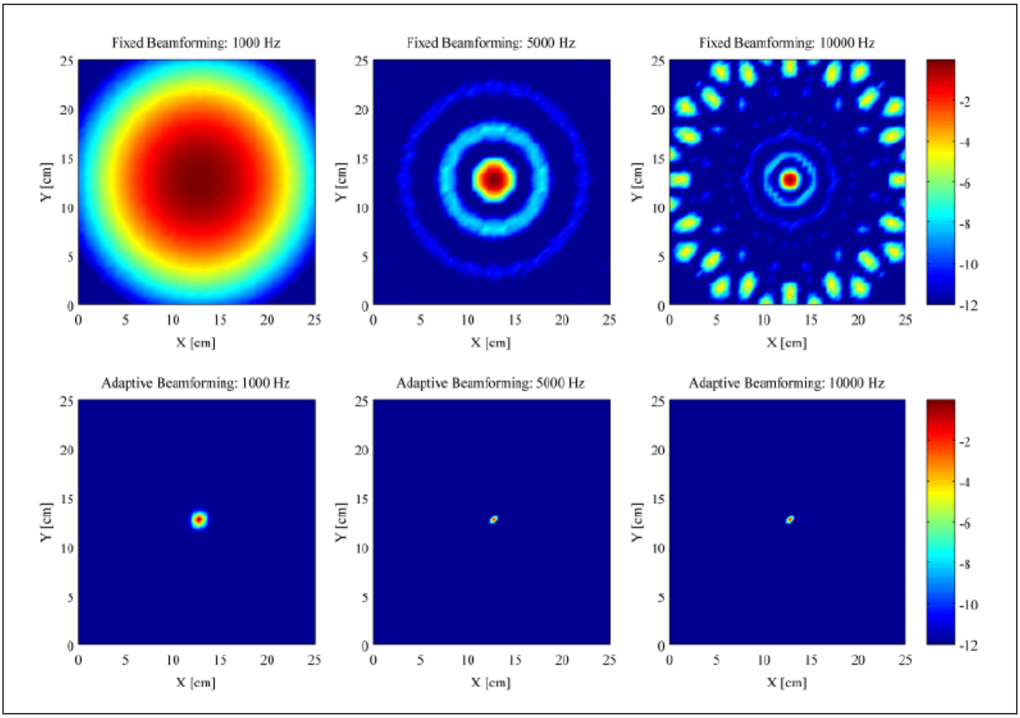

Figure 3. Beam pattern comparison of fixed and adaptive beamforming

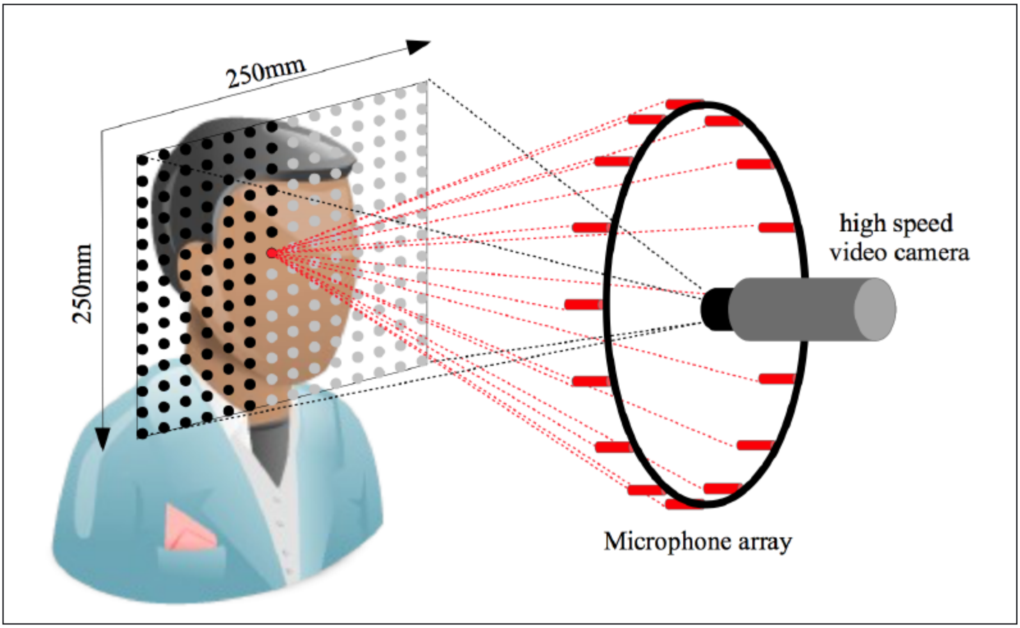

Figure 4. Scanning of the acoustic field distribution (50×50 points) by the circular microphone array

tive consonant [s] was a context proceeding after analysed vowels, therefore there were connections of the type [pˈVs], such as in the word *pąsy* ‘crimsons’ [pˈɔ̃sɨ] and *pęset* ‘tweezers’ [pˈɛ̃sɛt]. Examined vowel was expected to be realized three times by each speaker. Nasal vowels were assessed in a three-time-repeat of each word, because no other example of assumed phonetic structure was found. Examined speakers remembered word presented on the screen and were asked to speak out naturally on the agreed signal.

Results and Discussion

Used algorithms let calculate the move of the signal in time, as well as specify the frequency characteristics of analyzed segments. Thanks to that, each evaluated word got the picture of spatial distribution of acoustic field. Figures 5a, 5b and 5c enable to take a look at possibilities, that analysis with the use of acoustic camera gives.

The images illustrate the succeeding phases of nasal vowel realization [ɔ̃] in the word ‘pąsy’. First phase of articulation is realized with the use of oral energy (Fig. 5a), the second one with simultaneous resonance of oral and nasal cavity (Fig. 5b), and the last part goes exclusively in nasal cavity (Fig. 5c).

The diagram of spatial distribution of acoustic field was correlated with oscillogram, spectrogram, 3rd order polynomial approximation and histogram (Fig. 6). The position of articulograph sensor fixed above the red part of the upper lip (position 0 on the axis of spatial distribution of acoustic field top-bottom on the diagram 6) was taken as the point of division on the face of speaker. For the purpose of detection of stream of emission the threshold of decrease of acoustic pressure which amounts to 3 dB was taken.

In each signal the vertical space was analysed with the height of 61 mm and resolution of 1 mm. It was divided into three areas corresponding to different phases of articulation:

- Oral phase: acoustic signal was registered in the area starting from -30 mm to -8 mm (22 mm), the level of acoustic pressure was lower by at least 3 dB below 0 on the vertical axis (on the diagram 6 there is a section from 0 ms to around 70 ms);
- Oronasal phase (with simultaneous share of nasal and oral energy): acoustic signal was registered in the area from -7 mm to +7mm (15 mm, including 0), the difference in pressure above and below 0 was lower than 3 dB (on the diagram 6 there is a section from around 70 ms to around 140 ms and from around 250 ms to around 290 ms);
- Nasal phase: acoustic signal appeared in the area

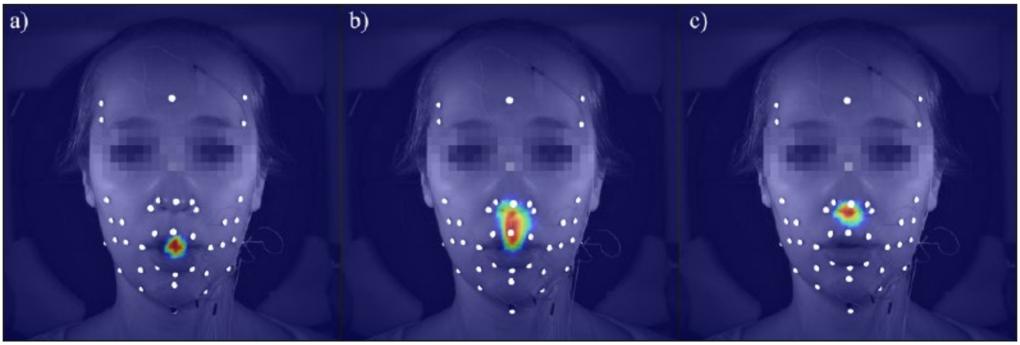

Figure 5. Spatial distribution of the acoustic field synchronised with the image from video camera during realization of nasal vowel [ɔ̃] in the word pąsy ‘crimsons’ in the different articulation phases: a) oral, b) oronasal c) nasal (speaker ZK_f, file 157)

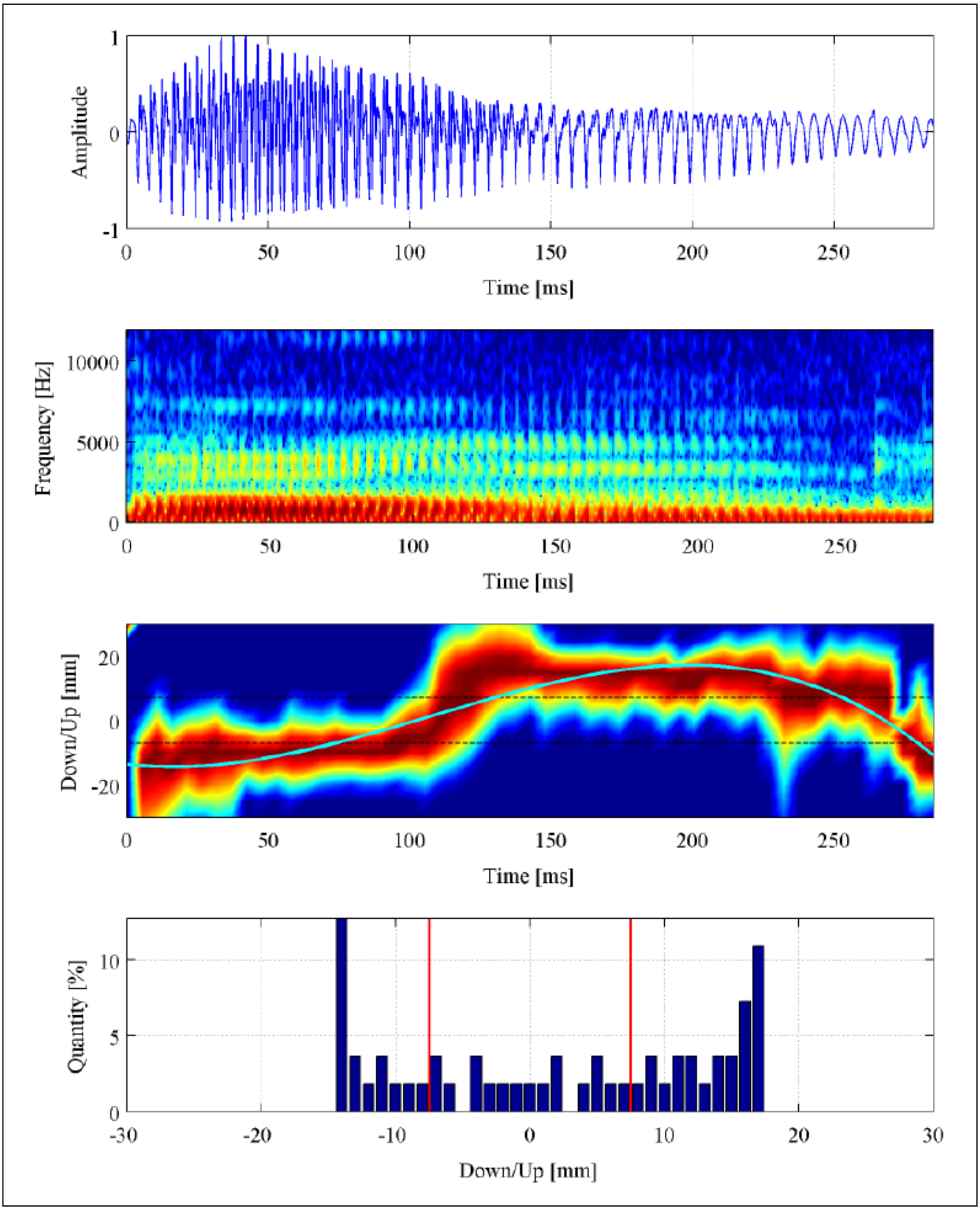

Figure 6. Spatial distribution of the acoustic field correlated with oscillogram, spectrogram, 3rd order polynomial approximation and histogram of realization of nasal vowel [ɔ̃] in the word *pąsy* 'crimsons' (speaker ZK_f, file 157)

from +8 mm to +30 mm (22 mm), the level of acoustic pressure was higher of around 3 dB above 0 on vertical axis (on the diagram 6 there is a section from 140 ms to around 250 ms and from 290 ms to 300 ms).

Presented way of detecting of sources of sound were used in interpretation of pronunciation of Polish nasal vowels in the onset of a word-medial stressed syllable before consonant [s]: [ɔ̃] in the word *pąsy* 'crimsons' and [ɛ̃] in the word *pęset* ‘tweezers’. All together 161 realizations of both vowels prepared by 20 speakers were examined. The table 1 below presents the average percentage and mean duration of distinguished phases (oral, oronasal and nasal) in the total duration of examined vowels.

Table 1. The average percentage [%] and mean duration [ms] of distinguished phases (oral, oronasal and nasal) in the total duration of examined vowels

Vowel	Phase (mean)		
	Oral	Oronasal	Nasal
[ɔ̃]	35.41 %	42.86%	21.72%
	88.17 ms	106.72 ms	54.08 ms
[ɛ̃]	65.09%	33.66%	1.25%
	148.40 ms	76.74 ms	2.85 ms

The pronunciation of the Polish nasal vowels consists of oral resonance combined with nasal resonance. The results of the analysis clearly indicate that the normative pronunciation of the Polish nasal vowels is asynchronous in that nasal resonance is delayed in relation to oral resonance (such a claim appeared in certain earlier phonetic accounts). Previous Polish instrumental phonetic studies neither described the exclusive nasal resonance phase nor provided information on the proportion of particular phases in the total duration of nasal vowels. In this light, the results presented in the article are innovative.

Conclusions and future work

A kind of novelty in the experiment described here was the use of microphone array, which allowed to examine the distribution of acoustic energy during articulation in a non-invasive way. The research presented in the study constitutes an important part of a normative description of contemporary Polish pronunciation in light of numerous controversies discussed over the last decades. The results may be applied in speech therapy; in particular, they may help diagnosticians and therapists of dyslalia. A detailed normative description will make it possible to accurately diagnose speech disorders and to precisely design treatment. The research results can be used in teaching phonetics (both general and clinical), spoken word culture at philology, pedagogy and journalism studies or at acting departments at theatre and film schools. They can also be applied in systems of automatic analysis and synthesis of speech. The future goal is a description of oral, oronasal and nasal segments in disordered speech.

Acknowledgments

Work described in this paper was supported by grant Nr 2012/05/E/HS2/03770 titled Polish Language Pronunciation. Analysis Using 3-dimensional Articulography with A. Lorenc as the principal investigator. The project is financed by The Polish National Science Centre on the basis of the decision Nr DEC-2012/05/E/HS2/03770.

References

1. R. Krakow and M. Huffman, Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory: An

introduction. In: M. Huffman and R. Krakow [eds.] Phonetics and Phonology: Nasals, Nasalization, and the Velum, Academic Press, San Diego 1993, pp. 3-59.

2. R. J. Baken and R.F. Orlikoff, Clinical measurement of speech and voice, 2nd edition, Taylor and Francis, New York, 2000.
3. F. Bell-Berti, *J. Speech. Hear. Res.*, 1976, **19**, 225-240.
4. M.P. Karnell, E.J. Seaver, and R.M. Dalston, *J. Speech. Hear. Res.*, 1988, **31**, 503-510.
5. H. Koneczna, W. Zawadowski, Przekroje rentgenograficzne głosek polskich, PWN, Warszawa, 1951.
6. B. Wierzchowska, Fonetyka i fonologia języka polskiego, Zakład Narodowy im. Ossolińskich, Wydawnictwo Polskiej Akademii Nauk Wrocław-Warszawa-Kraków-Gdańsk, 1980.
7. K.L. Moll, and R.G. Daniloff, *J. Acoust. Soc. Am.*, 1971, **50**, 678-684.
8. A. Serrurier, and P. Badin, A three-dimensional linear articulatory model of velum based on MRI data, Proceedings of the 6th Interspeech and 9th European Conference on Speech Communication and Technology, Lisboa, 2005.
9. M. Stone, T. Shawker, T. Talbot, A. Rich, *J. Acoust. Soc. Am.*, 1988, **83**, 1586-1596.
10. S. Horiguchi, and F. Bell-Bertti, *Cleft Pal. J.*, 1987, **24(2)**, 104-111.
11. O. Fujimura, J.E. Miller, S. Kiritani, A computer-controlled x-ray microbeam study of articulatory characteristics of nasal consonants in English and Japanese, Proceedings of 9th International Congress on Acoustics, Madrid, 1977.
12. K. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson, *J. Acoust. Soc. Am.*, 1992, **92**, 3078-3096.
13. J.J. Ohala, *J. Acoust. Soc. Am.*, 1971, **50(1)**, 140.
14. D. Warren, *Cleft Pal. J.*, 1967, **16**, 279-285.
15. M. Rothenberg, *J. Speech. Hear. Res.*, 1977, **20**, 155–176.
16. S.G Fletcher, and M.E. Bishop, *Cleft Pal. J.*, 1970, **7**, 610-621.
17. R.M. Dalston, D.W. Warren, and E.T. Dalston, *Cleft Pal-Craniofac. J.*, 1991, **28(2)**, 184-189.
18. D. Król, A. Lorenc, R. Święciński, 2015, Detecting Laterality and Nasality in Speech with the Use of a Multi-Channel recorder, Proceedings of 40th IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, 2015.
19. A. Lorenc, R. Święciński, D. Król, Assessment of Sound Laterality with the Use of a Multi-Channel Recorder, Proceedings of 18th International Congress of Phonetic Sciences, Glasgow 2015.
20. D. Król, Choice of analog-to-digital converters for audio measurements using MLS algorithm, Proceedings of 15th European Signal Processing Conference, Poznań, 2007.
21. D. Król, On superiority of Successive Approximation Register over Sigma Delta AD converter in standard audio measurements using Maximum Length Sequences, Proceedings of International Conference on Signals and Electronic Systems, Kraków, 2008.
22. D. Król, R. Wielgat, T. Potempa, P. Świętojański, Analysis of Ultrasonic Components in Voices of Chosen Bird Species, Proceedings of Forum Acusticum, Aalborg, 2011.
23. O. L. Frost, An algorithm for linearly constrained adaptive array processing. Proceedings of the IEEE, 1972.
24. P. Vary, R. Martin, Digital Speech Transmission: Enhancement, Coding and Error Concealment, John Wiley & Sons, 2006.
25. D. Król, Macierze mikrofonowe i głośnikowe, In: T.P. Zieliński, P. Korohoda and R. Rumian. Cyfrowe przetwarzanie sygnałów w telekomunikacji: Podstawy, multimedia, transmisja, PWN, Warszawa, 2014, pp. 665-695.
26. A. Lorenc, Diagnosis of the pronunciation norm, Logopedia 42, <http://www.logopedia.umcs.lublin.pl/images/1-278_Logop_42_ANG_ok.pdf>.