BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 66, No. 3, 2018
DOI: 10.24425/122106

# Fusing the electromagnetic articulograph, high-speed video cameras and a 16-channel microphone array for speech analysis

Ł. MIK[1]*, A. LORENC[2,3], D. KRÓL[1], R. WIELGAT[1], R. ŚWIĘCIŃSKI[4], and R. JĘDRYKA[1]

[1]Polytechnic Institute, State Higher Vocational School in Tarnow, Mickiewicza 8, 33-100 Tarnów, Poland
[2]Department of Speech Therapy and Applied Linguistics, Maria Curie-Skłodowska University, Sowińskiego 17, 20-040 Lublin, Poland
[3]Institute of Applied Polish Studies, University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland
[4]Amsterdam School of International Business, Amsterdam University of Applied Sciences, Fraijlemaborg 133, 1102CV Amsterdam, Netherlands

**Abstract.** Electromagnetic articulography (EMA) is one of the instrumental phonetic research methods used for recording and assessing articulatory movements. Usually, articulographic data are analysed together with standard audio recordings. This paper, however, demonstrates how coupling the articulograph with devices providing other types of information may be used in more advanced speech research. A novel measurement system is presented that consists of the AG 500 electromagnetic articulograph, a 16-channel microphone array with a dedicated audio recorder and a video module consisting of 3 high-speed cameras. It is argued that synchronization of all these devices allows for comparative analyses of results obtained with the three components. To complement the description of the system, the article presents innovative data analysis techniques developed by the authors as well as preliminary results of the system's accuracy.

**Key words:** electromagnetic articulography, microphone array, vision system, speech analysis.

## 1. Introduction

Methods for the analysis of articulatory movements (e.g. of the tongue and lips) have developed over many years. With the objective of capturing the behaviour of articulators during speech, researchers have adapted a number of medical techniques, such as magnetic resonance imaging (MRI) [1], ultrasonography (USG) [2] and cinefluorography [3] for this purpose. Out of these non-invasive methods, MRI seems to be gaining in popularity as it gives the possibility to obtain a 3-dimensional image of the vocal tract and does not cause dangerous ionizing radiation such as in the case of cinefluorography. This method has been applied, for instance, to determine the shape of the oral and pharyngeal cavity during extended articulation of vowels [4–6] and of the nasal cavity and surrounding structures during extended articulation of nasal consonants [7]. MRI data have been used to construct a three-dimensional articulatory model of the soft palate [8]. Also, this technique was employed in an investigation of Polish voiceless sibilants [9]. 3D models of the vocal tract obtained from MRI images of two Polish speakers were used to describe the relation between spectral formants and the shape of intraoral occlusions and cavities.

Despite the enormous technological advancement giving the possibility of three-dimensional modelling of articulatory structures, and the recent development of real-time magnetic resonance imaging techniques [10], the majority of available MRI systems provide unsatisfactory results in terms of image accuracy and the speed of sampling (speakers must extend the desired articulation for a number of seconds). Moreover, because recordings are performed with the speaker in the supine position, MRI articulatory studies are negatively affected by gravity [11]. Another disadvantage of this method is that due to loud noises and vibrations issued by the device during the test, it is difficult to obtain simultaneous audio signal of high quality.

Independently of the medical field, electromagnetic articulography (EMA) was created, which is a technique that uses a dedicated device and operates with satisfactory precision [12–18]. Electromagnetic articulography has been developed since the 1990s [11] and enables registration of spatiotemporal data from sensors placed on internal and external articulators in order to obtain information that reflects their positioning, shape and speed during utterance production. It is noteworthy to point out that simultaneous registration of EMA, audio and video data is not often encountered in the literature, as the articulograph is usually supported only by a single-channel audio recorder [13]. Occasionally, a video system involving 2 or 3 cameras is used to register the image of a speaker's face to complement articulographic data [14–18]. Such video recordings may be performed with markers placed on the face of the speaker to facilitate more precise assessment of external articulators (e.g. lips, jaw, cheeks) [14–16, 18].

Knowledge of the speech motor control obtained by means of using EMA and audio (optionally video) recordings is crucial for understanding the speech production mechanism. This, in turn, is important for speech disorder diagnosis and therapy.

Apart from difficulties posed by the tools in phonetic analysis, some articulatory phenomena constitute a challenge to the researcher. Problematic issues include *inter alia* the assessment of nasalization. The different methods of investigating nasality have been classified by Krakow and Huffman [19] into three types of techniques: (1) investigating the activity of muscles controlling velopharyngeal movements, (2) the shapes and movements themselves and (3) the effects of these movements.

*e-mail: l_mik@pwsztar.edu.pl

Manuscript submitted 2017-02-22, revised 2017-05-10 and 2017-08-07, initially accepted for publication 2017-08-10, published in June 2018.

One of the methods for assessing the source of velopharyngeal movements (1) is electromyography (EMG), which measures the electrical activity associated with muscle contractions. The signals are registered with pairs of electrodes attached to the surface of the examined muscles, but usually placed inside them [20]. Because of its invasiveness, the technique is rarely used in clinical trials.

The shapes and movements of the velopharyngeal structures (2) can be observed by means of imaging techniques and creating outlines. Such imaging methods include fibre-optic endoscopy, radiography (X-rays), magnetic resonance imaging (MRI), computed tomography (CT) and ultrasound. The above-mentioned techniques are utilised not only for the purpose of velar normative articulations but also for disordered speech such as cleft palate speech [21]. EMA sensors have also been used to register the movements of the velum [22].

Velopharyngeal movements produce both aerodynamic and acoustic effects that can be analysed (3). Aerodynamic events are usually recorded with sensors that register the pressure and flow of air. The degree of opening of the velopharyngeal port can be estimated by means of measuring oral and nasal air pressures and flows [23, 24]. Such studies of air flow are performed with the use of special masks that register air pressure fluctuations during speech [25, 26] on the basis of which average values of air flow from the nose can be estimated.

Acoustic signals are recorded with microphones. These include the nasometer (used especially in clinical assessment of velopharyngeal disorders and cleft speech; this device is based on TONAR – The Oral-Nasal Acoustic Ratio – a system developed in the 1970s by S. Fletcher [27]), wherein the acoustic signal is recorded simultaneously by two microphones – an oral and nasal one [28]. During the measurement, the speaker wears a special headset with a plate that separates the oral and nasal channel. A microphone is mounted on each side of the plate. Such separation of signals coming from the nose and mouth allows to assess nasalance, which is extremely difficult when the audio signal is collected with only one microphone.

Most methods of assessing velopharyngeal structures mentioned here have their weaknesses. The most common of them include: high cost (e.g. MRI, EMA), invasiveness (e.g. endoscopy, myography), atypical positioning during data collection (e.g. MRI scan in a horizontal position, having a mask on when speaking), or direct contact of the apparatus with the velum, which may hinder its natural movements (e.g. velotrace, myography, EMA).

With a view to obtaining more accurate speech analysis results, including those related to such problematic phenomena as nasality or laterality, an innovative system integrating EMA, 16-channel audio and high-speed video data recording was created. It acquires articulatory data from a Carstens AG500 articulograph. Video data are registered by 3 high-speed cameras that track movements of reflective markers glued to the speaker's face. Audio data are recorded by an audio recorder with a microphone array of 16 unidirectional microphones (a dedicated device developed specifically for this project). Due to the analysis of delays of the audio signal recorded by particular microphones, it is possible to obtain acoustic field intensity distribution imposed on the image from the video camera. Using this method, a relatively precise estimation of sound source intensity and localization is possible, which is highly insightful for lateral, central, oral and nasal articulation assessment [29–31]. It is also worth mentioning that this approach to speech analysis is novel and had not been applied earlier.

Fusing the electromagnetic articulograph, high speed video cameras and a 16-channel microphone array provides new opportunities for speech analysis. Examples of such analyses fostered by the system are presented below.

## 2. System description

A simplified block diagram of the whole system is presented in Fig. 1. The system is controlled by the EMArecorder computer application that sends commands to the EMA in order to

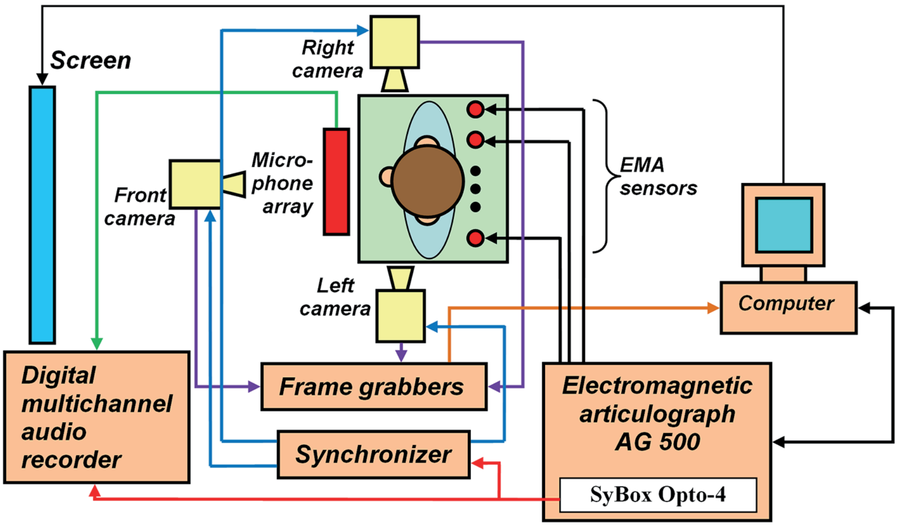


Fig. 1. Block diagram of the measurement system

generate START and STOP signals for the remaining system components. The sampling rate of EMA sensor signals being 200 Hz, the frame rate of the system cameras was also set to 200 fps. This allowed for better data quality (due to faster sampling) in comparison to other research [16], where the image acquisition rate was done at 60 fps and the sampling speed of EMA signals had to be decreased at the analysis stage.

The specifications of the component elements of the system described in this paper are summarized in Table 1.

Table 1
Specifications of the component elements of the system

| Electromagnetic Articulograph | Vision System | Multi-channel Audio Recorder |
|---|---|---|
| • Carstens Medizinelektronik AG500<br>• Sampling frequency: 200 Hz<br>• 12 channels with 3D position sensors<br>• Triger output for other devices<br>• Notebook with dedicated manufacturer software | • Three high-speed video cameras (Point Grey Gazelle GZL-CL-22C5M-C)<br>• Three frame grabbers with Camera-Link interface<br>• Three IR LED illuminators<br>• OptiTrack reflective markers on speaker's face<br>• Programmable synchronizer with external trigger<br>• Video frames per second: 200 | • Designed based on ARM Cortex M4<br>• 16-channel<br>• 16-bit resolution for each channel<br>• Sampling frequency: 96 kHz<br>• Microphone array with 16 directional microphones<br>• External trigger |

High efficiency PC workstation with RAID matrix, Matlab environment and StreamPix software

The AG500 articulograph allows for recording EMA signals from maximally 12 sensors. For this study, three of them were used for data normalization in order to correct head movements [32]. These reference sensors were placed on the left and right mastoid processes and on the ridge of the nose, close to its root. Another sensor was fixed to a wooden spatula and used for making palate, upper incisors and gums contours; one contour was traced with the speaker breathing through the nose (lowered velum), and one while breathing through the mouth (raised velum). This sensor was also used for temporomandibular joints localization. All the remaining sensors were attached to moving articulators (Fig. 2 and Fig. 3a).

Five sensors were placed on the tongue: four in the midsagittal plane (tongue tip – TT, tongue front – TF, tongue dorsum – TD and tongue back – TB) and one on the left side – TLS. Two sensors registered upper lip (UL) and lower lip (LL) trajectories. They were placed in the midsagittal plane just above and below the vermilion border. One sensor (J), fixed inside the oral cavity on the border of gums and lower incisors, registered mandible movements. Measurements of the spatial positioning of sensors attached to the tongue provided data on its shape and movements in time.

The video recording component consisted of three Point Grey cameras (Gazelle GZL-CL-22C5M-C model) recording movements of external articulators (e.g. lips, jaw and cheeks). One video camera was placed in front of the speaker and two others captured the image of both sides of the speaker's head. The number of the cameras used in this system allowed to obtain 3D positions of the facial OptiTrack markers through correlating the recorded 2D images. The arrangement of EMA sensors and reflective facial markers as used in the research is presented in Fig. 3.

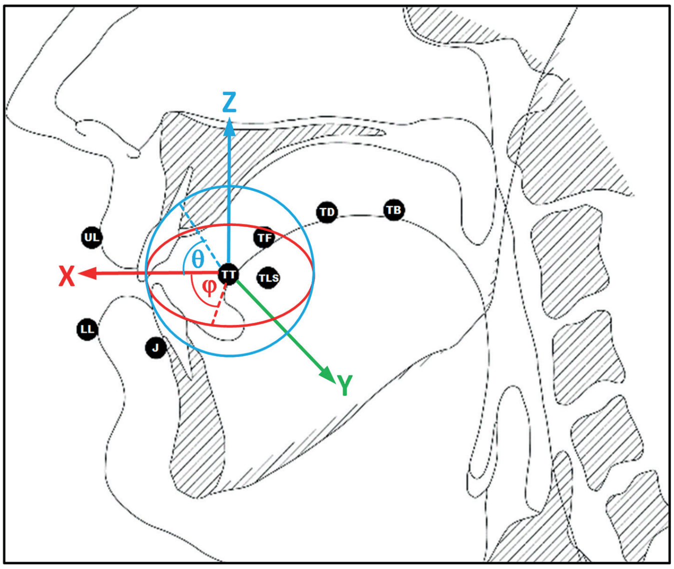


Fig. 2. EMA sensor placement on the tongue, lips and mandible. An example of the positioning of the coordinate system is shown with reference to the tongue tip sensor (TT)

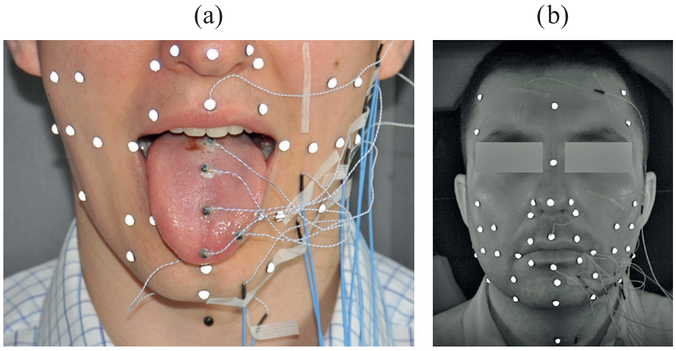


Fig. 3. Arrangement of articulograph sensors (a) and facial markers (b)

The video was recorded at the rate of 200 fps. The third component of the system is a microphone array with 16 directional microphones placed in a circle and connected to a 16-channel audio recorder. This device registers audio signal with a 96 kHz sampling rate and 16-bit resolution.

During the recording session, the speaker read out words displayed on the screen by the EMArecorder application. After receiving the command to start data acquisition, the articulograph generated synchronization signals for external devices in the TTL standard. The source of the synchroni-

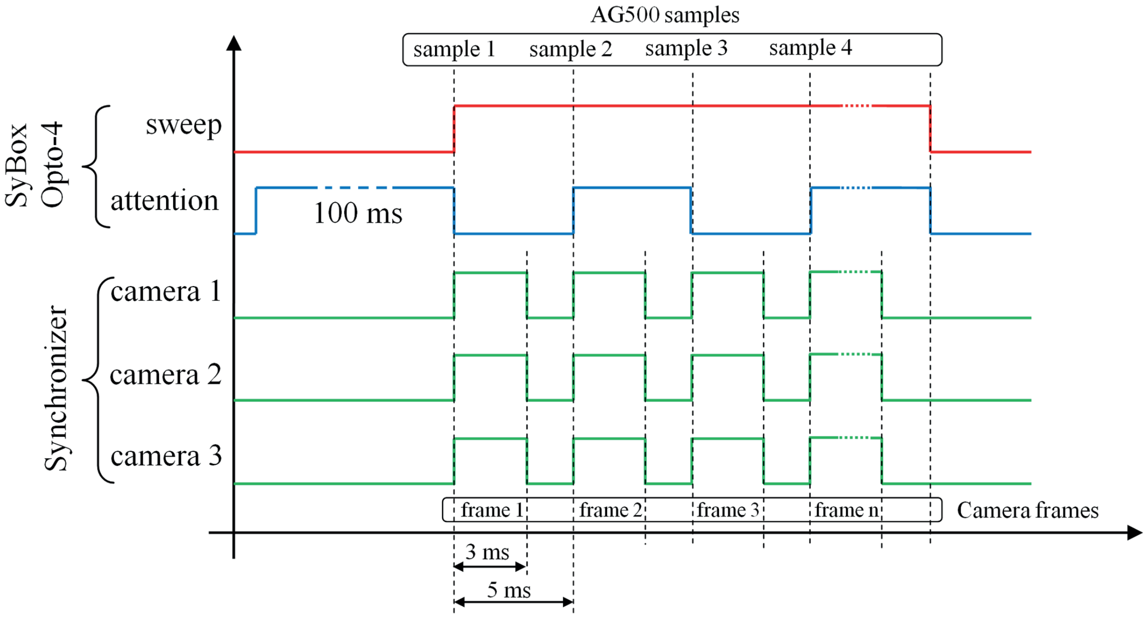


Fig. 4. The system's synchronization signals

zation signals was SyBox Opto-4, which is a standard component of AG500.

This module was responsible for activating the audio recorder and the synchronizer in the video system. Fig. 4 shows how the signals were generated during a recording sweep.

The sweep signal generated by the SyBox Opto-4 device carried information when the sampling was in progress [33]. The audio recorder and vision system used both edges of this signal to start and stop the data acquisition process. The synchronizer generated trigger signals for all video cameras as rectangular pulses with a 60% duty cycle and frequency of 200 Hz. The duration time of the high state was set at 3 ms to facilitate proper scanning of single images by the video cameras. All recordings stopped when sweep went into a low state. When *sweep* was in a high state, then the attention signal changed its value every 5 ms. The articulograph acquired samples on both edges of this signal.

Before the analysis, the data from AG500 were pre-processed so as to remove undesirable noise. Trials with the application of a second-order Butterworth low-pass filter with fixed cut-off frequency did not yield satisfactory results, which is shown in Fig. 5.

Instead, the Savitzky-Golay filter was applied, using a 4th order polynomial with a span of 21 samples. This kind of filter increased the signal-to-noise ratio without distorting the signal significantly. A double application of this algorithm was needed for more effective smoothing (see Fig. 5).

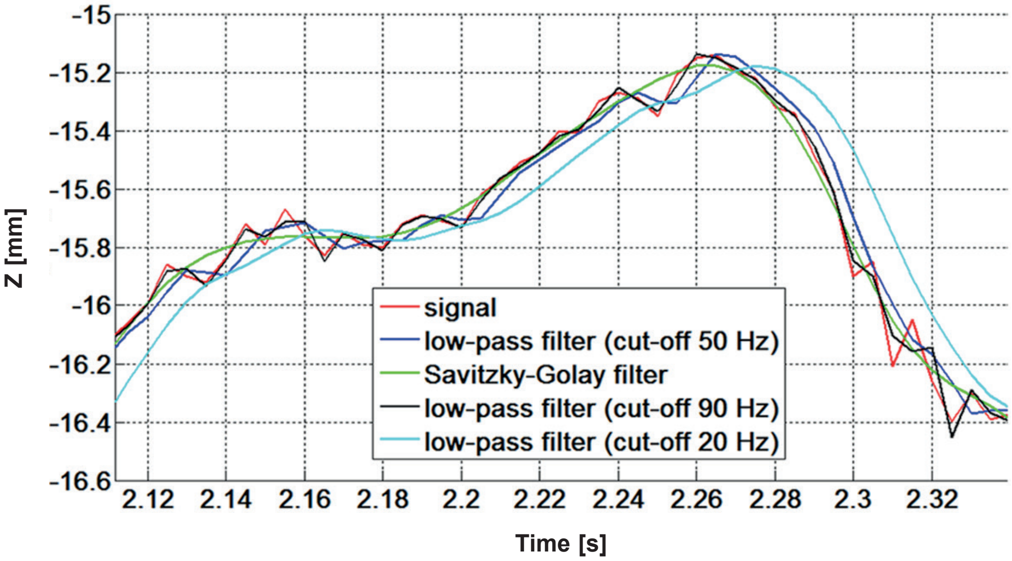


Fig. 5. Comparison of signal filtering techniques

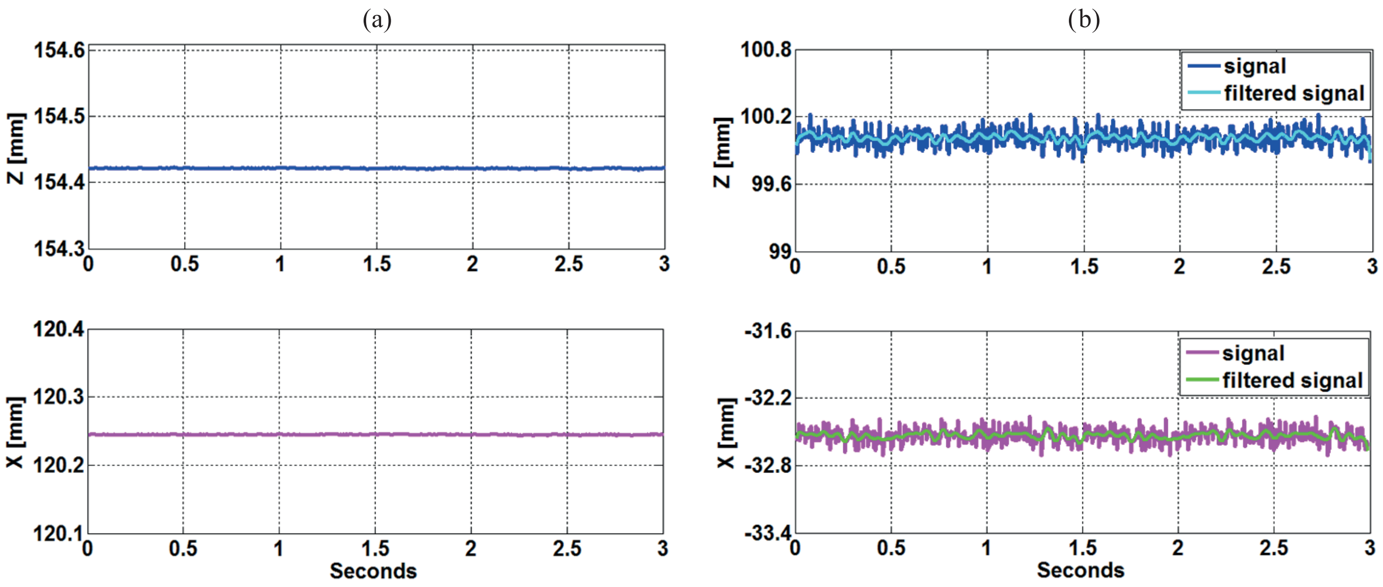


Fig. 6. Position data recorded for the X and Z coordinates in the static system: for a reflection marker (a – without filtering) and for an EMA sensor (b – before and after filtering)

## 3. System accuracy

The precise estimation of EMA sensors and video markers' position may be affected by various factors. The main sources of error for electromagnetic articulography are the following:

1) external sources of electromagnetic field (all devices in the neighbourhood that emit signals with frequencies that overlap the operating frequencies of EMA transmitters),
2) noise introduced by analogue-to-digital conversion,
3) slightly floating position of reference sensors caused by skin movements.

Meanwhile, main sources of errors affecting camera images include:

1) lighting,
2) noises introduced by analogue-to-digital conversion,
3) slightly floating position of reference video markers caused by skin movements.

In order to properly estimate the sources of errors mentioned above (except for floating reference sensor and marker positions), a rigid phantom with one EMA sensor and one reflection marker was prepared. The measurement was carried out in the static system. Standard deviations of all the measured coordinates were calculated in order to determine the error generated by the devices used (vision system and EMA).

The differences in the values on the vertical axes that are evident in Fig. 6 result from different positioning of the origin of the coordinate systems for the articulograph and the video system. The zero point of the system of coordinates for the camera image was placed in the lower left corner of the frame, whereas for articulographic data the beginning was in the middle of a sphere, whose intersection was imposed on the picture registered by the frontal camera. Table 2 presents standard deviation values for both axes in the static recording. X and Z dimensions were used for analysis of trajectory of sensors and facial markers.

Table 2
Standard deviation values for static recording of the position of video marker and EMA sensor

| Axis | Standard deviation [mm] | | |
|---|---|---|---|
| | *Marker* | *Sensor (non-filtered)* | *Sensor (non-filtered)* |
| X | 0.005 | 0.078 | 0.026 |
| Z | 0.005 | 0.068 | 0.030 |

As can be seen in Table 2, the precision of estimating the position of external articulators with cameras is somewhat better than with EMA, due to the high resolution of the cameras. However, the error at the level of micrometres does not significantly affect the accuracy of the sensor and marker's position estimation in either method. This demonstrates the proper configuration of equipment and the lack of significant noise sources. Filtering data from EMA sensors slightly improves the quality of data, which is presented in Table 2.

During the recording sessions, sudden jumps in sensor position from the current trajectory were observed. This problem is known and well described in other papers [34, 35], where authors show considerable positional errors in some regions of the EMA sphere. Such errors also occurred during intense articulation movements of the tongue (probably due to jerking the sensor wires) or when the sensors were at the boundary of the articulograph sphere. Due to the fact that trajectories were monitored in real time in JustView software (provided by the manufacturer of AG500), recordings in which such events occurred or in which the signal errors exceeded the manufacturer's norms were rejected. Another major problem

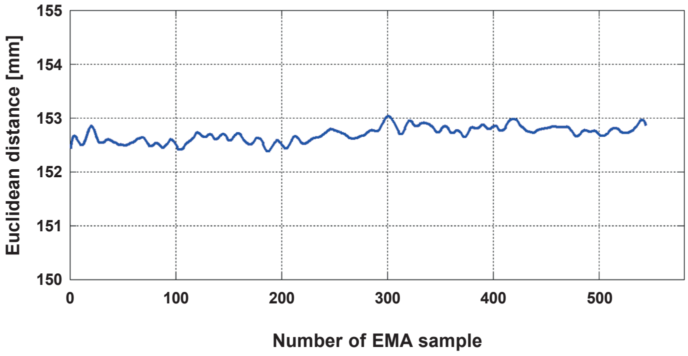


Fig. 7. Euclidean distance between 2 reference sensors (filtered signals) during recordings of a single word

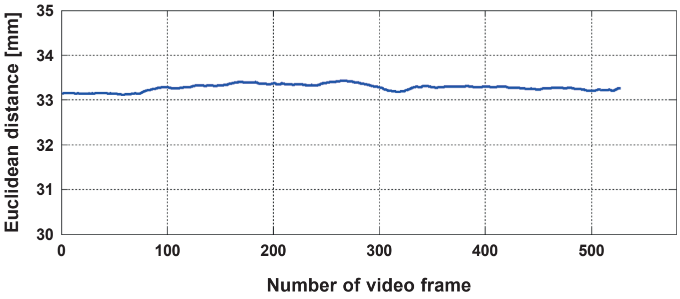


Fig. 8. Euclidean distance between 2 reference markers during a recording of a single word

were large numerical errors caused by the CalcPos software in some cases. Its numerical instabilities (described in [36, 37]) are probably caused by the Newton–Raphson method used to estimate sensor positions. Multi-channel analysis of articulographic data allowed to identify such cases and reject the tokens from analysis.

It is important to notice the high relative stability of the reference points used for head motion correction. The movements of the speakers' heads were unrestrained during recording sessions. Thus, while speaking, some of them nodded, shook or tilted their heads. In order to separate such movements from the traces of articulatory gestures, it is necessary to record data from reference sensors and markers that indicate the spatial position of the head at a given point in time. Due to these sensors and markers, it is possible to clear the data from non-reference sensors and markers of the shifts introduced by the movements of the head.

The *NormPos a*pplication [38], delivered with the AG500 articulograph, was used for head movement correction of EMA data; it corrects the position of all the sensors in relation to the position of reference sensors in a selected recording of a speaker.

The reference sensors are attached to three relatively stationary points of the head, as explained above. These points are not, however, completely stationary. Some slight displacements can be observed, which can be seen in Fig. 7 where changes in the values of the Euclidean distance between reference sensors are shown.

In the case of reference video markers, 6 of them were fixed on the forehead and the nasal bridge (relatively stationary points). Just 3 markers are sufficient for determining rotation and transformation matrices necessary for 3D position calculation and correction of the remaining markers. Due to the surplus number of reference markers, it was possible to select only those between which the Euclidean distance was most constant. The markers were not completely stationary during the recordings either, which is shown in Fig. 8.

Changes of Euclidean distance between reference sensors and markers usually did not exceed ±0.5 mm from the mean distance (in the worst case).

## 4. Examples of data analysis

**4.1. Analysis of EMA sensor movements**. In order to analyse articulographic data recorded by the system described here, a Matlab-based application named phonEMAtool was created. The application can visualize dynamic movement trajectories of all sensors (except reference ones) in the X (front – back) and Z (up – down) axes. Moreover, phonEMAtool was used to analyse and extract information related to the position of sensors at desired time frames, including values of 2 angles: $\varphi$ – azimuth and $\theta$ – elevation. Additionally, phonEMAtool calculated the velocity of sensors and identified their minimum and maximum values in time. Furthermore, a 20% threshold for rising or falling edges of the velocity function could be calculated, in accordance with the method presented in [32]. This option was used in order to avoid typical problems encountered in the estimation of the beginning and end of sensor movements.

The velocity of each sensor in the X (front – back) and Z (up – down) axes was calculated separately to estimate the dynamics of articulator movements along these axes. Instantaneous velocity was calculated by means of formula (1).

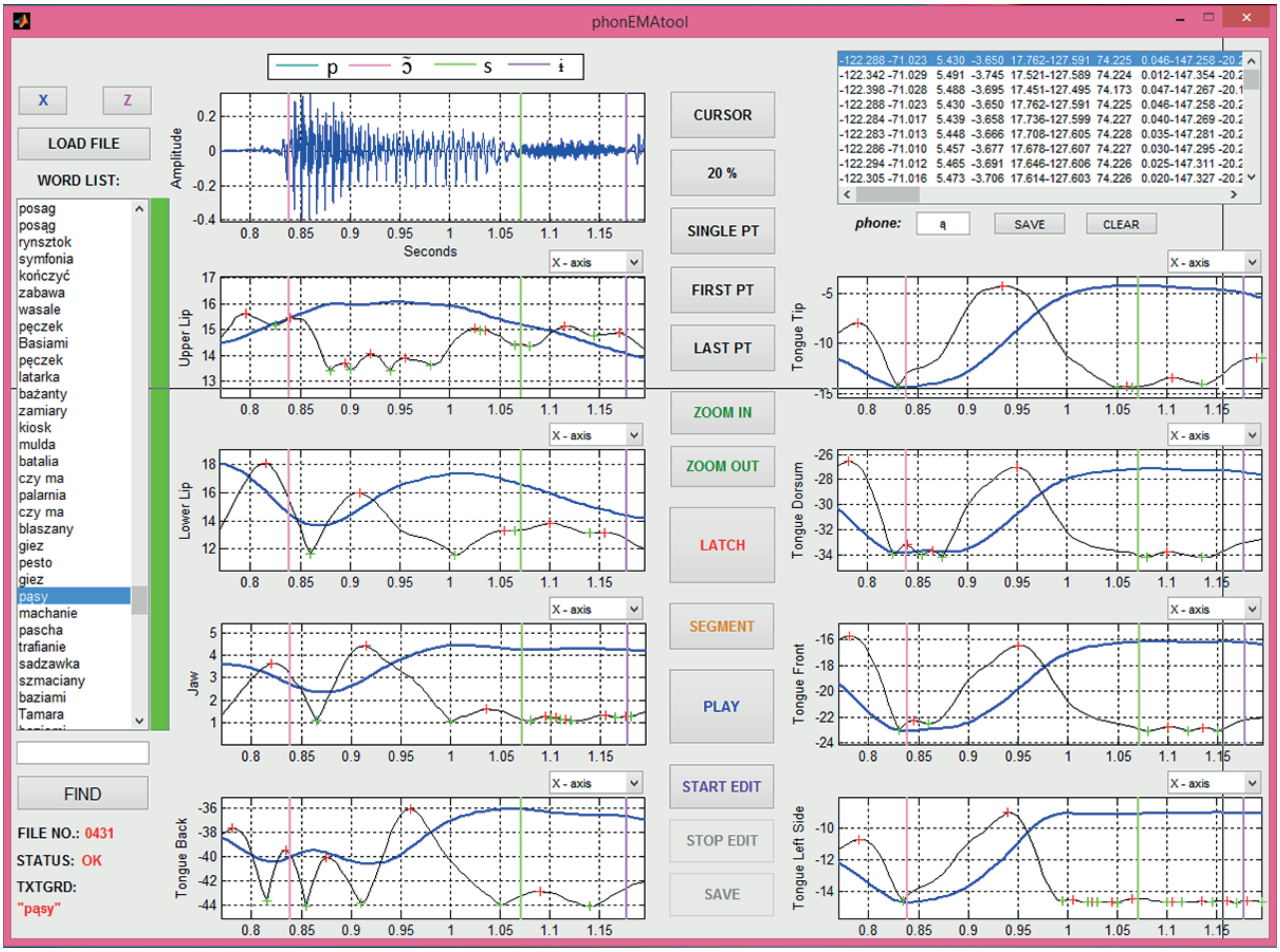
$$v_X = \sqrt{\left(\frac{dx}{dt}\right)^2} \quad \text{and} \quad v_Z = \sqrt{\left(\frac{dz}{dt}\right)^2}. \tag{1}$$

The absolute values were necessary to estimate zero and maximum readings regardless of the direction of the velocity vector. Speed curves based on such calculations were plotted in phonEMAtool's graphs. It was necessary to normalize their range so as to plot them in the same space as sensor trajectory tracings.

a) The phonEMAtool application facilitates simultaneous and synchronized processing of three data types:
b) audio – wav files
c) EMA – text files (binary data converted to column text format) acoustic segmentation – TextGrid files (Praat format) [39].

Figure 9 presents the workspace window of the phonEMAtool software during exemplary data analysis.

The software allows for isolating and exporting selected sets of data for further analysis.



posag
posąg
rynsztok
symfonia
kończyć
zabawa
wasale
pęczek
Basiami
pęczek
latarka
bażanty
zamiary
kiosk
mulda
batalia
czy ma
palarnia
czy ma
blaszany
giez
pesto
giez
pasy
machanie
pascha
trafianie
sadzawka
szmaciany
baziami
Tamara

```
-122.288 -71.023  5.430  -3.650  17.762-127.591  74.225  0.046-147.258 -20.2
-122.342 -71.029  5.491  -3.745  17.521-127.589  74.224  0.012-147.354 -20.2
-122.398 -71.028  5.488  -3.695  17.451-127.495  74.173  0.047-147.267 -20.1
-122.288 -71.023  5.430  -3.650  17.762-127.591  74.225  0.046-147.258 -20.2
-122.284 -71.017  5.439  -3.658  17.736-127.599  74.227  0.040-147.269 -20.2
-122.283 -71.013  5.448  -3.666  17.708-127.605  74.228  0.035-147.281 -20.2
-122.286 -71.010  5.457  -3.677  17.678-127.607  74.227  0.030-147.295 -20.2
-122.294 -71.012  5.465  -3.691  17.646-127.606  74.226  0.025-147.311 -20.2
-122.305 -71.016  5.473  -3.706  17.614-127.603  74.226  0.020-147.327 -20.2
```

FILE NO.: 0431
STATUS: OK
TXTGRD:
"pąsy"

Fig. 9. Workspace window of phonEMAtool software during articulatory gesture analysis of the word ‘pąsy’ (*blushes*, n.pl.) in the X axis

Figure 10 presents a part of the Polish word ‘kalambur’ (*charade*) with acoustic segment boundaries superimposed on the waveform and articulatory segment boundaries in the graph window displaying the tongue tip sensor trajectory (thick line) and its velocity (thin line).

The labels and criteria used for articulatory segmentation are based on the research of Best et al. [40]. The data displayed in Fig. 10 show that the onset of the tongue tip closing gesture (GONS) during the pronunciation of the consonant [l] (marked in the initial, 20-percent threshold on the rising slope of the velocity PVEL1), equivalent to tongue tip rise, begins already in the initial phase of the preceding vowel. The tongue tip sensor (TT) achieves maximum velocity (PVEL1) of this rising movement during the final part of the periodic waveform of the first vowel. The greatest tongue tip elevation corresponding to the formation of the apical closure during the articulation of the sound [l] coincides with the slump in the velocity (its minimum value marked in Fig. 10 as MinVEL1). The articulatory boundaries of this sound segment were determined at following time points:

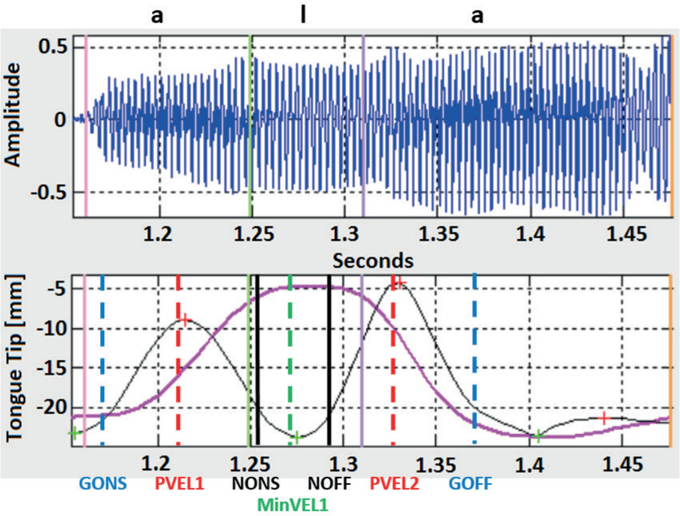


Fig. 10. Waveform and tongue tip (TT) sensor trajectory (pink colour) and velocity in Z axis (up-down) during pronunciation of the consonant [l] in the Polish word ’kalambur’ (*charade*)

- NONS (nucleus onset) – beginning of the articulatory gesture nucleus (final, 20-percent threshold PVEL1 on the falling slope: beginning of the obstruction)
- NOFF (nucleus offset) – end of the articulatory gesture nucleus (initial, 20-percent threshold PVEL1 on the rising slope: end of the obstruction)

During the articulation of [l], minimum velocity (MinVel) of the TT sensor is marked with a green dashed line within the nuclear phase in Fig. 10. After this phase, the sensor accelerates, which indicates the lowering of the tongue tip related to the release of the occlusion. Then the sensor reaches another maximum velocity value (PVEL2) after the beginning of the next vowel segment, very close to the left border of acoustic onset boundary of the [a] vowel. The end of the apical articulatory closing gesture (GOFF) executed to pronounce the consonant [l] (final, 20-percent threshold on the falling slope of the velocity function PVEL2) falls in the centre of the following vowel [a].

PhonEMAtool shares numerous functionalities with MVIEW – an application for displaying articulographic data created by Mark Tiede [41]. However, phonEMAtool was developed independently, solely for the purpose of the project described here so as to make the analysis more user-friendly. Operation of the software and its interface does not require its user to generate any Matlab commands.

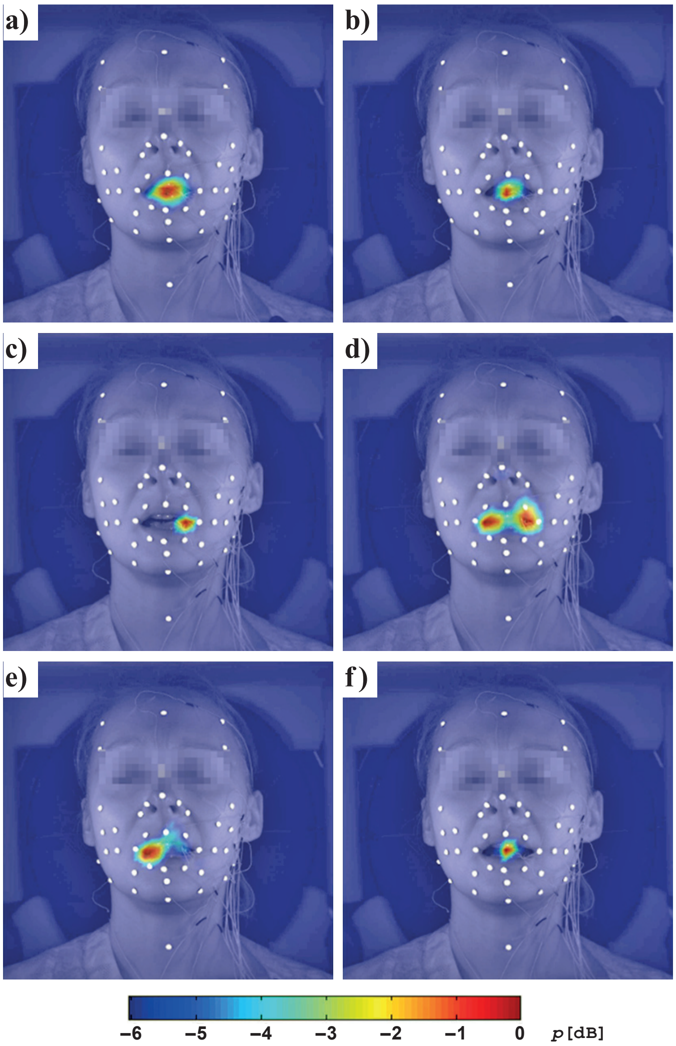


Fig. 11. Spatial distribution of acoustic field intensity synchronized with video images during the articulation of the sequence [ala] in the word ‘kalambur’ (*charade*). The frames show central release in the middle part of the first vowel [a] (a), the phase of [l] with central sound release (b), the phase of [l] with left-side unilateral release (c), the phase of [l] with bilateral release (d), the phase of [l] with right-side unilateral release (e) and the middle part of the second vowel [a] (f)

**4.2. Acoustic camera analysis**. The system described in this article allows to obtain more information about the articulation of the lateral consonant [l] as the acoustic camera enables analysis of spatial distribution of the acoustic field. Dividing the vertical facial plain into central (along the nose and central area of the lips), left and right (along the respective corners of the mouth) enables the researcher to pinpoint the sources of sound distribution and sound amplitudes in selected areas. This technique makes it possible to ascertain whether a sound is released centrally through the oral cavity or if its release is lateral. Moreover, it is possible to establish if an articulation is uni- or bilateral and, in the case of bilaterally released consonants, the dominance of either of the sides can be indicated at a given point in time.

Figures from 11a to 11f show all the aforementioned articulation types obtained with the use of the acoustic camera for the articulation of the sequence [ala] in the Polish word ‘kalambur’ (*charade*, n.sg.). The figures present video camera images of the speaker’s face synchronized with the acoustic energy distribution map. The colours, from dark blue through light blue, yellow, light red to dark red, represent increasing acoustic energy. Subsequent figures demonstrate central (Figs. 11a, 11b, 11f), unilateral right (Fig. 11c), bilateral (Fig. 11d) and unilateral left (Fig. 11e) articulations identified on the basis of the intensity and distribution of the acoustic field.

As can be seen in Fig. 11, the microphone matrix provides a possibility of differentiating between central and lateral articulations (including unilateral and bilateral ones). Moreover, this technique can be used for determining whether an articulation is oral, oronasal or nasal [29, 30], which would be impossible to assess solely with articulographic research.

Currently, the method used in the system described here emerges victorious as compared with the tools described by Krakow and Huffman [19] since it enables separate analysis of the oral and nasal signal. Moreover, it does not subject the recorded person to any inconvenience related to invasiveness, harmful effect of X-rays, or the discomfort of wearing a mask, or having an ultrasound device under the jaw.

**4.3. Video data analysis**. The vision system coupled with spatial analysis of the face markers allows to determine the location and movement trajectories of external articulators. The results may be compared to analogous data registered by the electromagnetic articulograph from sensors attached to the face of the speaker. In order to determine the alignment of data from the vision system and EMA, one sensor was attached to the upper and one to the lower lip, above and beneath vermillion borders, and face markers were fixed on top of those sensors. Due to this operation, it became possible to register lip movement trajectories by two different methods. The results are illustrated below with analysis of the minimum and maximum position of the lower lip sensor during articulation of the Polish word ‘sasanka’ (*pasque flower*). Figure 12 shows the movements of the lower lip sensor along the Z axis as registered by the AG500 articulograph and analysed in the *phonEMAtool* application.

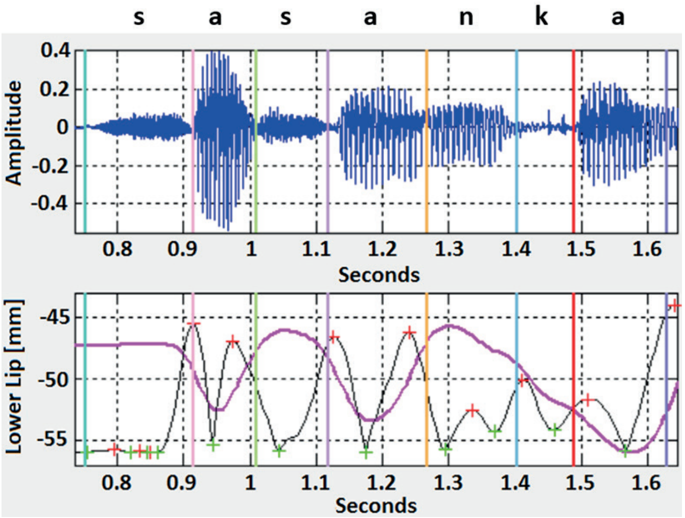


Fig. 12. The waveform and lower lip (LL) sensor trajectory (pink colour) and velocity (black colour) in the Z axis (up-down during pronunciation of the Polish word ‘sasanka’ (*pasque flower*)

Figure 13 presents the vertical movement of the same point (central, below the border of the lower lip) registered by the front camera.

When comparing the lower lip movement trajectories in Fig. 12 and Fig. 13, one can notice that their shape is almost identical. The main discrepancy between the figures consists in a ~3 ms delay in the onset and offset of the video marker trace as compared with the EMA trace; this difference results from the time necessary to scan the video image matrix. Despite this easily correctible delay, the video system on its own seems to provide high quality results and is a cheaper, alternative method for assessing the movements of external articulators.

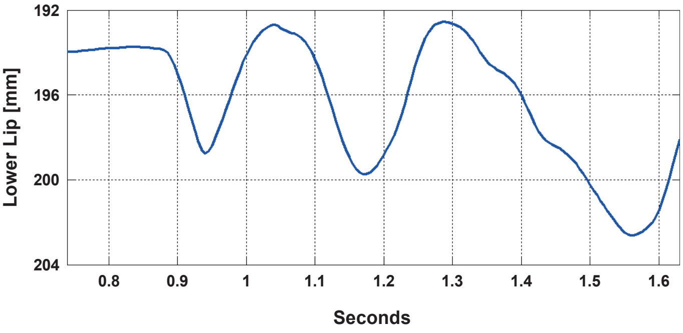


Fig. 13. Lower lip marker trajectory by the front camera during pronunciation of the Polish word ‘sasanka’ (*pasque flower*)

## 5. Conclusions

The measurement system described in this paper is an effective tool that facilitates simultaneous investigation of speech in 3 domains: articulatory movements, distribution of acoustic field intensity and movements of characteristic points on the face. The dedicated acoustic camera based on the multi-channel audio recorder and a microphone array is a particularly useful tool. The results obtained with this equipment are unique and provide multiple possibilities of practical application of the newly developed methods in the future.

The main goal of the article was to show possibilities and benefits of using measurement tools in parallel. For this reason, exemplary results for only one speaker are presented. In future research, it is intended to conduct comparative data analysis among speakers in order to find universal dependencies for normative pronunciations of Polish sounds.

**Acknowledgments**. Work described in this paper was supported by grant No. 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 decision No. DEC-2012/05/E/HS2/03770.

## References

[1] K. Mathiak, U. Klose, H. Ackermann, I. Hertrich, W.E. Kincses, and W. Grodd, “Stroboscopic articulography using fast magnetic resonance imaging”, *Int. J. Lang. Commun. Disord*., 35(3), 419–425 (2000).

[2] L. Davidson, “Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance”, *J. Acoust. Soc. Am*. 120(1), 407–415 (2006).

[3] K.L. Moll, “Cinefluorographic techniques in speech research”, *J. Speech Hear. Res*. 3, 227–241 (1960).

[4] T. Baer, J. Gore, S. Boyce, and P. Nye, “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels”, *J. Acoust. Soc. Am*. 90, 799–828 (1991).

[5] C. Moore, “The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images”, *J. Speech Hear. Res*. 35, 1009–1023 (1992).

[6] B. Wein, W. Angerstein, C. Neuschaefer-Rube, A. Obrębowski, and S. Klajman, “Badanie obwodowego narządu mowy przy wymowie głosek polskich za pomocą jądrowego rezonansu magnetycznego (NMR)”, *Polish Journal of Otolaryngology* (*Otolaryngologia Polska*) 48(2), 178–198 (1994).

[7] J. Dang, K. Honda, and H. Suzuki, “Morphological and acoustical analysis of the nasal and paranasal cavities”, *J. Acoust. Soc. Am*., 96(4), 2088–2100 (1994).

[8] A. Serrurier and P. Badin, “A three-dimensional linear articulatory model of velum based on MRI data”, *Proceedings of the 9th Eurospeech*, 2161–2164 (2005).

[9] M. Toda, S. Maeda, and K. Honda, “Formant-cavity affiliation in sibilant fricatives”, *Turbulent sounds: An interdisciplinary guide*, eds. S. Fuchs, M. Toda, M. Zygis, De Gruyter Mouton, 343–371 (2010).

[10] T. Sorensen, A. Toutios, L. Goldstein, and S.S. Narayanan, “Characterizing vocal tract dynamics with real-time MRI”, *15th Conference on Laboratory Phonology*, (2016)

[11] M. Stone, G. Stock, K. Bunin, K. Kumar, and M. Epstein, "Comparison of speech production in upright and supine position", *J. Acoust. Soc. Am.*, 122, 532–541 (2007).

[12] J.S. Perkell, M.H. Cohen, M.A. Svirsky, M.L. Matthies, I. Garabieta, and M.T. Jackson, "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements", *J. Acoust. Soc. Am.* 92(6), 3078–3096 (1992).

[13] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion", *Advances in Nonlinear Speech Processing – Lect. Notes Comput. Sc.* 4885, 263–272 (2007).

[14] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data", *Proceedings of the 5th AMDO*, 132–143 (2008).

[15] O. Engwall and J. Beskow, "Resynthesis of 3D tongue movements from facial data*"*, *Proceedings of the 8th Eurospeech*, 2261–2264 (2003).

[16] J. Beskow, O. Engwall, and B. Granström, "Simultaneous measurements of facial and intraoral articulation", *Proceedings of Fonetik*, 57–60 (2003).

[17] H. Kjellström and O. Engwall, "Audiovisual to articulatory inversion", *Speech Communication* 51(3), 195–209 (2009).

[18] D. Schabus, M. Pucher, and P. Hoole, "The MMASCS multimodal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech", *Proceedings of the 9th LREC*, 3411–3416 (2014).

[19] R.A. Krakow and M.K. Huffman, "Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory: An introduction", *Phonetics and Phonology 5: Nasals, Nasalization and the Velum*, 3–59 (1993).

[20] F. Bell-Berti, "An electromyographic study of velopharyngeal function in speech", *J. Speech. Hear. Res.* 19, 225–240 (1976).

[21] T. Bressmann, B. Radovanovic, G.V. Kulkarni, P. Klaiman, and D. Fisher, "An ultrasonographic investigation of cleft-type compensatory articulations of voiceless velar stops", *Clinical Linguistics & Phonetics*, 1028–1033 (2011).

[22] S. Rossato, P. Badin, and F. Bouaouni, "Velar movements in French: An articulatory and acoustical analysis of coarticulation" *Proceedings of the 15th ICPhS*, 3141–3144 (2003).

[23] D. Warren, "Velopharyngeal orifice size and upper pharyngeal pressure-flow patterns in normal speech", *Plast. Reconstr. Surg.* 33, 148–161 (1964).

[24] D. Warren, "Velopharyngeal orifice size and upper pharyngeal pressure-flow patterns in cleft palate speech: a preliminary study", *Plast. Reconstr. Surg.*, 34, 15–26 (1964).

[25] D. Warren, "Nasal emission of air and velopharyngeal function", *Cleft Palate J.*, 16, 279–285 (1967).

[26] M. Rothenberg, "Measurements of air flow in speech", *J. Speech Hear. Res.* 20, 155–176 (1977).

[27] S.G. Fletcher and M.E. Bishop, "Measurement of nasality with TONAR", *Cleft Palate J.* 7, 610–621 (1970).

[28] R.M. Dalston, D.W. Warren, and E.T. Dalston, "Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairment", *Cleft Palate J.* 28(2), 184–189 (1991).

[29] A. Lorenc, „Wymowa normatywna polskich samogłosek nosowych i spółgłoski bocznej", *Dom Wydawniczy ELIPSA*, Warszawa 2016.

[30] D. Król, A. Lorenc, and R. Święciński, "Detecting laterality and nasality in speech with the use of a multi-channel recorder", *Proceedings of the 40th IEEE ICASSP*, 5147–5151 (2015).

[31] A. Lorenc, R. Święciński, and D. Król, "Assessment of sound laterality with the use of a multi-channel recorder", *Proceedings of the 18th ICPhS*, (2015)

[32] P. Hoole, C. Mooshammer, and H.G. Tillmann, "Kinematic analysis of vowel production in German", *Proceedings of the 3rd ICSLP*, 53–56 (1994).

[33] Carstens Medizinelektronik GmbH, "SyBox Opto-4 Manual ver. 0", http://ag500.de/Sybox_man.pdf

[34] C. Kroos, "Measurement accuracy in 3D electromagnetic articulography (Carstens AG500)", *8th International Seminar on Speech Production*, 61–64 (2008).

[35] Y. Yanusova, J.R. Green, and A. Mefferd, "Accuracy assessment for AG500, electromagnetic articulograph", *J. Speech Lang. Hear. Res.* 52(2), 81–84 (2009).

[36] P. Hoole and A. Zierdt, "Five-dimensional articulography" *Speech Motor Control: New developments in basic and applied research*, eds. B. Maassen and P.H.H.M. Van Lieshout, 331–349 (2009).

[37] M. Stella, P. Bernardini, F. Sigona, A. Stella, M. Grimaldi, and B. Gili Fivela, "Numerical instabilities and three-dimensional electromagnetic articulography", *J. Acoust. Soc. Am.* 132(6), 3941–3949 (2012).

[38] Carstens Medizinelektronik GmbH, "NormPos Program Description", http://www.ag500.de/manual/ag500/NormPos.pdf

[39] P. Boersma and D. Weenink, "Praat: doing phonetics by computer" [software], (2016). Different versions retrieved in 2016 from: http://www.praat.org/.

[40] C.T. Best, C. Kroos, R.L. Bundgaard-Nielsen, B. Baker, M. Harvey, M. Tiede, and L. Goldstein, "Articulatory basis of the apical/laminal distinction: tongue tip/body coordination in the Wubuy 4-way coronal stop contrast", *Proceedings of the 10th ISSP*, 33–36 (2014).

[41] M. Tiede, "MVIEW: Multi-channel visualization application for displaying dynamic sensor movement" [software], (2010).