Possible Methods for Combining Tongue Contours of Dynamic MRI and Ultrasound Records

Réka Trencsényi[1] and László Czap[2]

[1]Department of Electrical and Electronic Engineering, Institute of Physics, Faculty of Science and Technology, University of Debrecen, Bem tér 18/a, 4026 Debrecen, Hungary, trencsenyi.reka@science.unideb.hu

[2]Institute of Automation and Infocommunication, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Egyetemváros, 3515 Miskolc, Hungary, czap@uni-miskolc.hu

Abstract: One of the trends of the current generation of machine speech, is articulatory speech synthesis, that is based on the processing of visual and geometric information, related to voice production. Accurate knowledge of the static and dynamic geometric parameters of the vocal organs, plays a fundamental role in the realization of speech synthesis. Appropriate sources of visual extraction of these data can be MRI and ultrasound (US) records made during speech, which can be described by different geometries. Harmonization of the geometries of MRI and US frames is not a trivial task. In this publication, we present one possible method for the transformation between the two sources. The starting point of the transformation process is formed by tongue contours obtained by automatic algorithms. Beyond this exact method, we also follow statistical procedures, by applying machine learning to interconnect MRI and US records.

Keywords: articulatory speech synthesis; tongue contour tracking; machine learning; dynamic MRI and US records; harmonization of MRI and US sources

1 Introduction

Speech synthesis is one of the most dynamically developing fields in speech research, with ever more complex technical and methodological challenges, which even today, forms an integral part of the human-machine relationship. In this regard, the communication role of the machine is crucial, since its basic designation is the implementation of text-to-speech transformation, i.e. the realistic imitation of the acoustic product forming during natural human speech. In the extended version of this, the model can be further refined by taking into account the supra-segmental elements of speech (rhythm of speech, voice level,

pitch, tone, intonation, stress), which can have high importance in the domain of speech recognition, as well [1]. Currently, research is ongoing, in the field of speech synthesis, with focus on the creation and improvement of text-to-speech systems, that allow the spread of such applications as, e.g. passenger information systems, speaking smart devices, belletristic readers, screen readers, sound weather forecast or telephonic directory enquiry services. In the case of text-to-speech readers representing the traditional trend of researches, speech construction occurs by direct or indirect utilization of human voice samples. The success of these endeavors is proven by numerous publications of the literature [2-7] which report on speech synthesis based on different speech databases or corpuses, in the case of Hungarian, German, or multilingual synthesizers. In addition to the classical concepts, there are also such fields starting to evolve, which are less elaborate and many open problems are still expected to be solved. For instance, articulatory [8-9] or machine-learning-based speech synthesis [10-11] can be classified here.

Articulatory speech synthesis, instead of the application of human voice samples, tries to implement the imitation of the acoustic product by machine imaging of human voice production and articulation. One of the modern technological streams of this is the experimentation trending to the articulatory electromechanical speech generators needed for the production of speech of robots [12] [13]. The starting point of synthesis is the execution of articulatory-acoustic conversion that is built upon visual information relating to speech [14]. Consequently, different imaging procedures (e.g. Magnetic Resonance Imaging (MRI), Computer Tomography (CT), Ultrasound (US)) have essential roles, which supply new information channels in the process of scientific research. Accordingly, MRI or US records made during speech can be potential sources of visually supported extraction of the parameters describing human articulation. Since most actively the tongue takes part in voice production, it is expedient to monitor primarily the motion of the tongue as accurately as possible. In recent years, besides the mentioned MRI, CT, and US, popular tools of the investigations are electropalatography (EPG) or electromagnetic articulography (EMA). Applying the simpler accessible US, EPG, and EMA procedures, information about the dynamic features of speech can be obtained mostly along certain plane sections, although three-dimensional US technique is available, as well, which provides information in multiple planes [15]. Nevertheless, by dint of MRI and CT equipment demanding clinical conditions, three-dimensional morphological data can be acquired. Recently, several studies have dealt with elaboration and development of dynamic tongue contour tracking algorithms [16-18], which can form one of the keystones of research performed in the topic of articulatory speech synthesis. Dynamic scanning of the tongue contour is worth doing in the sagittal plane, where the up-down and forward-backwards motion of the tongue is visible in a two-dimensional section. The most convenient tools of the investigations can be US and MRI records, the advantage being the good spatial and temporal resolution, the ability for synchronization of the image and sound materials, and

the protection of the speaker from harmful exposure. Designation of the tongue contour can be done manually or by automatic algorithms, though hundreds or even thousands of frames, creating a given record to justify the preference of dynamic programming against manual operations. The precision of tongue contour fitting is largely determined by the quality of the record and the type of contour tracking algorithm, thus, the ambition for refinement of image processing and tongue contour tracking is still a key task for research.

Beyond this, the application of machine learning algorithms designates an important direction, during which the machine produces output results from the set of certain input parameters, based on information gained from the environment, while it learns and improves performance. Machine learning algorithms try to imitate the behavior of the human brain, so the knowledge and realistic modelling of the operation of neural networks plays a key role. Biological neural networks realize a learning process based on different patterns, which can be mapped by creating appropriate algorithms, in the case of machine learning. In the field of speech synthesis, the set of input parameters of the machine can be formed by, for example, human voice samples or data retrieved from visual sources, which performs the training and the auditory product can be vocalized. Thus, the possibility of neural networks, trained by visual information, offers the linking of methods of articulatory speech synthesis and machine learning in a natural way. Opportunities are actually unlimited, and the procedures and their combinations are mostly, as of yet, not revealed fully.

Our work herein examines the transformational relationships between the geometries of US and MRI frames and the simultaneous application of tongue contour tracking and machine learning algorithms.

2 US and MRI Frames

2.1 Starting Points

Our current research focuses on the simultaneous analysis of US and MRI records made during speech, that can facilitate the visually supported complex retrieval of the static and dynamic parameters that describe human articulation. The MRI records were selected from the free-access multimedia package, on the website of the University of Southern California, the US records were available in the form of audiovisual materials created by the Micro system of the Lingual Articulation Research Group of the Hungarian Academy of Sciences and Eötvös Loránd University [19]. The dynamic moving images can be decomposed into static frames, as a result of that, the subsequent moments of speech generation can be studied step by step. Figure 1 presents a US and an MRI frame which visualize

tongue positions corresponding to sound *k* arising from a female and a male speaker.

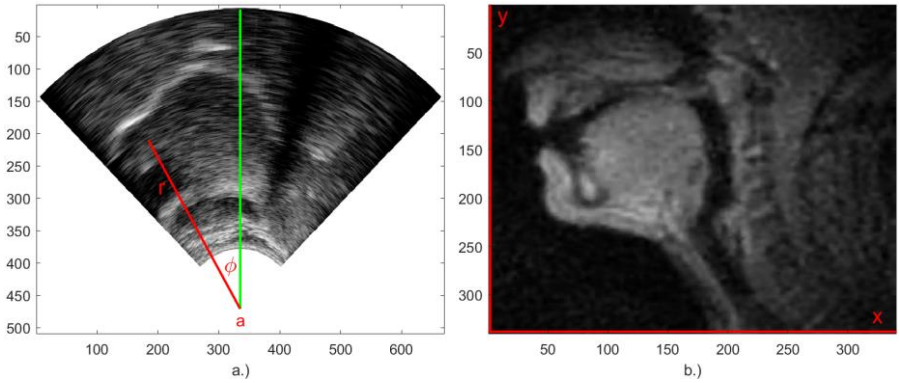

Figure 1

The side-view position of the tongue in a US (a) and an MRI (b) frame described by radial (a) and Cartesian coordinates (b)

The records display the region of the oral cavity in the sagittal plane dividing the human body into left- and right-hand parts, so in a two-dimensional section, the up-down and back-forth motions of the tongue become seeable. In the US record, the tongue contour appears as a bright band that is produced by the US waves reflected at the boundary of the tongue and the air above and the line of the edge of the tongue can be traced at the lower border of the bright band. Since the hyoid bone and the mandible partially shield the US waves, the US transducer is not capable of probing the region of the oral cavity entirely. This deficiency shows up in the form of a dark band emerging on the left and right sides of the image, at the front and rear parts of the tongue that hides the movement of the tongue root and the tongue tip, thus, in contrast to MRI records displaying the total region of the oral cavity, only partial information can be obtained about the shape and movement of the tongue. The fact can be noted as a further difference that the contour of the palate cannot be identified in the US frames, while in the MRI frames, the contour of the hard palate can be determined with sufficient accuracy, and also, the movement of the soft palate can be detected. In Figure 1, it can be observed that the US frames are spread in a zone covered by a sector of a circle, so the two-dimensional polar coordinates can be conveniently applied by the description of the position, of each pixel. These coordinates can be defined by radius **r** measured from center **a** of the circle and angle $\boldsymbol{\phi}$ relative to the vertical symmetry axis of the image. Thereby, the location of a pixel, taken in the plane of the frame, is determined by the pair of coordinates $(\mathbf{r},\boldsymbol{\phi})$ unambiguously. In the case of the used US frames, the value of angle $\boldsymbol{\phi}$ can change between -45° and 45°. However, the most comfortable frame of reference needed for the treatment

of MRI frames can be given by a two-dimensional Cartesian coordinate system, in which, the position of the designated point of the frame is fixed by the pair of coordinates (**x**, **y**). One of the aims of research is to harmonize the radial and rectangular arrangements of US and MRI records, by finding the appropriate geometric transformations.

Geometric transformations can be realized through the conversion of the relevant anatomic contours of US and MRI records. These curves can be obviously given by the tongue and palate contours, since in the dynamic description of articulation, the change of relative positions of the surface of the tongue and the palate plays an essential role in the region of the oral cavity. Hence, these examinations provide the most accurate data concerning tongue and palate contour required.

For determination of the contour of the edge of the tongue we developed and improved automatic tongue contour tracking algorithms, based on dynamic programming. The primary aim of tongue contour tracking, is the dynamic description of tongue positions, belonging to different speech sounds, and the investigation of tongue movements characterizing sound transitions created during co-articulation. Besides the qualitative analysis, the tongue contour can also be a good starting point for the quantitative study of speech, since the numeric values derived from tongue contour, can support the deeper understanding and development of articulatory models. Algorithms elaborated for detection of the tongue contour can be extremely diverse depending on the applied procedures. The edge of the tongue is drawn as a bright band in US records, while in MRI records it can be experienced as a contrast coming into existence between the dark domain of the air in the oral cavity and the bright domain of the tongue tissue, so contour tracking means the search for the pixels at the boundary of the dark and bright domains, determining the line of the edge of the tongue, in both cases. Using our approach, the application of our algorithm is preceded by the preprocessing of records, that tends to cancel the noise and discontinuities, resulting from imaging techniques. The most effective instruments of reducing the mentioned errors are edge-enhancement and averaging operations, that mathematically can be realized by convolution [20]. The found pixels of maximal brightness, adjusting to the uneven line of the edge of the tongue, produce a rough curve, the smoothing of that can be solved by a discrete cosine transformation. The images of Figure 2 show automatically fitted tongue contours in an MRI (a) and (b) and US (c) and (d) frames, respectively. In Figure 2a, the tongue position belonging to sound ***o*** can be observed, while Figure 2c renders the tongue position corresponding to sound ***ɔ*** by highlighting the smoothed tongue contour. In Figures 2b and 2d, the magnified details of the unsmoothed tongue contours drawn in frames 2a and 2c, can be seen.

Figures 2b and 2d can be created by a special transformation starting from Figures 2a and 2c.

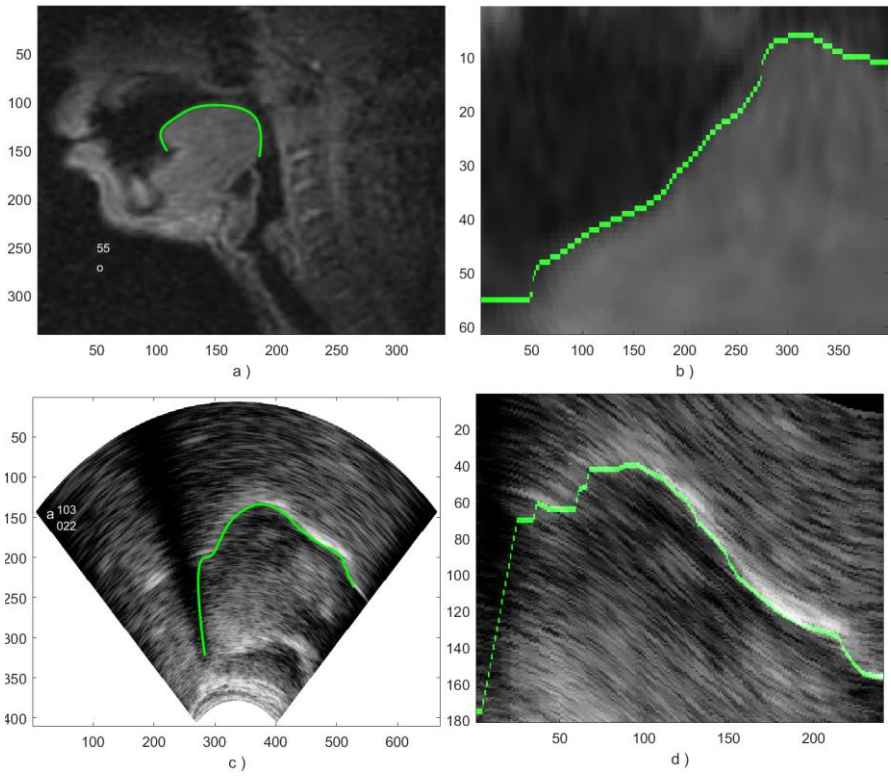

Figure 2

Automatically fitted tongue contours in MRI (a, b) and US (c, d) frames, showing also the magnified details of the unsmoothed tongue contours (b, d)

The substance of the transformation procedure is illustrated by dint of the US frame seen in Figure 3. As a first step, in Figure 3a of radial geometry, originating from the center of the circle, radial sections are formed in the range -45° and +45° defined by the record. Along these sections, the image is practically resampled. The sections produced in this manner are arranged in columns, resulting in such an image matrix, that most conveniently can be described in the Cartesian **x**-**y** plane. Figure 3b is generated on the track of shaping the matrix structure. Investigations show that sampling performed by 1/4° is the ideal, since this time, a change in the contour, greater than two pixels, does not occur between adjacent columns of the matrix. For the sake of clarity, the sections are depicted only by 5° that are demonstrated by the white lines in Figure 3. The procedure works in the case of MRI frames in a similar way, by applying the center and angular domain (usually wider than the range -45° – 45°) designated in the MRI frame properly.

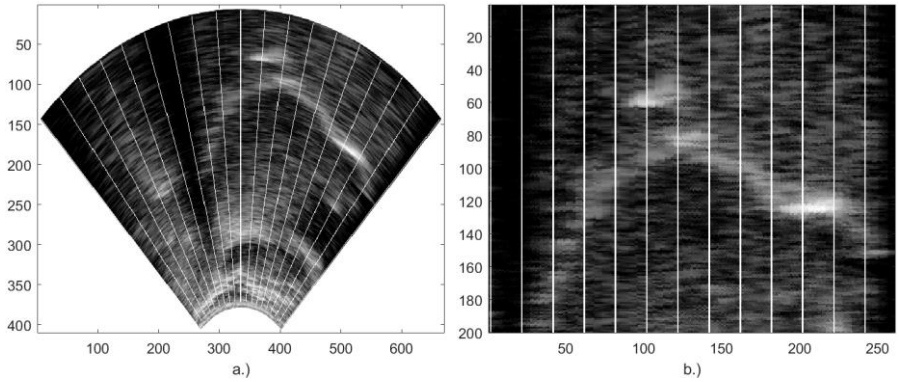

Figure 3
Some radial sections (drawn by white lines) of US frames in the original (a) and in the matrix-structured transformed plane (b)

The speaker for the MRI records is a native American English male speaker (John Esling), who vocalizes series of sounds of type VCV with vowel V and consonant C. In the US records, series of sounds arising from native Hungarian and Chinese female speakers are recorded, which are of CVCV, CVC, and VCV structures. All of these speakers are young adults, and the Chinese participant speaks in a Shaanxi Xi'an dialect. According to the presented frames of Figure 2, the obtained curves follow the line of the edge of the tongue authentically.

2.2 Geometric Transformations

Due to the screening effect of the hyoid bone and the mandible, US images are able to visualize the movement of the tongue only partially, that leads to a more confined data set regarding the position of the tongue compared to MRI frames. Since the production of a more extended parameter set from a narrower one, is much more challenging than the reverse, we specified the contours of US records, as the base of transformations.

As mentioned above, in addition to the tongue contour, the curve fitted to the palate plays a key role in the examination of articulation. Hence, before implementation of the transformation, the palate contour is needed to be ascertained in US frames. It is not a trivial task, because it cannot be revealed immediately in the US records. The location of palate, however, can be given via estimation by presuming the boundary of the tongue and palate by selecting the points being in the highest positions, and touched by the surface of the tongue during articulation. This requires, of course, the investigation of such consonants during the utterance of that the tongue surely touches the hard or soft palate. This condition is fulfilled automatically in the case of the available US package containing various audio items, since, during the articulation of consonants being

present in the recorded sentences, the tongue comes into contact with the palate at different places. We implemented the drawing of the contour of the palate essentially by the solution of an extremum search problem, the result of that is presented by the red curve of Figure 4, and the tongue contour belonging to the frame is demonstrated by the green curve.

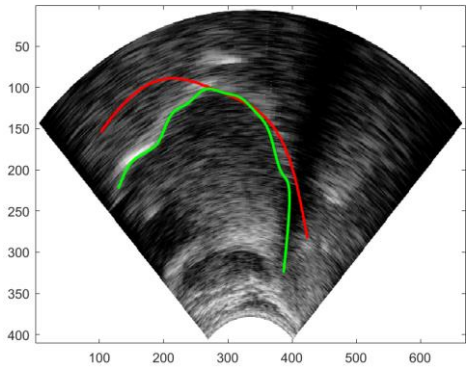

Figure 4
Tongue contour (green) fitted to the surface of the tongue and palate contour (red) arising from an extremum search problem in the case of US frames

For the transformation of the curves of Figure 4, we searched for such a reference point that can be identified with convincing certainty in the US and MRI frames, as well. We defined this point at the peak of the epiglottis, the position of that is marked by the red circle drawn in the images of Figure 5.

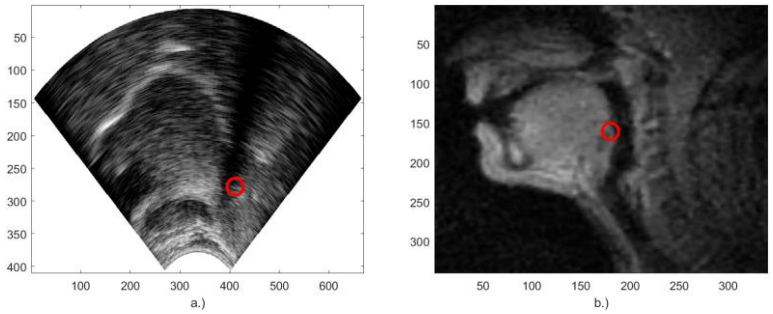

Figure 5
The peak of the epiglottis localized by red circles in the US (a.) and MRI (b.) frames

In the course of the visual study of dynamic US records we concluded that, during articulation of certain sounds, the tongue is pulled back insomuch that it touches the epiglottis. Hence, we designated the peak of the epiglottis as the starting point of the palate contour, and we determined the angular range covered by the tongue contour belonging to the selected sound *k*, which is limited by the values -39.6° and 19.4°. We performed the transformation of the curves of the tongue and palate

contour in the polar coordinate system by scaling the radial and angular range given by the pairs of values (**r**, $\boldsymbol{\phi}$) describing the points of the curves, and by shifting the initial angle $\boldsymbol{\phi_o}$ of the angular range according to the formulas

$$r' = Rr$$

$$\varphi' = FI\,\varphi$$

$$\varphi_0' = \varphi_0 + FIKORR \tag{1}$$

The scale factors R and FI of relationships (1) enable the normalization of the radial and angular range, and the term FIKORR is responsible for the rotation of the angular range. By fixing the values R=0.31, FI=1, $FIKORR$=12.6°, it is allowed to transplant the tongue and palate contours to the MRI frame. According to Figure 6, the tongue and palate contour fit to the MRI frame in an acceptable way, where the angular range of the tongue contour extends between the values -27° and 32°. Ultimately, the radial geometry of US frames is embedded into the rectangular geometry of MRI frames by the transformations (1) executed in the system of polar coordinates.

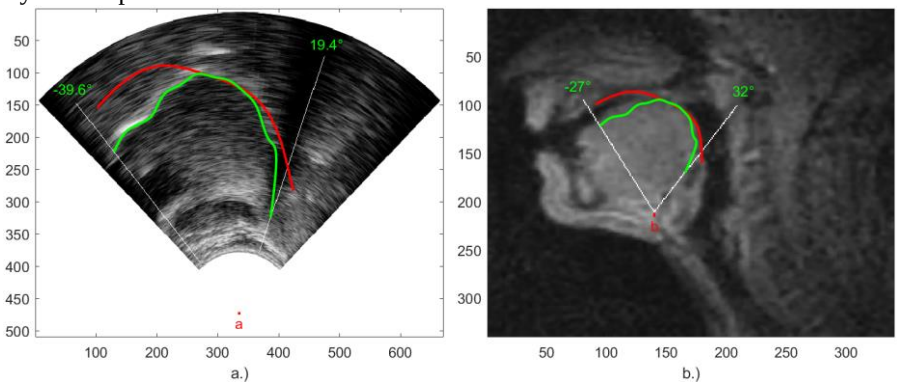

Figure 6

The angular range of the fitted (a) and transformed (b) US tongue contours drawn in the US (a) and MRI (b) frames, where the US palate contour is presented by red curves

By means of the transformation, the biunique correspondence of the points of tongue contours fitted to the US and MRI frames becomes possible that can be traced by dint of Figure 7. Due to the factor FI=1, the transformation is isogonal, therefore, the four contour points specified by the four inner radial sections selected in the US frame can be mapped along the same four radial sections drawn in the MRI frame to the MRI tongue contour illustrated by the blue curve. Thus, passing along a given section, two points can be found on the green and blue curves that can be assigned in pairs unambiguously.

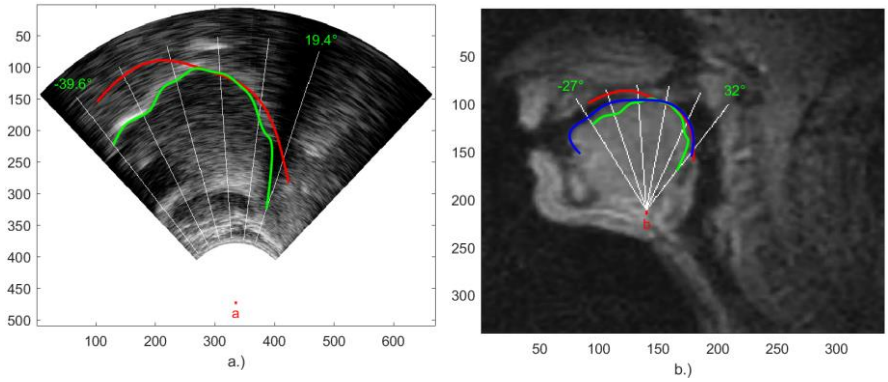

Figure 7

The biunique correspondence of the points of the transformed US tongue contour (green) and the fitted MRI tongue contour (blue) along the designated radial sections drawn by white lines in the US (a) and MRI (b) frames

The transformation can be realized in the reverse direction, as well, which means that a contour of an MRI frame can be projected to a US frame. For this purpose, the inverse transformations of (1) should be applied in the form of

$$r = r'/R$$

$$\varphi = \varphi'/FI$$

$$\varphi = \varphi_0' - FIKORR \qquad (2)$$

Using the transformations of (2), the tongue and palate contour of an MRI frame can be transferred to the appropriate US frame, as it is exemplified by Figure 8, where the tongue contour of Figure 7b is projected. The transformed curves accurately demonstrate those sections of the tongue and hard palate which do not appear in the US record because of the screening effect of the mandible.

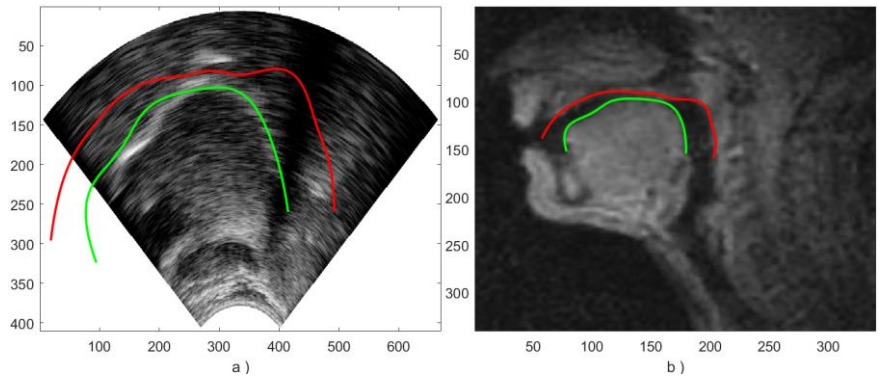

Figure 8

The fitted (b.) and transformed (a.) MRI tongue contours drawn in the MRI (b.) and US (a.) frames, where the MRI palate contour is presented by red curves

It is important to emphasize that the transformation of (1) and (2) containing exact steps cannot be applied uniformly in the case of all speech sounds, since the parameter set describing the transformation can change sound by sound. This circumstance makes the compact combination of US and MRI records more difficult, but this problem can be resolved by the optimization of the parameters of the transformation, extending to several speech sounds, or by involving machine learning algorithms belonging to the forefront of statistical methods. In the next chapter, the application possibilities of machine learning are presented.

3 Machine Learning

We created our programs in the MATLAB environment, and we implemented machine learning by such an algorithm that determines the weight factors of the neural network, by the scaled conjugate gradient method [21]. Knowing the input parameters, this optimization procedure solves the system of equations assigned to the problem by an iterative method, while the output parameters calculated by the procedure converge to the prescribed values. The advantage of the method is the fast convergence that can be ensured by minimizing the number of steps of the iterative algorithm, thus, machine learning training can be carried out in a relatively short time. The iterative steps are realized along such a direction that enables faster convergence than the most negative gradient corresponding to the steepest descent, while it preserves the error minimization obtained in the previous steps. Training stops when the maximum number of epochs is reached, or the maximum amount of time is exceeded, or performance is minimized to the goal, or the performance gradient falls below the minimum performance gradient, or validation performance has increased more than maximum validation failures times since the last time it decreased.

We placed two hidden layers in the neural network, which individually contained 30 neurons. We designated the input parameters needed for learning by dint of four chosen points of the dynamically changing tongue contour, to that we assigned the discrete cosine transform of the tongue contour in the output side of the system. The four feature points coincide with those four points that are determined by the four inner radial sections of the angular range, as shown in Figure 7b. As illustrated by Figure 9, the feature points of the US and MRI tongue contours are stamped by magenta and yellow markers, respectively, together with the ordinal numbers of the given points along the green and blue curves. In this manner, the feature points of the US and MRI tongue contours correspond to each other pairwise, unambiguously, along a given radial section. It can be seen that the magenta markers follow each other in reverse order compared to the yellow markers. This effect is caused by the vertical reflection of the US tongue contour when embedding it into the MRI frame. The relative positions of the four feature points are identical in each frame, in the sense that the four points can be found at

about 20%, 40%, 60%, 80% of the angular range [-27°, 32°], in the case of all tongue contours. So the feature points are fixed automatically in all frames.

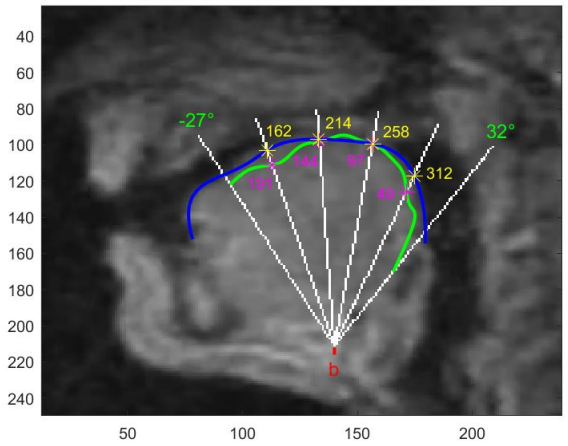

Figure 9

The feature points of the US (green) and MRI (blue) tongue contours stamped by magenta and yellow markers along the selected radial sections indicated by white lines

We executed the learning, first by fixing the input and output parameters arising from the MRI source, and then we tested the results in the same MRI frames. Based on a similar principle, we repeated the procedure for the US frames. Finally, combining the input parameters gained from the US source with the output parameters originating from the MRI source, we ran the algorithm again, then we tested the results in the MRI frames. The following sub-sections discuss the three different approaches.

3.1 MRI-MRI Learning

The subsection summarizes the results of machine learning accomplished in the case of the MRI records. The base of learning is formed by the phonemic configurations belonging to the speech sounds *ɔ, a:, e͡s, e͡ʃ, d, d͡z, d͡ʒ, ɛ, e:, g, ɟ, i, j, k, l, n, ɲ, o, ø, r, ʃ, s, t, c, u, y, z, ʒ*. The input parameters are given by the **y** coordinates of the four selected points of the tongue contour, measured in the plane of the image, while the set of output parameters is determined by the first twenty coefficients of the discrete cosine transform of the tongue contour. After running the learning algorithm, the trained tongue contour can be reconstructed by inverse discrete cosine transform. It practically means that the production of the complete curve occurs by using just four points. Our results are demonstrated through the example of sounds *j* and *t*.

Figures 10a and 10c present tongue contours fitted to the tongue positions corresponding to sounds *j* and *t*. Figures 10b and 10d display trained tongue contours belonging to the same sounds *j* and *t*. When comparing the fitted and trained tongue contours, no significant visual distinction shows up, the difference is minimal between the two curves, which can be determined also quantitatively for example by the values of the Mean Absolute Difference (MAD), Root Mean Squared Distance (RMSD), Mean Sum of Distances (MSD), or Nearest Neighbor Distance (NND).

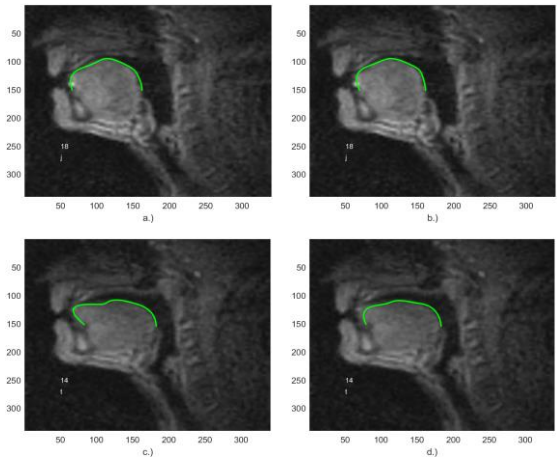

Figure 10

Fitted (a, c) and trained (b, d) MRI tongue contours in the case of sounds *j* (a, b) and *t* (c, d)

The results illustrated in Figure 10 reflect that the learning algorithm works effectively, confirmed by as well by the graphs of Figure 11 and showing the mean squared error of training, testing and validation. It can be seen that, besides rapid decrease, the errors of learning and testing are essentially identical.

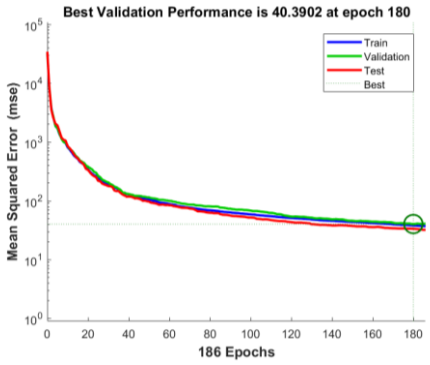

Figure 11

The mean squared error of training, testing, and validation in the case of MRI-MRI learning

3.2 US-US Learning

The subsection summarizes the results of machine learning performed in the case of the US records. In this case, learning is built upon utterances of CVCV type. The interpretation of the input and output parameters is the same as in the previous subsection, and at this time, the steps are led through the example of sounds *g* and *ʃ*.

Figures 12a and 12c demonstrate tongue contours fitted to the tongue positions corresponding to sounds *g* and *ʃ*. Figures 12b and 12d depict trained tongue contours belonging to the same sounds *g* and *ʃ*. Comparing the fitted and trained tongue contours, no considerable distinction can be observed between the two curves.

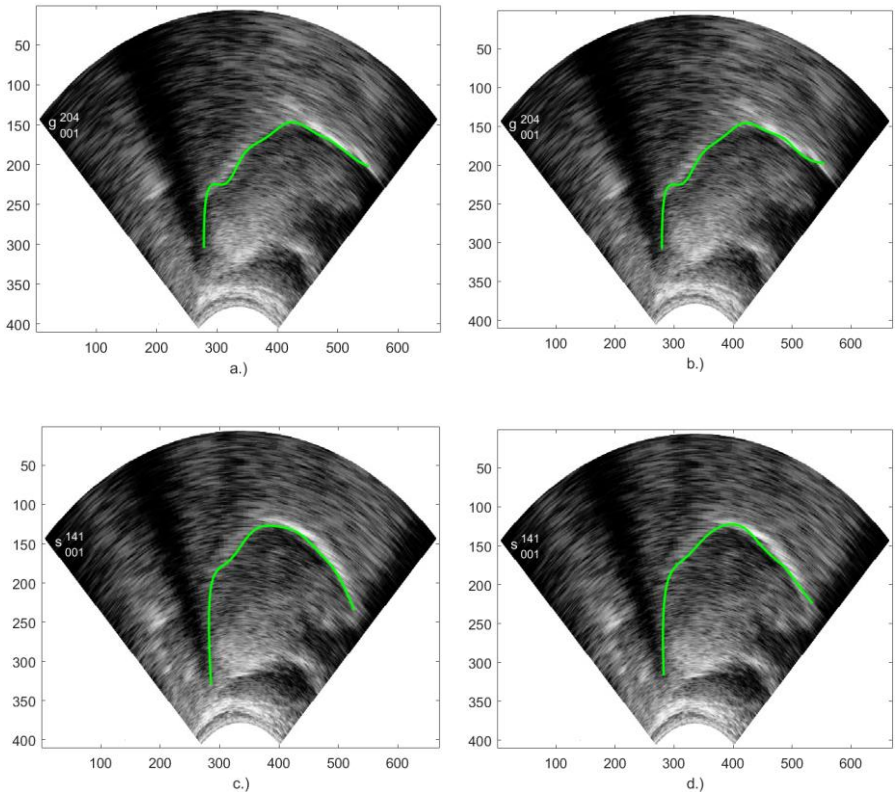

Figure 12
Fitted (a, c) and trained (b, d) US tongue contours in the case of sounds *g* (a, b) and *ʃ* (c, d)

Figure 13 illustrates the formation of the mean squared error of training, testing, and validation, the tendency of that is similar to the curves obtained during learning implemented by the MRI records.

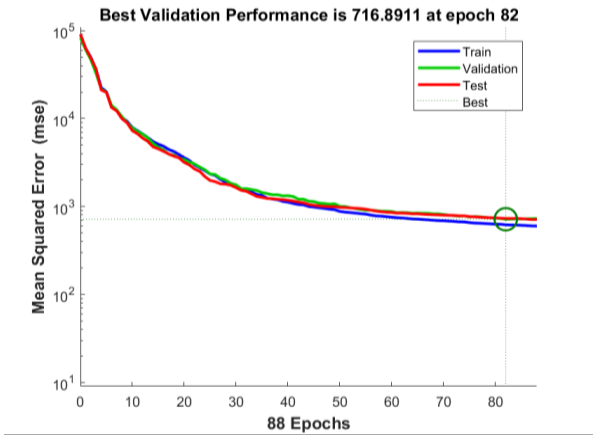

Figure 13
The mean squared error of training, testing, and validation in the case of US-US learning

3.3 US-MRI Learning

In the previous two subsections, the input and output parameters of machine learning originated from the same source, since MRI tongue contour was trained by MRI data, and US tongue contour was trained by US data. It is also worth examining how successful the parameters of the two different sources can be connected. It is quite challenging because the tongue contours of a female and male speaker of different anatomies need to be harmonized. Based on our expectations, however, even the parameters of the transformation carry quantitative information about the differences of the anatomies of the two speakers. For this purpose, we constructed the neural network such a way that its input parameters are created by the four selected points of the US tongue contour, and its output parameters are generated by the discrete cosine transform of the MRI tongue contour. Thereby, such a learning mechanism can be established in which MRI tongue contour can be produced by the utilization of US data. We note that the size of the used database lags behind the cases discussed in the previous two subsections by orders of magnitude. The reason for this is that the MRI and US records hold, not the same utterances, in all cases, furthermore, the number of frames assigned to the individual speech sounds does not match and that makes the harmonization of the parameters for the learning algorithm more difficult. Synchronization of the utterances and number of samples, however, is also currently in progress.

Figure 14a exemplifies the tongue contour fitted to the tongue position corresponding to sound *k*. Figure 14b presents the trained tongue contour belonging to the same sound *k*. The result can be interesting even from several viewpoints, since beyond the fact that the input and output parameters connected

by the neural network arise from records of utterers of various native language and different gender made by different imaging techniques, neither the condition can be neglected that learning produces a wider data set starting from a narrower one. Namely, as mentioned earlier, US records are not able to display the rear part of the tongue and the region of the tongue tip that is visible in MRI records without any obstacles. This predicts that, using the partial data originating from US records and involving learning algorithms, the contour of the complete edge of the tongue can be estimated effectively.

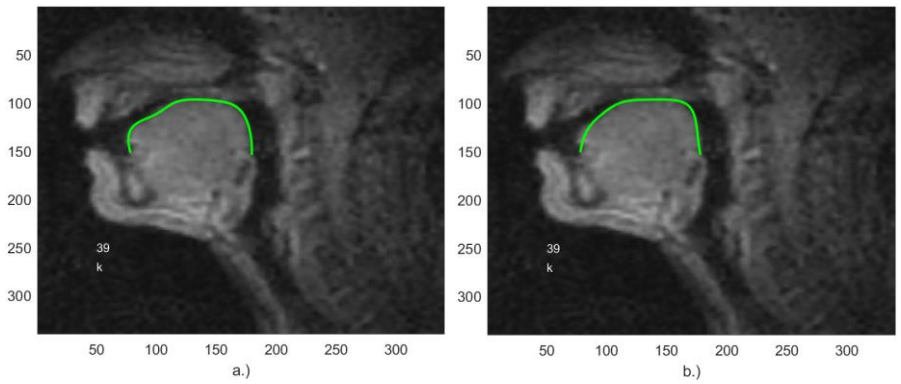

Figure 14
Fitted (a) and trained (b) MRI tongue contours in the case of sound *k*

Conclusions

The main goal of this work was the development and refinement of methods that can be applied towards articulatory speech synthesis. The tools of investigation are constituted by dynamic MRI and US records. The examinations basically run along two threads that approach the problem of harmonization of the relevant anatomic contours of MRI and US frames from different viewpoints. At the starting level, geometric transformations are performed, which interconnect the tongue and palate contours of the MRI and US frames in a bi-unique way. Although this procedure is based on exact mathematical considerations – according to the present stage of research work – it cannot be applied for all speech sounds, in a uniform manner, because the parameter set of the transformation does not contain the same values for each speech sound. So this solution seems to be quite tedious. In pursuance of our future plans, it will be resolved by the optimization of the parameters of the transformation, to produce satisfactory matching of MRI and US contours. Therefore, by way of statistical methods, machine learning is involved in the study, the application is associated with our automatic tongue contour tracking algorithms. Machine learning is implemented in respect of MRI-MRI, US-US, and US-MRI sources by the appropriate combining of the input and output parameters of the neural network. Currently, only a limited number of training and testing configurations are available, but the source data are being gradually expanded. The actual results

exhibit only a narrow slice of the ongoing research work, since the fields of articulatory speech synthesis and machine learning, raise, in themselves, a large number of problems, that can be regarded as temporarily partially solved. Accordingly, the future trends of research can be determined by the perfection of the models of speech synthesis created by statistical or rule-based algorithms and built on visual information. It has a potential fundamental importance, for example, in speech therapy for clinical purposes, in the shaping of non-native language learning trainings or in the construction and development of the synthesizers needed for vocalizing silent speech.

Acknowledgement

We would like to thank the MTA-ELTE Lendület Lingual Articulation Research Group, for providing the recordings with the Micro system.

References

[1] Czap, L., Pintér, J. M.: Intensity feature for speech stress detection. Proceedings of the 16th International Carpathian Control Conference Miskolc, Hungary: IEEE IAS/IES/PELS, 2015, 91-94

[2] Olaszy, G.: Making Speech Database for Machine Speech Production. (Beszédadatbázisok készítése gépi beszédelőállításhoz) Beszédkutatás99, 1999, 68-89

[3] Olaszy, G., Németh, G., Olaszi, P., Kiss, G.: Profivox: the Most Modern Native Speech Synthesiser (Profivox: a legkorszerűbb hazai beszédszintetizátor) Beszédkutatás 2000, 2000, 167-179

[4] Németh, G., Olaszy, G., Fék, M.: Development and Experimental Results of a Novel Corpus-based Machine Text-to-Speech System (Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei) Beszédkutatás, 2006, 183-196

[5] Sproat, R. W.: Multilingual text-to-speech synthesis, KLUWER Academic Publishers, 1997

[6] Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Tech., 6, 2003, 365-377

[7] Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Comm., 56, 2014, 85-100

[8] Zappi, V., Vasuvedan, A., Allen, A., Raghuvanshi, N., Fels, S.: Towards real-time two-dimensional wave propagation for articulatory speech synthesis. Proceedings of Meetings on Acoustics 171ASA, 26, 2016, 045005

[9] Czap, L., Pintér, J. M., Baksa-Varga, E.: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children. Speech Comm., 106, 2019, 7-20

[10] Wu, Z., Valentini-Botinhao, C., Watts, O., King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2015, 4460-4464

[11] Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Andrew, N., Raiman, J., Sengupta, S., Mohammad, S.: Deep voice: Real-time neural text-to-speech. Proceedings of the 34th International Conference on Machine Learning, 70, 2017, 195-204

[12] Roehling, S., MacDonald, B., Watson, C.: Proceedings of the Australasian International Conference on Speech Science and Technology, 2006, 130-135

[13] Li, X., MacDonald B., Watson, C. I.: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, 5009-5014

[14] Czap, L., Mátyás, J.: Virtual speaker. Proceedings of 6th International Carpathian Control Conference ICCC 2005 Miskolc, Hungary: University of Miskolc, 2005, 351-358

[15] Lulich, S. M., Berkson, K. H., de Jong, K.: Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. Journal of phonetics, 71, 2018, 410-424

[16] Li, M., Kambhamettu, C., Stone, M.: Automatic contour tracking in ultrasound images. Clinical linguistics and phonetics, 19, 2005, 545-554

[17] Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A., Varjasi, G.: Synchronised Speech and Tongue Ultrasound Records by the Sono-Speech System. 13th Conference on Hungarian Computational Linguistics (Szinkronizált beszéd- és nyelvultrahang-felvételek a Sono-Speech rendszerrel) University of Szeged, Institute of Informatics, Szeged, 2017, 339-346

[18] Zhao, L., Czap, L.: Automatic tracking of tongue contours in ultrasound records (A nyelvkontúr automatikus követése ultrahangos felvételeken) Beszédkutatás, 27, 2019, 331-343

[19] Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based Ultrasound-to-Speech Conversion for a Silent Speech Interface, Interspeech 2017, Stockholm, Sweden, 2017, 3672-3676

[20] Czap, L., Image processing (Képfeldolgozás), Miskolc-Egyetemváros, Hungary: Miskolci Egyetem, 2007

[21] Moller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural networks, 6, 1993, 525-533