Impact of Preprocessing Features on the Performance of Ultrasound Tongue Contour Tracking, via Dynamic Programming

László Czap

University of Miskolc, Egyetemváros, 3515 Miskolc, Hungary
czap@uni-miskolc.hu

Abstract: The automated assessment of ultrasound images for speech processing is a difficult process. The number of frames processed, amounts to several hundred thousand, this makes assessment nearly impossible to process manually. Tongue contour tracking is indispensable for the dynamic modelling of articulation. The difficulty of the task lies in the fact that the images have a noisy background and the contour curve is discontinuous. An algorithm based on dynamic programming has been developed to track the movement of the back of the tongue. With an extreme size edge enhancing and averaging construction, the procedure addresses the problems of break of discontinuity and noise, simultaneously. In the image obtained after smoothing, the brightest curve is sought, from the left to the right edges of each image. The points of the curve thus obtained, follow the uneven line of the tongue contour. To smooth the curve, filtering based on Discrete Cosine Transformation (DCT) is applied. With the appropriate selection of universal parameters and processing the signals of several speakers in an identical way, the accuracy of edge detection can be enhanced considerably. We have optimized and qualified the results, comparing them with manual contour tracking. The accuracy of contour tracking may be improved by applying speaker-specific adjustments. The results of this analysis define temporary data for articulation key frames for visual speech synthesis (a talking head). Beyond the static analysis, we also investigate the trajectories of articulation features over time. We refined our previously created dynamic model, in order to construct a full dataset for the articulation.

Keywords: ultrasound imaging; tongue contour tracking; dynamic programming; image preprocessing; visual speech synthesis

1 Introduction

Several studies reliably prove that the visual information obtained involving the physiological processes of human speech, greatly promotes the understanding of the complex mechanism of speech formation, and through this, the efficient development of speech synthesis methods [1]. Radiological and monitoring processes currently available, like magnetic resonance imaging (MRI), computer tomography (CT), ultrasound, electropalatography (EPG), electromagnetic articulography (EMA) or electroglottography (EGG) are indispensable in getting to know the dynamic features of articulation. Using the morphological and geometric data obtained with the help of imaging techniques, it is possible to explore the articulatory movements belonging to a particular speech sign, which is of crucial importance in parametrizing a talking head, imitating articulation. In this research quantitative data from a series of MRI and ultrasound 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 training to improve the speech production of deaf and hard of hearing children. Such a system can well be used in, for example, speech therapy, in the elaboration of non-native language learning training or in the construction of synthesizers necessary to convert articulation features into silent speech [2] [3] [4].

This paper is concerned with automated tongue contour tracking on the basis of the processing of ultrasound images. The ultrasound is a method comfortably and simply accessible as in contrast to the MRI and CT equipment, limitedly available in medical centers, an ultrasound head fixed on a portable helmet is sufficient for the tests so the images and sound materials necessary for the analyses can be made flexibly without the speaker being adversely affected by any harmful radiation.

The determination of the boundaries of objects may form the basis of the separation of objects or segmentation. Several image processing tasks are connected to segmentation, especially in medical imaging. Its use is widespread in the checking of blood vessels [5] [6] as well as in the measuring of bones [7]. The analysis of the changes in the brain [8], in the thyroid nodules [9], or the prostate gland [10] can be regarded as typical applications. The analysis of ultrasound image patches is also supported by universally applicable procedures [11] [12] [13].

With images taken for the purpose of speech processing, it is advantageous, that with ultrasound imaging, images of high resolution (almost a thousand pixels in a radial section) and high speed (80-85 images per second) can be made. Good spatial resolution is indispensable for the shape of the tongue to be displayed as sharply as possible while good temporal resolution supports the possibility of studying the rapid co-articulation changes occurring during continuous speech in a reliable way. It should not be left out of consideration, either, that ultrasound is particularly suitable for analyzing continuous speech as the time necessary for scanning the vocal tract is only a fragment of the time required by e.g. MRI imaging. However, work is made more difficult by the circumstance that in contrast to the MRI and CT images helping collect three-dimensional morphological information, ultrasound only provides information about the position of the tongue in the two-dimensional midsagittal plane so the contour of the palate and the tip of the tongue are not displayed in the image. A further technical problem is that the surface contour of the tongue should be defined with the greatest possible accuracy during post-processing, which is not a trivial task.

During our work, the processing of the ultrasound images was carried out with the help of our software written in a MATLAB environment, in the course of which an auxiliary curve was fitted on the surface of the tongue, relying on dynamic programming. The verification of contour tracking results was performed on the basis of the article by Tamás Csapó and Steven Lulich [14]. During the movements of the tongue, the position of the auxiliary curve changes dynamically, therefore such a data set was obtained the elements of which varied in both space and time.

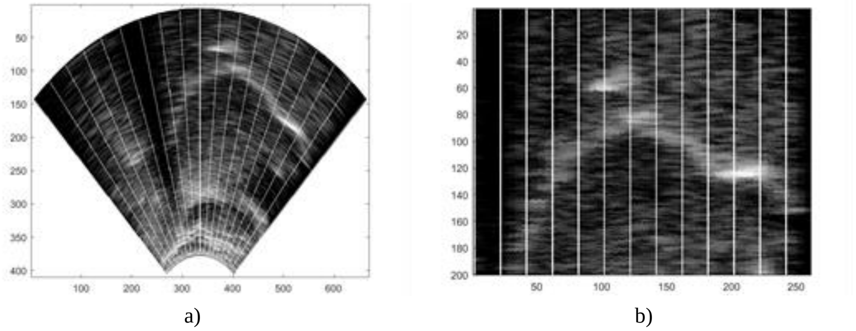
One type of procedures tracking tongue contour requires training with the contours manually marked in a large number of images, involving the application of artificial intelligence methods. The number of training images is e.g. 5,000 [15] or 700 radiographs and 400 ultrasound images [16]. Mozzafari et al. use 80% of the ultrasound frames for training [17] [18]. The accuracy of solutions not requiring any training usually falls behind of methods that requires machine learning. We have developed such a tongue contour tracking procedure not requiring training, the accuracy of which is competitive with that of the procedures requiring it.

2 The Method of Tongue Contour Tracking

An algorithm based on dynamic programming, has been developed herein, to track the movement of the back of the tongue. In the ultrasound image, tongue contour is created by the radiation reflected at the boundary of the tongue and the air above it. The curve of the back of the tongue is detected at the lower boundary of the bright band, thus obtained. The analysis requires several steps. First, the preprocessing of the image is completed. The extreme size Prewitt kernel, addresses the problems of break of discontinuity and noise, simultaneously. In the image obtained after filtering, the curve which has a maximum accrued brightness from the left to the right edge of the image is identified. The points of the curve thus obtained, follow the uneven line of the tongue contour. To smooth the curve, filtering based on Discrete Cosine Transformation (DCT) is applied. The smoothing of the curve improves the continuity of the contour and makes comparison possible with the results of manual contour detection.

2.1 Preprocessing Steps

As a first step, the image is resampled forming radial sections starting from the center of a circle. By arranging the sections thus obtained in a Cartesian column graph, a matrix is gained. Sampling performed by quarter degree has been arrived at through experience so that there should not be a bigger difference than two pixels between the neighboring columns in the contour. Figure 1 a) shows the detection of the center of the circle and the line of the radial sections. For the sake of clarity, sections are only represented by five degrees. The matrix obtained after sampling is shown in Figure 1 b).



a)

b)

Figure 1

a) Radial ultrasound image

b) Column graph after resampling

Moving top-down, in a column of the image, a falling edge is searched for, where brightness decreases. In image processing, search for edges is made through examining the variation of the brightness of pixels and developing differences. With noisy images, however, enhancing differences leads to the improvement of noise. Noise removal and search for edges may be performed in one step in the way that the difference in brightness is not developed pixel by pixel but brightness is averaged for a bigger patch – thus noise reduction is performed – and the difference is developed for another patch of similarly averaged brightness.

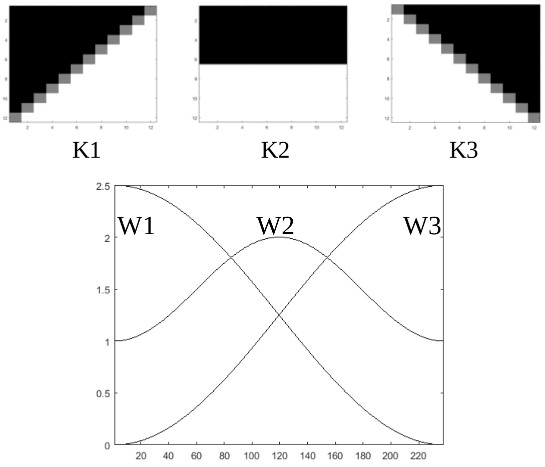


Figure 2

Prewitt kernels performing averaging and edge detecting for edges of different directions (above). Weighting functions of tri-directional edge enhancement (below).

In the top part of Figure 2, 12x12 pixel Prewitt kernels can be seen. Black stands for -1, white for +1 and grey for 0. The kernel performs averaging on the patch covered by the white area and then subtracts the mean of the patch marked with black from it. (The 1/144 multiplier to be applied for leaving the dynamic range unchanged is not included in the figure.) By applying the kernel (convolution), the difference between the averaged brightness of the two patches is obtained for a pixel of the image. Among the edge enhancement procedures, the Prewitt kernel proved to be the most effective in simultaneously handling noise and discontinuity. It can be observed in the ultrasound images, that on the left side of the image, the edges of the tongue contour are close to the direction of the minor diagonal of an imaginary square matrix, they are horizontal in character in the center of the image while on the right side of the image, they approximate the direction of the main diagonal. For each column, the result of the convolution performed with three kernels is weighted with the functions that can be seen in the bottom part of Figure 2. (W1: quarter-wave $\cos^2$, W2: full-wave raised -cos, W3: quarter-wave $\sin^2$. We have not examined any other weighting functions.) As a result, on the left side of the image, the edge enhancement performed with kernel K1, in the middle that performed with kernel K2 and on the right side that performed with kernel K3 is given greater weight. In the three weighted edge enhancement matrixes, the maximum is searched for each pixel, which is regarded to be the edge marker for each pixel. In comparison with W2, the weighting of W1 and W3 has been developed as a result of optimization.

After resampling, it is useful to apply vertical offset with the mean of the tongue contour measured in a lot of images in each column (Figure 3 a)). This step has a double role.

1) The slope range of the edges decreases and thus the efficiency of edge enhancement improves.
2) The area of the tendon, which might often lead to false edges, is shifted out of the image (see the thin white patch in the bottom third of image a) Figure 3).

The white band that can be seen at about row 350 of part a), Figure 3 is caused by the reflection of the tendon.

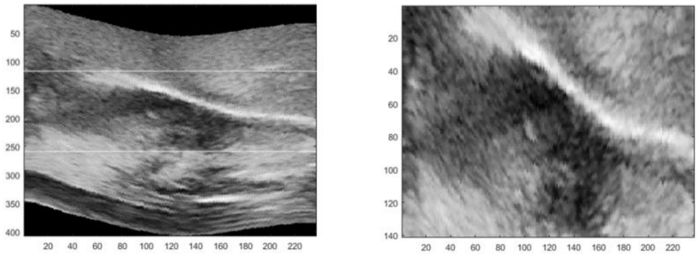


Figure 3

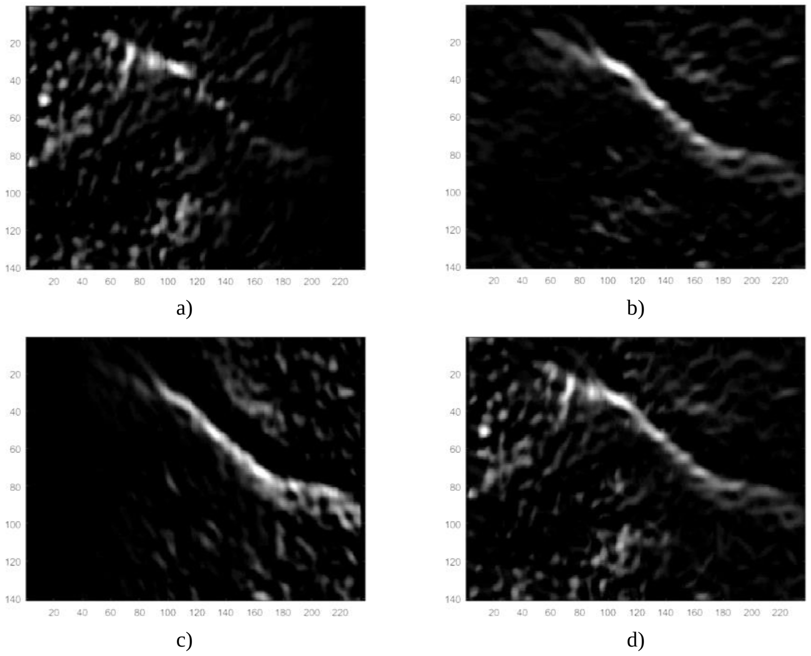
a) Offset column graph

b) The area to be processed

After this, it becomes possible to detect the region of interest (ROI) (Figure 3 b). The axes of the figure indicate the row number of the pixels and the serial number of the column.

Performing convolution with the kernels in Figure 2 after the application of the relevant weighting, the edge enhancements distinguished by direction in Figure 4 are obtained. In Figure 4 a), weighting enhances the left side of the image suppressing the right side and reinforcing the edges close to the direction of the minor diagonal. In Figure 4 b), the center of the image is enhanced, keeping the edges of the horizontal direction at the edge of the image, as well. In Figure 4 c), the weighting reinforces the edges characteristic of the right side of the image, close to the direction of the main diagonal, attenuating the false edges that may be obtained on the left side of the image. Figure 4 d) shows the edge marker derived by forming the maximum pixel by pixel.

In part d) of Figure 4, especially in the third of the image on the left, the edge marker does not show continuity. We may improve continuity with a further averaging filter (11 x 11 matrix). Figure 5 shows the results.



a)

b)

c)

d)

Figure 4
The result of edge enhancement after the convolution performed with kernels

a) K1
b) K2
c) K3
d) The unified edge marker

In the image thus obtained, the curve with the highest cumulated brightness is searched for from the left to the right edge of the image.

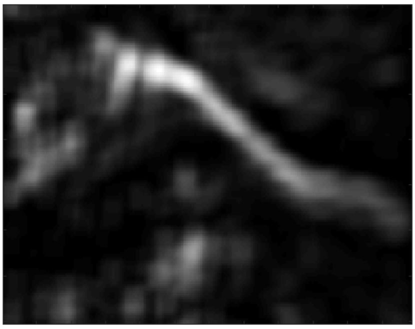


Figure 5
The image of edge enhancement with improved continuity, obtained after further averaging

The algorithm of dynamic programming has free end-to-end connectivity so it may start at any row of the first column and may end at any row of the last column. The cumulated sum should be defined for each pixel, moving from left to right. Cumulated brightness and the information from which pixel it has been accessed should be stored in each pixel. Therefore, it is examined for each pixel, which of the altogether five pixels among those in the preceding two rows and the following two rows in the previous column has the greatest cumulative brightness. You get the maximum brightness of the pixel being examined if the brightness of it is added to this sum. If you move on from the previous one or two rows of the previous column of the matrix representing the image, you go one or two rows downwards. If you obtain the maximum moving on from the following one or two rows of the preceding column, you go one or two rows upwards. You move on horizontally in the same row. If you find the maximum of the last column, the curve of the tongue contour may be reverse-engineered from this pixel. In Figure 6, the tongue contour detected in the column graph is indicated with white dots.

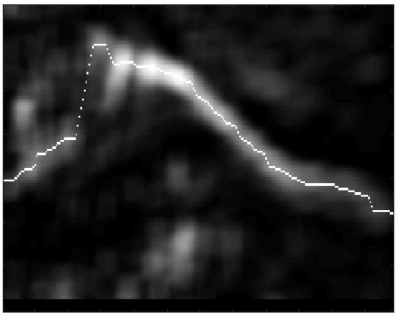


Figure 6
Uneven tongue contour obtained in the column graph

Figure 7 shows the contour projected back on the radial figure. The curve obtained follows the unevenness of the edge. The curve may be filtered with discrete cosine transformation. The smoothed curve is indicated with a line of white dots.

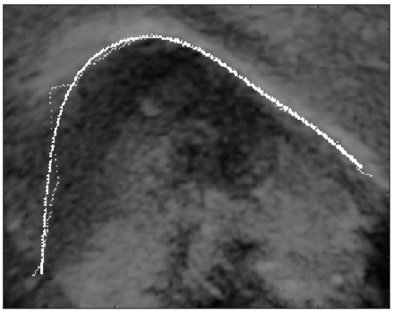
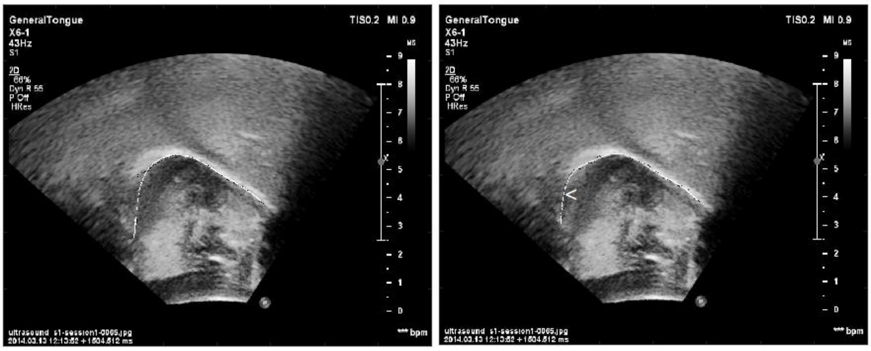


Figure 7
The uneven (thin grey) curve and the smoothed (thick, white) contour projected back on the radial image, enlarged

2.2 Parameter Initialization

In the ultrasound image, inaccuracy is not manifested in a little displacement of the edges but in some sections of the tongue contour, the search algorithm “gets lost” and finds an alien edge the brightest. Figure 8 gives examples for both correct and false edge detection, indicated the tongue contour found with a white curve.



a) b)

Figure 8

a) Correct b) False edge detection (< shift)

In our procedure, the success of edge detection is influenced by four parameters:

1) The size of the edge enhancing kernel.
2) The weighting of W1 for the edges parallel with the minor diagonal.

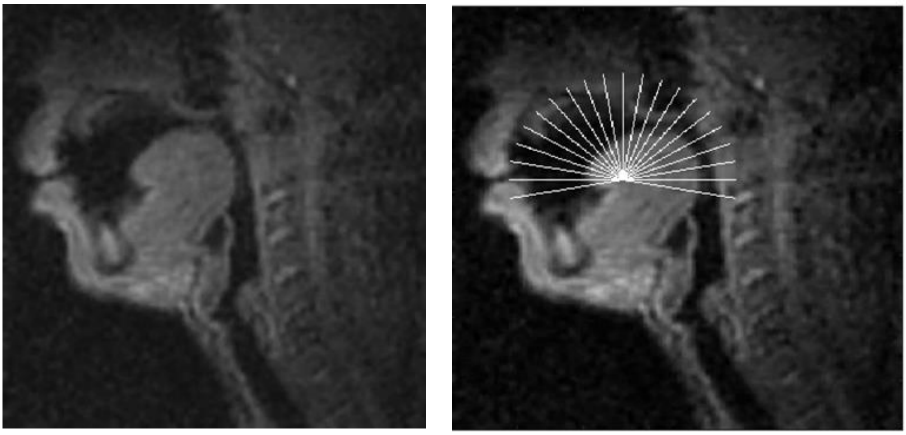
3) The weighting of W3 for the edges parallel with the main diagonal.

4) The size of the smoothing kernel.

These features can be changed independently. The right size kernel results in more spectacular weighted edges that yields the most highlighted edge after smoothing. The optimization has been performed with a comparison with manual contour detection.

2.3 Extending the Method to the Processing of MRI and CT Images

The procedure described is not only suitable for the processing of ultrasound images but the borderlines of the objects may be detected with it in images obtained with other imaging procedures, too. During the processing of the MRI and CT images, the difference is only that resampling is performed with a different center and radius, and moving upwards from the bottom, in contrast to ultrasound images it is not increasing (rising edge) but decreasing (falling edge) brightness transitions that are searched for. Part a) of Figure 9 shows the original MRI image while in Figure b), the radiuses of the radial resampling can be seen. (For the sake of traceability, radiuses are plotted by 10° while resampling is performed by 1°.)

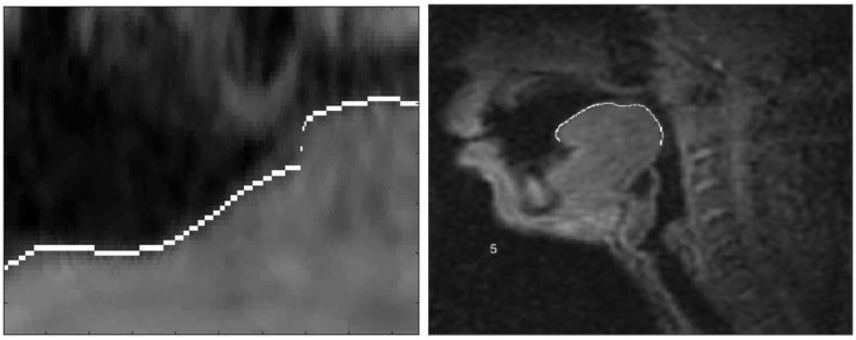


a) b)

Figure 9

a) The original MRI image b) Resampling radiuses

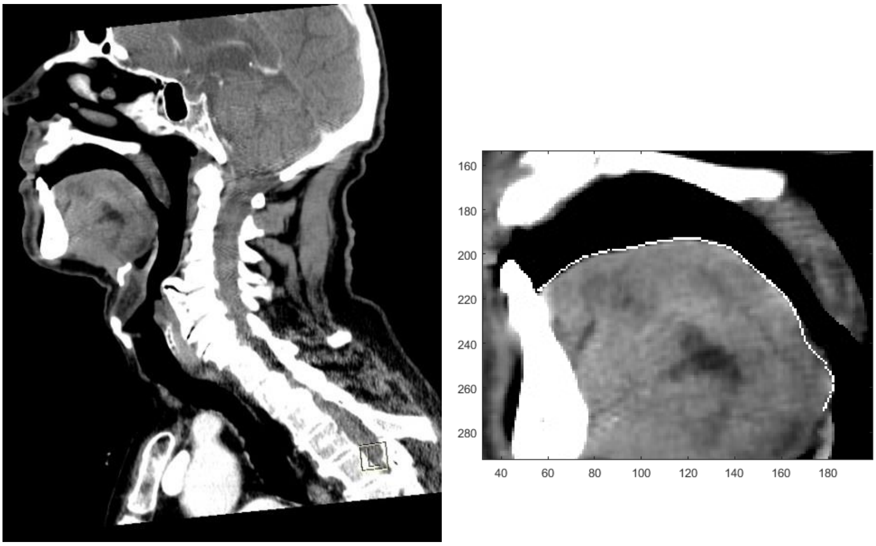
In image a), Figure 10, the resampled image is shown spread in the Cartesian coordinate system The detected contour is indicated with white dots. In Figure b), tongue contour can be traced projected back on the original image.



a) b)

Figure 10

a) The spread image with the white contour b) The tongue contour detected in the original image

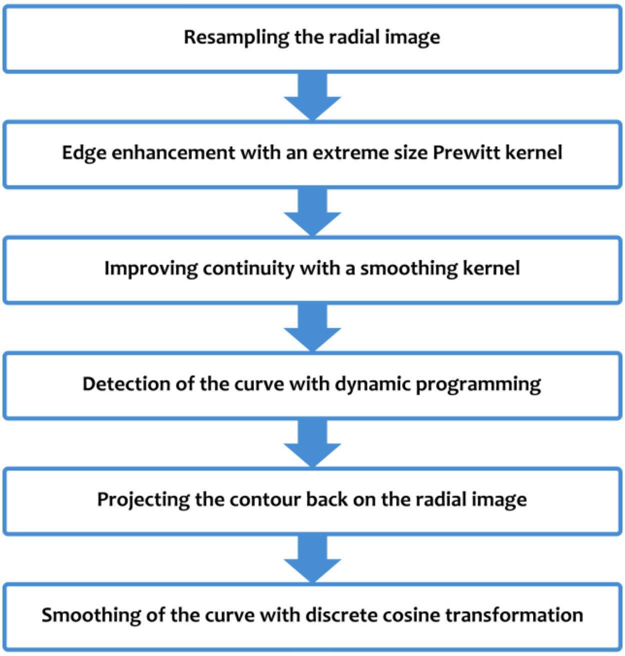


a) b)

Figure 11

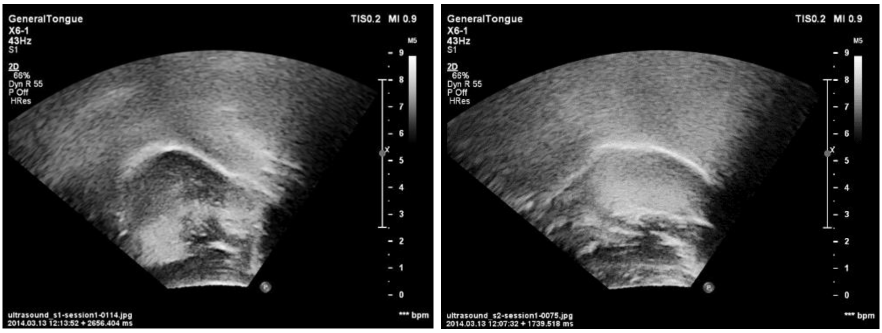
a) CT image b) The enlarged area of the tongue with the white contour

Steps of the applied contour tracking procedure:


Resampling the radial image
Edge enhancement with an extreme size Prewitt kernel
Improving continuity with a smoothing kernel
Detection of the curve with dynamic programming
Projecting the contour back on the radial image
Smoothing of the curve with discrete cosine transformation

3 Results

The verification of the results of tongue contour tracking was performed on the same set of ultrasound images on which Tamás Csapó and Steven Lulich [14] had compared contour tracking processes. Two female (F1, F2) and two male speakers (M1, M2) said the sentence ‘I owe you a yoyo’ twice, one after the other, in the recording. Figure 12 shows the variation in the tongue contour sharpness of the four speakers.



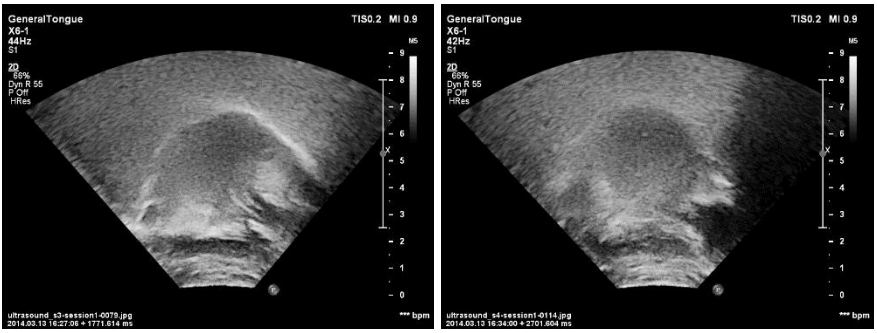


Figure 12

Ultrasound images of the four speakers in the middle of the first ‘o’ sound of the word ‘yoyo’ (From left to right: top: F1, F2, bottom: M1, M2)

It can clearly be seen that the sharpness of tongue contour is different for the four speakers. The success of manual and automatic contour tracking is strongly influenced by edge visibility. Csapó and Lulich [14] performed manual contour detection with seven volunteers. Mean Absolute Error was used to characterize manual contour detection (Table 1). The table confirms the visual qualification of the tongue contours in Figure 5. During the assessment of automatic tongue contour detection, the accuracy that can be expected depends on image quality, which can be characterized with the mean error of manual processing.

Table 1

Mean absolute error of manual contour detection (mm)

Speaker	**F1**	**F2**	**M1**	**M2**
Mean Absolute Error	0.95	1.09	1.17	2.11

Four free access methods were tested for automatic contour tracking. The best result was yielded by the AutoTrace3.5 setting. Images were divided into two parts, so that the sample sentence appeared once in every series of images. The two samples of the four speakers yield altogether eight series of images. In the case marked as AutoTrace3.5, training samples were represented by the average of the manual contour detections of seven image series and testing was performed on the eighth image series. When the speaker’s own image series was not involved in the training (AutoTrace3 but six image series from the other three speakers were used for training and the speaker’s own two image series were used for testing), the accuracy of contour tracking deteriorated drastically. Table 4 includes the Root Mean Square Error (RMSE) values for the different procedures. Averaging was performed on the logarithm of square error, and then the result was restored with an exponential function (logRMSE). Data are given in mm, in the enlargement applied in the images, 1 mm = 4.24 pixels.

In our approach, with optimizing the parameters of the proposed contour tracking algorithm, we strove for a universal setting: the four features listed in Section 2.2 are identical for all the four speakers. This is presented next.

Table 2
logRMS error in contour tracking with the universal setting (mm)

Speaker	**F1**	**F2**	**M1**	**M2**	**Mean**
logRMS error	0.66	0.88	1.10	2.13	1.19

As an example, Figure 14 shows the impact of the size of the smoothing filter, apparently influencing the accuracy of contour tracking considerably.

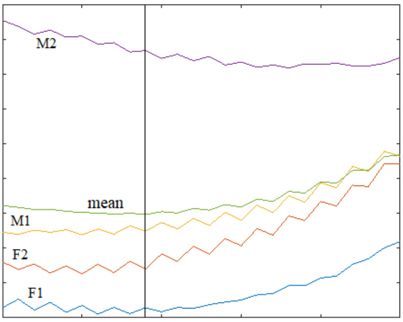


Figure 13
Mean error for each speaker and for the mean of the four speakers as a function of the size of the smoothing screen

With the setting yielding the smallest mean error, 11x11 pixels was the optimum size of the edge enhancing Prewitt kernel and 2.5 was the value of W1 and W2 weighting. As Figure 13 testifies, for the four speakers, the optimum size of the smoothing filter ensuring the smallest mean error, regarding the average of the four speakers, was 14x14 pixels (vertical line).

Performing the optimizing of parameters for each speaker after this, we obtained individual, speaker-specific settings with the speaker-specific selection of the edge enhancing kernel and the smoothing kernel. The result can be further improved by the individual setting of W1 and W2 weighting. Table 3 shows logRMS errors obtained with optimum settings for each speaker. The speaker selected for optimization is indicated in the 'Reference' column.

Table 3
logRMS error of contour tracking with the individual settings (mm)

Reference	**F1**	**F2**	**M1**	**M2**	**Mean**
F1	**0.55**	0.91	1.12	2.45	1.26
F2	0.91	**0.80**	1.05	2.65	1.35
M1	0.75	0.81	**1.02**	2.51	1.27
M2	0.79	1.17	1.31	**1.94**	1.30

As regards the optimum size of the edge enhancing kernel, there was no considerable difference between the speakers. W1 weight had to be increased for speaker F1 and W2 weight for speakers F2 and M2. For the size of the smoothing kernel, for speaker F1 the optimum proved to be 11x11 pixels, for speakers F2 and M1 10x10 pixels and for speaker M2 24x24 pixels. The optimum values of universal parameters lie between the individual optimum values.

Table 4
logRMS errors of automatic contour tracking (mm)

Speaker	**F1**	**F2**	**M1**	**M2**	**Mean**
AutoTrace3	5.85	7.06	5.59	9.94	7.11
EdgeTrak	1.95	3.46	1.89	5.15	3.11
TongueTrack	1.96	3.15	2.76	3.6	2.87
AutoTrace3.5	1.15	1.93	1.78	2.19	1.76
Proposed	**0.55**	**0.80**	**1.02**	**1.94**	**1.08**

Logarithm formation compresses the dynamic range with the weight of outstandingly high errors decreasing.

Figure 14 shows the values of Table 4 in graphic representation.

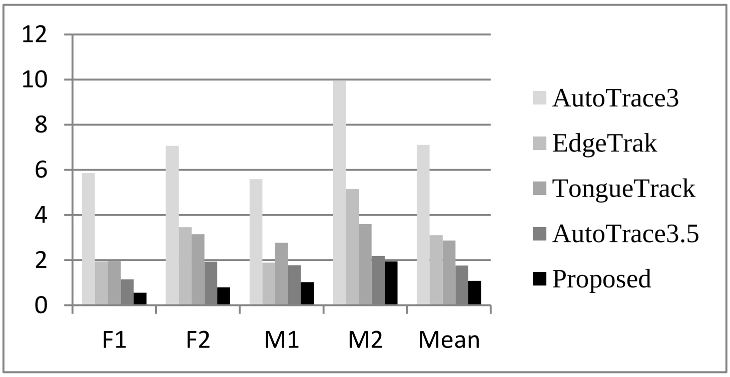


Figure 14
logRMS errors of the specific contour tracking procedures

Table 5 presents the different means of the errors in proposed contour tracking as compared with the mean of manual contour detection (logRMSE, linear RMSE, Mean Absolute Error).

Table 5
Different errors in proposed contour tracking (mm)

Speaker	**F1**	**F2**	**M1**	**M2**	**Mean**
logRMSE	0.55	0.80	1.02	1.94	1.08
linRMSE	0.63	0.99	1.24	2.36	1.31
Mean Absolute Error	0.33	0.41	0.45	0.64	0.46

It is a significant result that our tongue contour tracking algorithm based on dynamic programming yielded a smaller error than the other procedures tested by Csapó and Lulich [14]. The mean error calculated for the four speakers also proved to be the smallest with the method developed by us. In addition, with the AutoTrace3.5 application, having proved the best, contour detection requires training. Researchers have not examined the number of training samples necessary for a reliable result. Our method, based on dynamic programming, only requires the detection of the region of interest to avoid false edges.

As regards manual contour tracking, own logRMSE values for speakers F1 and F2 are better than that of each assessor. For speaker M1, 2 manual assessors out of the seven gave a better result and five worse than the ones designated as ’proposed’. With speaker M2, two assessors tracked tongue contour with a smaller while a third with identical error and four assessors achieved weaker tracking.

In lack of the set of images examined, we are unable to compare our proposed results with other published data, furthermore, the measures of accuracy are different. Mean Sum of Distance (MSD) is an often applied distance measure, especially for curve pairs with different indexing. For each point, the formula takes into account the nearest point in the other curve.

$$\mathrm{D}(\mathrm{U},\mathrm{V}) = \frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}} \min_{\mathrm{j}} |\mathrm{u}_\mathrm{i} - \mathrm{v}_\mathrm{j}| + \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \min_{\mathrm{j}} |\mathrm{v}_\mathrm{i} - \mathrm{u}_\mathrm{j}| \tag{1}$$

In calculating the absolute error in Table 6, we considered radial distances but nothing guarantees that this is also the nearest point. This way, the absolute error calculated by us cannot be smaller than MSD.

In the relevant literature, Zhu et al. [15] specify the most favorable MSD error as 0.85 mm. Xu et al. [19] obtained 0.87 mm MSD error for the most precisely tracked speaker. For ultrasound contour tracking, [20] reported an absolute error of 0.5 mm. Without providing the distance measure, Tang et al. [20] tracked tongue contour with a 3 mm mean error. Mozaffari et al. [17] published 0.91 and [18] 0.61 mm MSD errors. Jaumard-Hakoun et al. [22] reported a 0.67 mm MSD error. Roussos et al. [16] gave the RMS error of contour tracking in a graphic form, which shows a minimum 1.5 mm as can be read in the relevant figure.

The description of tongue contour and the assessment of its data, cannot yet be regarded to be completely elaborated [23]. The results closest to the mean of manual detection were obtained by using the first seven coefficients of the discrete cosine transformation used for contour smoothing. Consequently, the complete tongue contour can be characterized with seven data but it is difficult to connect DCT coefficients with the geometric data. The detailed analysis of tongue contours requires further investigations.

Conclusions

In the field of speech processing, it is an immense advantage of the ultrasound imaging systems, of being suitable for tracking rapid movements, image and sound synchronization and it presents only a minimum inconvenience to the speaker, who is also not affected by any harmful radiation. Its disadvantage is that it does not provide a full three-dimensional image, but only shows either a longitudinal or transverse section and the tip of the tongue cannot be seen. The longitudinal (midsagittal) section is the most suitable for tracking tongue movements.

In the course of preprocessing, we simultaneously cope with the problems of edge enhancement and noise removal, which are two contradicting required goals. The parameters of the pre-processing steps are of decisive importance. With their appropriate selection, contour tracking accuracy, surpassed all the other procedures investigated. Performance can be improved further with the application of speaker-specific settings. The results show that the analysis of the articulation on the basis of ultrasound images not only makes it possible to define the static data of tongue position but also offers the opportunity to perform a dynamic description of tongue movements [24].

Ultrasound and MRI tongue contour analysis has paved the way for the designation of a new research direction. We explore the possibility of combining the benefits of the two imaging methods: Good spatial and temporal resolution of the ultrasound recordings and the three dimensional representations of the tongue, by MRI imaging.

Acknowledgements

The described study was carried out as part of the EFOP-3.6.1-16-00011 “Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialization” project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

We would also like to thank the Lingual Articulation Research Group (Hungarian Academy of Sciences), for providing the recordings with the SonoSpeech system, and Tamás Gábor Csapó for the consultation on their method and providing the series of images and the data set they used.

References

[1] Barnaud, M-L., Schwartz, J-L., Bessière, P., 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 pp. 1-34

[2] Csapó, T. G. (2020) Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract, In Proc. Interspeech 2020, 2722-2726, DOI: 10.21437/Interspeech.2020-0015

[3] Gábor Gosztolya, Tamás Grósz, László Tóth, Alexandra Markó, Tamás Gábor Csapó, (2020) Applying DNN Adaptation to Reduce the Session Dependency of Ultrasound Tongue Imaging-Based Silent Speech Interfaces', Acta Polytechnica Hungarica, Vol. 17, No. 7, pp. 109-124, 2020

[4] Douros, I. K., Kulkarni, A., Dourou, C., Xie, Y., Felblinger, J., Isaieva, K., Vuissoz, P., Laprie, Y. (2020) Using Silence MR Image to Synthesise Dynamic MRI Vocal Tract Data of CV. Proc. Interspeech 2020, 3730-3734, DOI: 10.21437/Interspeech.2020-1173

[5] Liang, Q; Wendelhag, I; Wikstrand, J; Gustavsson, T. (2000) A multiscale dynamic programming procedure for boundary detection in ultrasonic artery images. IEEE Trans Med Imaging 2000, 19, 127-142

[6] D. Cheng and X. Jiang, (2008) Detections of Arterial Wall in Sonographic Artery Images Using Dual Dynamic Programming, in IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 6, pp. 792-799, Nov. 2008, doi: 10.1109/TITB.2008.926413

[7] Pezhman Foroughi, Emad Boctor, Michael J. Swartz, Russell H. Taylor, and Gabor Fichtinger (2007) Ultrasound Bone Segmentation Using Dynamic Programming. IEEE Ultrasonics Symposium. New York, NY, USA, 28-31 Oct. 2007, pp. 2523-2526

[8] Ahmad S, Wu Z, Li G, Wang L, Lin W, Yap PT, Shen D. (2019) Surface-constrained volumetric registration for the early developing brain. Med Image Anal. 2019 Dec; 58:101540. doi: 10.1016/j.media.2019.101540. Epub 2019 Aug 1. PMID: 31398617; PMCID: PMC6815721

[9] Liu T, Guo Q, Lian C, Ren X, Liang S, Yu J, Niu L, Sun W, Shen D. (2019) Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. Med Image Anal. 2019 Dec; 58:101555. doi: 10.1016/j.media.2019.101555. Epub 2019 Sep 5. PMID: 31520984

[10] Ghavami N, Hu Y, Gibson E, Bonmati E, Emberton M, Moore CM, Barratt DC. (2019) Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. Med Image Anal. 2019 Dec; 58:101558. doi: 10.1016/j.media.2019.101558. Epub 2019 Sep 11. PMID: 31526965

[11] T. McInerney and D. Terzopoulos, (1996) Deformable models in medical image analysis: a survey, Med. Image Anal., Vol. 1, No. 2, pp. 91-108

[12] Jingfeng Jiang and Timothy J. Hall (2009) A Generalized Speckle Tracking Algorithm for Ultrasonic Strain Imaging Using Dynamic Programming. Ultrasound Med Biol. 2009 November; 35(11): pp. 1863-1879

[13] Kathrin Ungru, Xiaoyi Jiang (2017) Dynamic Programming Based Segmentation in Biomedical Imaging. Computational and Structural Biotechnology Journal 15, pp. 255-264

[14] Csapó T. G., Lulich S. M. (2015) Error analysis of extracted tongue contours from 2D ultrasound images. Interspeech, pp. 2157-2161

[15] Jian Zhu, Will Styler, Ian C. Calloway (2019) A CNN-based tool for automatic tongue contour tracking in ultrasound images. ArXiv 2019, Vol. 1907.10210

[16] A. Roussos, A. Katsamanis and P. Maragos, (2009) Tongue tracking in Ultrasound images with Active Appearance Models, 16th IEEE International Conference on Image Processing (ICIP), Cairo, 2009, pp. 1733-1736, doi: 10.1109/ICIP.2009.5414520

[17] Mohammad Hamed Mozaffari, Shuangyue Wen, Nan Wang, Won-Sook Lee: (2019) Real-time Automatic Tongue Contour Tracking in Ultrasound Video for Guided Pronunciation Training, 14th International Conference on Computer Graphics Theory and Applications, Prague, Czech Republic

[18] M. Hamed Mozaffari, Md. Aminur Rab Ratul, Won-Sook Lee: IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision, arXiv preprint arXiv:1911.03972

[19] Xu K, Yang Y, Stone M, Jaumard-Hakoun A, Leboullenger C, Dreyfus G, Roussel P, Denby B. (2016) Robust contour tracking in ultrasound tongue image sequences. Clin Linguist Phon.; 30(3-5):313-27, doi: 10.3109/02699206.2015.1110714. Epub 2016 Jan 20. PMID: 26786063

[20] Changsheng Yang, Maureen Stone (2002) Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances. Speech Communication 38, pp. 201-209

[21] Tang, L., Bressmann, T. & Hamarneh, G. (2012) Tongue contour tracking in dynamic ultrasound via higher-order mrfs and efficient fusion moves. Medical Image Analysis, 16(8), pp. 1503-1520

[22] Aurore Jaumard-Hakoun, Kele Xu, Pierre Roussel-Ragot, Gérard Dreyfus, Maureen Stone, et al. (2015) Tongue contour extraction from ultrasound images based on deep neural network. The International Congress of Phonetic Sciences, Aug 2015, Glasgow, United Kingdom ⟨hal-01366237⟩

[23] Fabre, D., Hueber, T., Girin, L., Alameda-Pineda, X., Badin, P. (2017) Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. Speech Communication, Vol. 93, pp. 63-75S

[24] Zhao, Lu, Czap, László (2019) Visemes of Chinese Shaanxi Xi’an Dialect Talking Head, Acta Polytechnica Hungarica, Vol. 16, No. 5, pp. 173-193