Verification of Articulatory Phonetics Features with Quantitative Data

László Czap

Institute of Automation and Infocommunication, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary, czap@uni-miskolc.hu

Abstract: This paper aims to refine the base data set of visemes – the visual counterparts of phonemes – with quantitative data to provide accurate input for visual speech synthesis (a talking head that supports the training of speech production of deaf and hard-of-hearing children). Measurement-based features extend the existing data and refine our previously used dynamic model of articulation. This requires the definition of two major types of data simultaneously: the shape of the mouth, which can be examined relatively simply in an ordinary camera image, and the position of the tongue, the analysis of which requires the use of medical-level imaging devices and the processing of their signals. Articulatory phonetics can be divided up into three areas to describe consonants. These are voice, place, and manner respectively. This study aims to confirm the description of the place of articulation with measurement data. Data derived from the shape and position of the tongue is suitable for determining the place of articulation of sounds. In the case of vowels, we estimated the tongue position with the centroid of the tongue while in the case of consonants, we define the place of articulation with the measured distance of the tongue from the palate. To measure these, we used MRI and US images and determined tongue contours with an automated process. The results of this analysis statically define data for articulation keyframes for visual speech synthesis. We applied our results to improve the existing Hungarian transparent talking head with a more accurate model based on the clarification of the dynamic features. We also adapted the same model to the Chinese Shaanxi Xi'an dialect.

Keywords: Quantitative tongue description; Articulatory phonetics; Place of articulation; Talking head; Viseme features

1 Introduction

Previous studies show that visual information on the physiological processes of human speech greatly contributes to understanding the complex mechanism of speech formation and, through this, to the effective development of speech synthesis methods [1]. The radiological and monitoring processes currently available, such as magnetic resonance imaging (MRI) [2], computer tomography

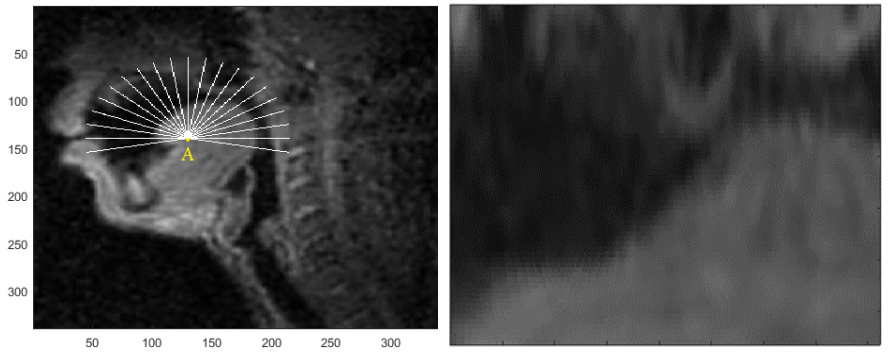
(CT) [3], ultrasound (US) [4], electropalatography (EPG) [5], or electromagnetic articulography (EMA) [6] are indispensable in getting to know the dynamic features of articulation. This is because the morphological and geometric data obtained with the help of imaging techniques can be used to map the articulation movements belonging to the given speech signal. This is essential, for example, in the parameterization of a talking head imitating the articulation. In this research quantitative data from a series of MRI and US images have been derived. Thus, we provided appropriate parameters for our animation algorithm. The main feature of this application is to show the tongue movements in a transparent-faced talking head. The basic items of this animation are the visemes. Such a system can be used well in speech therapy, in the design of non-native language learning training, or even in the construction of synthesizers to convert articulation features into silent speech [7]. Adaptation for a Chinese dialect has been examined as well.

The paper aims to provide a new quantitative method for analyzing tongue movements. Deaf and hard-of-hearing people are accustomed to lipreading, but unable to observe the invisible tongue movements. Without full acoustic perception, they rely on the visual modality of speech to be able to form their special speech signals. The quantitative data obtained helped in the better realization of a transparent talking head.

2 Methods and Material

The processing of MRI and US images was performed during the static and dynamic analysis. The programs for this were written in MATLAB environment, in the framework of which we fitted an auxiliary curve to the surface of the tongue based on dynamic programming [8].

The resolution of the raw MRI images is 320×320 pixels, as Figure 1a shows. As the first step of preprocessing, the image is resampled radially in the midsagittal MRI cross-section image by forming radial lines from a visually selected circle center (see point A in Figure 1a). (Here and in the following, scaled figures are in pixels.) This is necessary to avoid the appearance of two contour points in the same column of the image – where the tongue contour bends back – that the edge-detecting algorithm could not handle. For the sake of clarity, lines are only shown by ten degrees in Figure 2a but in reality, resampling is done by one degree. Arranging the sections thus obtained in a Cartesian column, a matrix is gained. The resampled image is represented in the Descartes coordinate system (Figure 1b). The bottom line contains the center point while the top line represents the points of the circumference of the circle.

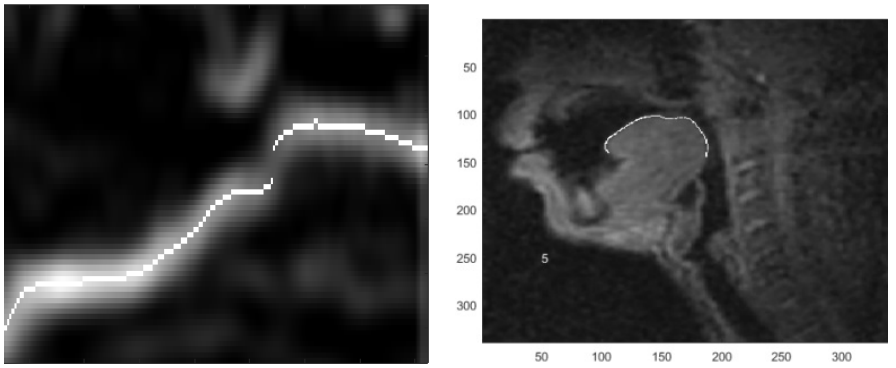


(a) (b)

Figure 1

Preprocessing, Step 1: (a) resampling the MRI image radially, (b) the Cartesian column matrix of the resampled image

In the second step, in the matrix, we find the largest cumulative luminance curve in the image obtained after edge enhancing with dynamic programming (Figure 2a). Processing is done from the left column to the right column of the image. The identified contour is indicated with white points in Figure 2a. The uneven tongue contour is smoothed by filtering before further processing. The smoothed tongue contour represents the base for the further analysis of articulation. In the image of Figure 2b, the tongue contour can be followed by projecting it back to the original image. The definition of the tongue contour offers an opportunity to perform various analyses.



(a) (b)

Figure 2

Preprocessing, Step 2: (a) The highlighted edge of the Cartesian image marks the found tongue contour with white points, (b) the tongue contour is projected onto the original image

2.1 Analysis of Tongue Position

The definition of the tongue contour makes it possible to calculate geometric features for a segmented part of it. Due to a lack of Hungarian recordings, a multilingual MRI visual database [9] was used to determine the tongue position associated with each speech sound as the male speaker produced vowels and VCV sound sequences (V: vowel, C: consonant). Through the exploration of the place of articulation in MRI images, we obtained static viseme data for each speech sound.

2.1.1 Method of Defining Tongue Position of Vowels with Quantitative Data

The idea was that by defining the centroid of the cross-section of the tongue body, we could obtain quantitative data about the horizontal and vertical positions of the tongue characteristic of the current speech sound. The centroid (C_{xy}) of the tongue is derived as the first-order momentum of the horizontal and vertical coordinates of the white points of the filled-up tongue body (1) as shown in Figure 3 [10, 11].

$$\boldsymbol{Cxy} = [\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}];\ \overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \boldsymbol{x} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}),\quad \overline{\boldsymbol{y}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \boldsymbol{y} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

where $f(x,y)=1$ in the white area, $f(x,y)=0$ outside the white area, and n is the number of white points.

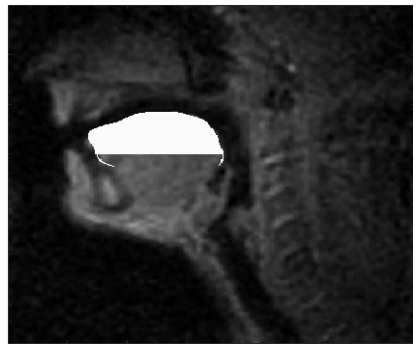


Figure 3
Filling the section of the tongue (sound /ɛ/)

We also need to determine the optimal number of pixel rows to fill the tongue body downwards from the top point of the tongue to define the centroid. Filling too few rows might lead to inaccurate measurement of the tongue position while filling too many rows might lead to oversimplification and losing the characteristic tongue position of distinct sounds.

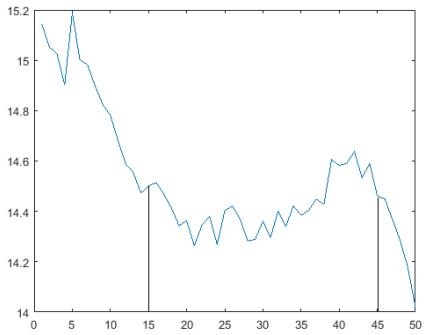


Figure 4
The variance of the centroids in pixels of the 28 vowels (vertical axis) of the database as a function of the depth of filling (horizontal axis)

To determine the filling depth, the standard deviation of the center of gravity of all vowels of the multilingual video database was examined, looking for a maximum for the highest distinction. In Figure 4, the variance (the average of deviations from the mean) decreases by filling over 45 rows of pixels as the tongue's root is less discriminative. Selecting less than 15 rows defines a cross-section representing just the top of the tongue and not the mass of it.

Based on the argumentation above, the tongue centroid for the vowels was investigated filling the depth of 42 rows of pixels for each vowel by the given resolution of the MRI image. In physical dimensions, the upper 22 millimeters of the tongue cross-section were selected.

2.1.2 Method of Defining the Place of Articulation for Consonants with Quantitative Data

The articulation of consonants is substantially different from that of vowels. This can be characterized by the place of articulation, which is determined by a gap or closure formed by the lip-tongue-jaw movement.

The place of articulation is determined by the narrowing or closure formed by the articulatory movements [12]. Thus, the place of articulation can be assigned to the place of the narrow part formed by the tongue and the unmoving part of the oral cavity.

The contour of the alveolar ridge and the palate can be defined analogously to the definition of tongue contour. The only difference is that moving away from the circle center we need to find not decreasing brightness – a falling edge – but rather increasing brightness – a rising edge. In images where the palate has no sharp edge the palate contours are defined by averaging several images.

Once we know the tongue contour and the palate contour, the distance of the two curves can be defined point by point. The distance measure derived from the definition of Nearest Neighbor Distance (NND) is suitable for determining the distance of curves consisting of a different number of points [13]. Let us take the two curves defined by their samples: U=[u_1, u_2, …, u_n] and V=[v_1, v_2, …, v_m]. Let the distance of one point of U from curve V be the distance of the point belonging to V nearest to it, and conversely, according to (2).

$$DU(i) = \min_j |u_i - v_j| \qquad DV(i) = \min_j |v_i - u_j| \tag{2}$$

Figure 5 shows the nearest neighbors of the tongue and palate.

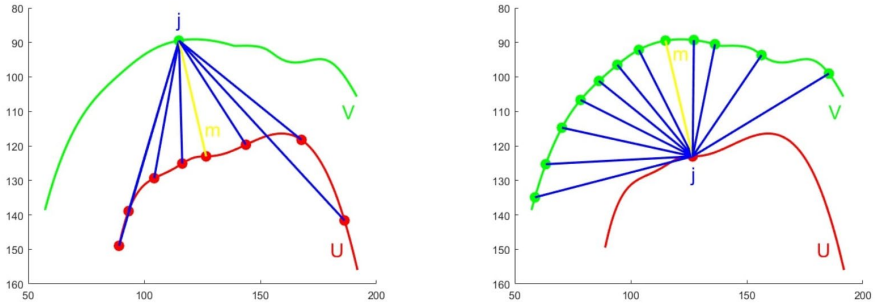
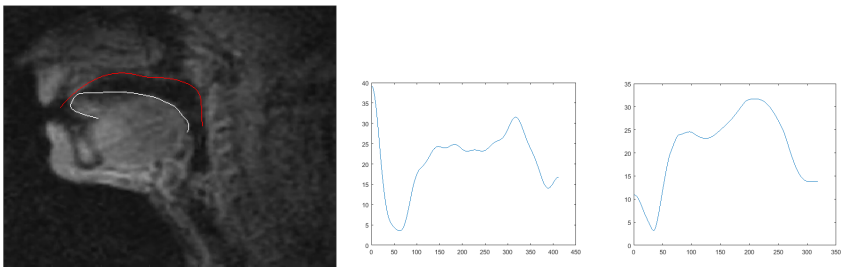


Figure 5
Graphical representation of NND

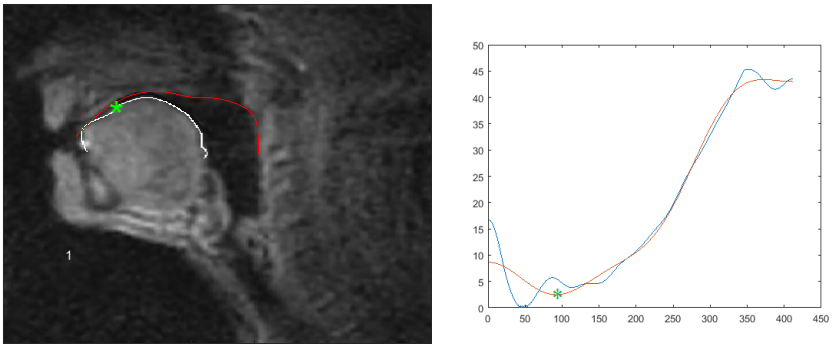
Figure 6a shows the alveolar ridge-palate contour (red line) and the tongue contour (white line). Figure 6b shows the minimum distance measured from the palate (vertical axis) to the points of the tongue (horizontal axis) while 5c shows the minimum distance measured from the tongue contour to the points of the palate. On the horizontal axis, the serial number of the tongue and palate contour points respectively, on the vertical axis the NND corresponding to the current contour point can be seen, measured in pixels. The place of articulation is considered the point of the tongue belonging to the smallest distance.



(a) (b) (c)

Figure 6
For the sound /r/: (a) The contour of the palate (red) and tongue contour (white), (b) the distance of the palate measured from the tongue, (c) distance of the tongue measured from the palate

With fricatives and approximants, a longer section of the tongue is close to the palate. With such sounds, it seems appropriate to regard the center as the middle of the whole near section. With low-pass filtering of the distance function, the place of the curve minimum can be shifted to the middle of the narrow section as is shown in Figure 7b. Discrete cosine transformation was used to filter the curve.



(a) (b)

Figure 7

(a) The tongue contour of sound /j/ and (b) the filtering of the distance function, (the blue line represents the original curve, and the red is the filtered one)

Selecting the appropriate frame of the video stream to represent the sound is a crucial point of this analysis. In the case of stop sounds and affricates, the frame before the burst is marked as the representative frame of sound; for the other consonants, the middle point of the time interval of the sound is selected.

3 Results

In our experiments, we obtained quantitative data from the determination of the tongue contour during speech. The results are shown separately for vowels and consonants, comparing them to traditional descriptive phonetic data.

3.1 Vowels

In the case of vowels, cross-section data formed along the longitudinal axis of the vocal tract are affected by jaw openness and tongue position. The narrower and wider sections of the vocal tract and the lips' shape determine the spectral properties of the excitation signal coming from the larynx [14]. Figure 8a shows the position of vowel articulation according to the phonetic parameters with the conventional representation in literature. This figure is adopted from the website of the International Phonetic Alphabet, (IPA).

On MRI recordings of vowel announcements, the tongue contour of the image taken from the center of the stationary phase of the sound was filled up to a line depth of 22 millimeters after automatic contour selection. Figure 8 shows the centroid of the tongue while pronouncing the sound /ɛ/.

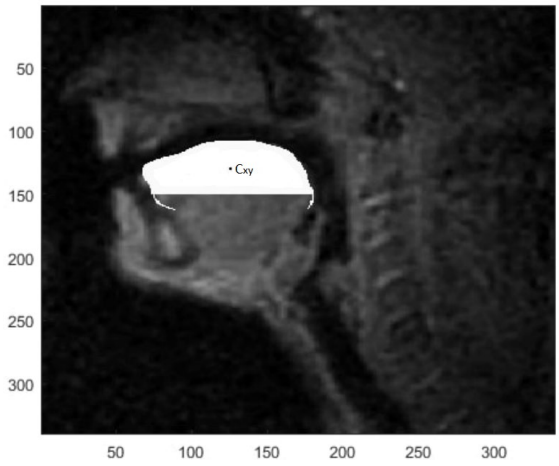


Figure 8
The centroid of sound /ɛ/

Figure 9b shows the C_{xy} centroid of the filled-up tongue in the oral cavity according to the 320×320 pixel coordinate system in, e.g. Figure 1a and Figure 2b, which visually reflects the phonetic arrangement of Figure 9a.

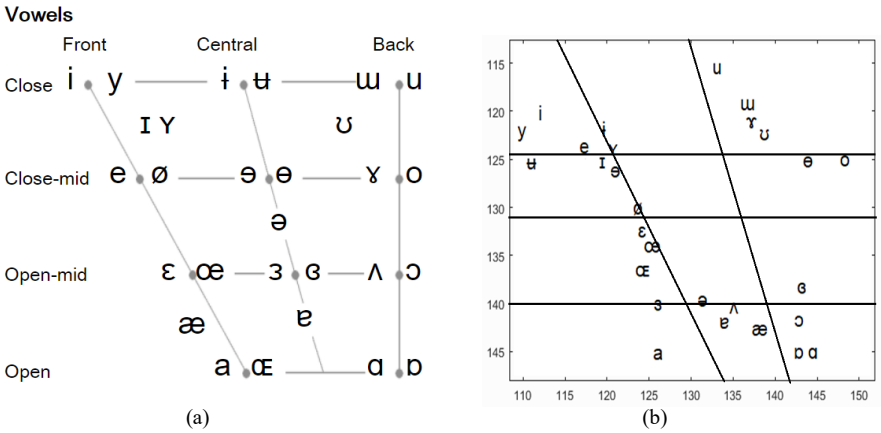


Figure 9
(a) Articulation map of vowels [15], (b) tongue centroids in the MRI images (the back part of the oral cavity is shown on the right, the front part on the left)

The traditional map of tongue position – divided vertically into four and horizontally into three sections – was compared with the measured centroid data.

In Table 1, the tongue positions located in the correct section are indicated in unshaded cells. Gray shading indicates a one-box difference either horizontally or vertically. The more white cells are in the table, the better the theoretical and the measured quantitative data match each other. No differences greater than one box were measured.

Table 1
The accuracy of the tongue positions

	horizontal	vertical		horizontal	vertical
i			o		
y			ə		1
ɨ			ɛ		
ʉ	1	1	œ		
ɯ			ɜ	1	
u			ɞ	1	
ɪ		1	ʌ	1	
ʏ			ɔ		
ʊ			æ	1	
e			ɐ		
ø			a		
ɘ			ɶ		1
ɵ	1		ɑ		
ɤ		1	ɒ		

3.2 Consonants

Table 2 shows the place of articulation of consonants. The tongue position of the bilabial and labiodental sounds was not examined. The reason for this is that in the formation of these sounds, the position of the tongue is indeterminate, that is, it adapts to the tongue position of the adjacent sounds. In these cases the place of articulation is not determined by the tongue but by the teeth and lips.

Table 1
Place of articulation of consonants [16]

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 10 shows the places of articulation obtained with the method described in 2.1.2.

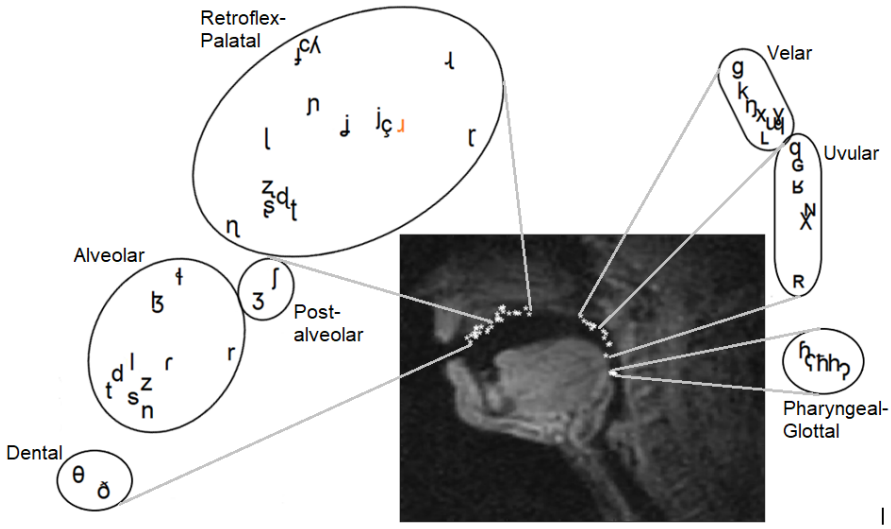


Figure 10
Calculated place of articulation of consonants marked in the MRI image

The theoretical and quantitative defined places of articulation differ only for a single sound /ɹ/ (marked with red in the figure). This means that the places of articulation defined with quantitative data match the physiological definitions well.

3.3 Application of the Results in the Speech Assistant System

The Speech Assistant system – elaborated for the Hungarian language to support the deaf in learning to speak – was further refined by incorporating the results of this work [17]. Figure 11 shows the visualized image of the reference pronunciation (bottom) and that of the sound recorded during practice (top), while on the right side, it displays the transparent talking head in two views. The talking head for the Shaanxi Xi'an dialect of Chinese and its Speech Assistant system is under development.

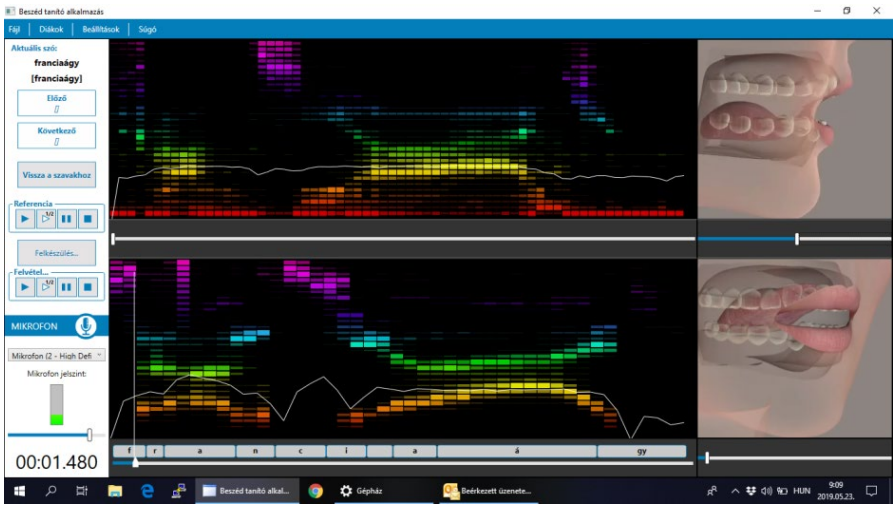


Figure 11
Screen view of the Speech Assistant during practice

Conclusions

A quantitative analysis of articulation was performed mainly to make the articulation visible in a transparent talking head. In the case of vowels, the tongue position was described with first-order momentums derived from MRI images. By consonants, the place of articulation was identified with the place of the closure or narrowing between the tongue and the alveolar ridge or the palate. The results of the approach confirm the suitability of quantitative analysis for verifying descriptive phonetic classifications. Thus, an important step towards the quantitative description of the articulation of speech production was taken. The traditional phonetic classification of speech sounds, the calculated place of articulation, and the tongue position defined by measurements are consistent with descriptions reported in the relevant literature. Supporting articulatory phonetics with quantitative data requires further, more detailed investigations. The native English speaker perfectly articulated the sounds of the International Phonetic Alphabet. The obtained results do not contradict the findings of the Hungarian

descriptive phonetic classifications. According to the testimony of Figure 9 and Table 1, the place of articulation of the vowels matches the descriptive phonetics data with at most one classification section error. The extension of the analysis to a larger number of speakers and different sound environments offers the possibility of improvement.

The results show that the analysis of the articulation of the Chinese Shaanxi Xi'an dialect speaker based on ultrasound images makes it possible to define the static data of visemes but also offers the opportunity to perform a dynamic description of the articulation. The presented analyses support visual speech synthesis with quantitative data that go beyond phonetical considerations.

References

[1] Barnaud, M. L., Schwartz, J. L., Bessière, P., and Diard, J. (2019) Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. PLoS ONE, 14, 1, https://doi.org/10.1371/journal.pone.0210302

[2] Miller, N. A., Gregory, J. S., Aspden, R. M., Stollery, P. J., & Gilbert, F. J. (2014) Using active shape modeling based on MRI to study morphologic and pitch-related functional changes affecting vocal structures and the airway. Journal of Voice, 28(5), 554-564, https://doi.org/10.1016/j.jvoice.2013.12.002

[3] Baum, S. R., Blumstein, S. E., Naeser, M. A., and Palumbo, C. L. (1990) Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech. Brain and Language, 39(1), pp. 33-56, https://doi.org/10.1016/0093-934X(90)90003-Y

[4] Recasens, D. (1991) On the production characteristics of apicoalveolar taps and trills. Journal of Phonetics, 19(3-4), pp. 267-280, https://doi.org/10.1016/s0095-4470(19)30344-4

[5] Czap, L. (2021) Impact of Preprocessing Features on the Performance of Ultrasound Tongue Contour Tracking, via Dynamic Programming. Acta Polytechnica Hungarica, Vol. 18, No. 2

[6] Serrurier, A., Badin, P., Barney, A., Boë, L. J., & Savariaux, C. (2012) The tongue in speech and feeding: Comparative articulatory modelling. Journal of Phonetics, 40(6), 745-763, https://doi.org/10.1016/j.wocn.2012.08.001

[7] Hueber, T., Benaroya, E. L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010) Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication, 52(4), pp. 288-300, https://doi.org/10.1016/j.specom.2009.11.004

[8] Zhao, L., Czap L. (2019) A nyelvkontúr automatikus követése ultrahangos felvételeken. (Automatic tracking of the tongue contour on ultrasound

recordings.) *Beszédkutatás* 27(1), pp. 331-343, https://doi.org/10.15775/ Beszkut.2019.331-343

[9] sail.usc.edu/span/rtmri_ipa/je_2015.html, Accessed 18.02.2020

[10] Hu, M. K. (1962) Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, 8(2), pp. 179-187, https://doi.org/10.1109/TIT.1962.1057692

[11] Mukundan, R., and Ramakrishnan K. R. (1998) Moment functions in image analysis. Singapore: Word Scientific Press. pp. 11-24

[12] Erdogan, N., & Wei, M. (2019) Articulatory Phonetics: English Consonants. In Erdogan, N., & Wei, M. (Ed.), Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners pp. 263-284, IGI Global, http://doi:10.4018/978-1-5225-8467-4.ch011

[13] Zharkova, N., & Hewlett, N. (2009) Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /a{script}/. Journal of Phonetics, 37(2), pp. 248-256, https://doi.org/10.1016/j.wocn.2008.10.005

[14] Ivanova, S. A., & Hasko, V. (2019) Articulatory Phonetics: English Vowels. In Erdogan, N., & Wei, M. (Ed.), Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners pp. 285-301, IGI Global, http://doi:10.4018/978-1-5225-8467-4.ch012

[15] https://www.internationalphoneticalphabet.org/ipa-charts/vowels/ Accessed 22.04.2020

[16] https://www.internationalphoneticalphabet.org/ipa-charts/consonants/ Accessed 22.04.2020

[17] Czap, L., Pintér, J. M., & Baksa-Varga, E. (2019) Features and results of a speech improvement experiment on hard of hearing children. Speech Communication, 106, pp. 7-20, https://doi.org/10.1016/j.specom. 2018.11.003