UNIVERSITY OF MISKOLC
FACULTY OF MECHANICAL ENGINEERING
AND INFORMATICS

JÓZSEF HATVANY DOCTORAL SCHOOL FOR
COMPUTER SCIENCE AND ENGINEERING

Head: PROF. DR. JENŐ SZIGETI

Relationship between geometric and acoustic characteristics of articulation

Theses of Ph.D. dissertation

Author: DR. RÉKA TRENCSÉNYI

Supervisor: PROF. DR. LÁSZLÓ CZAP

Miskolc
2024

Members of the evaluation committee:

President:

- Prof. Dr. Jenő Szigeti, professor,
  University of Miskolc

Members:

- Prof. Dr. Klára Vicsi, private professor,
  Budapest University of Technology and Economics
- Dr. Attila Trohák, associate professor,
  University of Miskolc
- Dr. Attila Varga, associate professor,
  University of Miskolc

Reviewers:

- Prof. Dr. György Takács, honorary professor,
  Pázmány Péter Catholic University
- Dr. Erika Varga Dr. Baksáné, associate professor,
  University of Miskolc

1. Short exposition of the research topic and the tasks set

My research topic is connected to the areas of speech science. One of the most important thematic domains of speech research is speech synthesis, which can form an elementary constituent of the human-machine relationship. In this case, the communication role of the machine is manifested in the fact that it becomes a coding transmitter, i.e. it produces speech. Nowadays, the most widespread direction of speech synthesis is making text-to-speech readers, which vocalise written texts of general topics or confined to a concrete subject. For instance, belletristic readers, passenger information systems, newsreaders, sound weather forecast, or telephonic directory enquiry services can be classified into this application category. The aim of the creation of speech synthesisers is the realistic imitation of the acoustic product formed during natural human speech. In this approach, the starting point is given by the waveform of speech, which is applied in two kinds of solutions to the production of machine speech. So-called source coding techniques belong to one of the groups, by dint of which the essential information is extracted from the speech signal, and these are treated as input data series during the synthesis. The other solution utilises human voice for building speech directly in such a way that waveform segments of different length are cut from the speech signal and stored, then, by the appropriate selection and concatenation of the elements obtained, the desired speech wave is constructed. Furthermore, based on broader methodological viewpoints, one can already

distinguish rule-based and statistical speech production procedures, respectively. In the case of the previous one, each step of the synthesis is coordinated by rules established by observations and experiences, and, in the case of the latter one, one can get to speech production through internal system states based on probabilities. One of the typical variants of the statistical methods is the construction and application of machine learning algorithms, which can be counted as one of the most actively prospering tendency of the present scientific researches.

Text-to-speech systems represent the classical branch of speech synthesis. Besides this, however, also such fields are starting to come into the limelight ever more lively that are less elaborated, and a lot of open problems are still expected to be solved. For example, articulatory speech synthesis can be classified here, which, instead of human voice samples, tries to implement the imitation of the acoustic product by machine imaging of voice production and articulation. One of the technological streams of this is the experimentation aiming at articulatory electromechanical speech generators needed for the production of speech of robots. Speech synthesis built on modelling of the articulation channel, alias vocal tract, ranging from the glottis to the oral and nasal aperture, favours also future trends, which mainly relies on visual information. Visual information gained about the physiological processes of human speech largely promote the understanding of the complex mechanism of speech production and thereby the effective development of the methods of speech synthesis. Radiological and monitoring procedures available in these days – such as magnetic resonance imaging (MRI), computer tomography (CT), ultrasound (US), electropalatography

(EPG), electromagnetic articulography (EMA), or electroglottography (EGG) – play an indispensable role in the treatment of the problem of the acoustic-articulatory conversion. Using the morphological and geometric data generated by the dint of the above mentioned imaging and monitoring techniques, the articulatory movements belonging to the given speech signal can be mapped completely. Connecting articulation to acoustics, i.e. the implementation of speech production based on morphological and geometric data of the vocal tract is, however, not a trivial task. The actuality of the problem raising is shown by the fact that revealing and practical imaging of the articulatory-acoustic system of relationships can have fundamental importance for example in the speech therapy of clinical purpose, in the shaping of non-native language learning trainings, or in the construction and development of synthesisers needed for vocalising silent speech, which can serve the rehabilitation of people undergoing laryngectomy, as well.

During the research work, it is also worth paying attention to the comparison and harmonisation of the visual information produced by dint of different imaging procedures, since the simultaneous analysis of data arising from different sources can further deepen the knowledge associated with the articulatory and acoustic perspectives of speech. The simultaneous comparative application of the monitoring techniques is not a trivial task at all, as the appropriate and authentic combination of the sources demands very serious and professional anatomic, geometric, engineering, and informatical knowledge. The effort, however, is practically essential for the detailed unravelling of the operation of the vocal tract and the deeper understanding of the

articulatory-acoustic relationships, respectively.

In light of the above, one of the main goals of my doctoral research work was the harmonisation of the radial and rectangular geometries of two-dimensional US and MRI records made during speech, the ground of which was formed by tongue and palate contours fitted to the image frames by automatic algorithms. I intended to realise this task in approaches relying on analytic rules and artificial intelligence. Following the analytic line, I worked out such geometric transformations that connect and embed into each other the anatomic environments of the two sources in a mutual unambiguous and bidirectional manner so that, by the optimisation of the parameters of the mathematical operations, the best possible coincidence between the tongue and palate contours of the US and MRI frames should be attainable. To enforce artificial intelligence, I envisaged the application of machine learning algorithms with the construction of such neural networks that realise the learning of data arising from the MRI tongue contours, resting on the parameters extracted from the US tongue contours.

The other larger topic of my researches was the carrying out of articulatory speech synthesis, to which I made use of the above mentioned US and MRI records. In the first step, I aimed at the dynamic gaining of the relevant geometric data of the imaging sources serving as the base of synthesis, in possession of which I planned to create stand-alone speech sounds and continuous speech, respectively. In the course of this, I wished to invoke artificial intelligence again. I endeavoured to build such machine learning algorithms that, starting from the visual geometric features, execute the learning of articulatory parameters

derived from the acoustic signals of the US and MRI records.

2. The methods and models applied

I implemented the tasks set during my research work possessing the relevant theoretical knowledge by dint of computer methods. The treatment of the US and MRI records could not lack the deployment of some tricks of image processing, and finding of the anatomic contour lines connecting to the visual information was allowed by the usage of automatic tongue contour tracking algorithms based on dynamic programming. In my investigations aimed at the structures and the harmonisation of the US and MRI records, such mathematical approaches had a key role by means of which I enforced analytical geometric considerations and transformations described by exact relationships in the biunique correspondence of the visual elements of the two sources. I supplemented the realisation of the geometric transformations by the elaboration of algorithms based on optimisation principles. I applied machine learning at several points of my research work, i.e. I constructed neural networks producing artificial intelligence for the training of parameters belonging to the given task. When executing the speech synthesis, I relied on the acoustic tube model and the principle of linear prediction, respectively, counted as one of the most essential tools of speech technology. I wrote all of the codes needed for the programming of each sub-process of the analyses on MATLAB interface, in some phases of my researches, however, also simple manual calculations were involved in the investiagions.

3. Results, theses

1. In the first phase, I focused on the simultaneous analysis and harmonisation of US and MRI records, the aim of which was the mutual unambiguous correspondence of the radial geometry of the US frames and the rectangular geometry of the MRI frames. One possible manner of the harmonisation of the two sources is the bidirectional conversion of the contour lines of the oral cavity of the same type relevant from the viewpoint of articulation. I implemented this idea by carrying out special geometric transformations, the obvious tools of which were the tongue and palate contours. After designating the mathematical framework including the visual schemes and the characteristic parameters of the transformations, I created such optimisation techniques that, based on different perspectives, search for the most favourable values of the parameter set defined by the transformations. By dint of the transformations and the optimisation, I projected the environment of the two sources onto each other so that their relative position should be ideal, i.e. the global distance between the US and MRI contours should be minimal, ensuring by this the largest possible overlap between the curves. The transformation mechanism encompasses fundamentally three operations, which concerns the geometric regions covered by the tongue and palate contours. The three operations are composed by elements declared by scaling of the radial range, scaling of the angular range, as well as rotation of the angular range. One-one scale factor is responsible for the normalisation of the radial and angular ranges, which are the radial magnification and the angular deformation. The rotation

of the angular range can be taken into account by a translational factor, which provides the angular displacement. I chose the angular deformation as a unit, i.e. I interpreted the transformation as a conformal mapping, so the parameters to be optimised were created by the radial magnification and the angular displacement, to which I also added the coordinates of the centre serving as the origin of the transformations.

Thesis T_1: The radial and rectangular structures of the two-dimensional US and MRI records can be embedded into each other by means of such geometric transformations that realise the bidirectional conversion of the geometric regions defined by the tongue and palate contours of the records: the scaling of the radial range and the rotation of the angular range, respectively. Optimising the radial magnification, the angular displacement, as well as the centre of the transformations connected to the operations, the best overlap between the anatomic domains of the US and MRI records can be achieved. [1,2,4,5,9,11]

2. In the second phase, I still aimed at the harmonisation of the US and MRI records, the accessories of which were the tongue contours. This time, however, I replaced the optimisation approach based on the analytical geometric view and the resulting exact mathematical transformations by the application of artificial intelligence. It means that I connected the tongue

contours of the two sources by dint of machine learning algorithms, during which I created the neural network by distinct constructions, and I took into account the formal characteristics of the contours by two parameters of different kinds. On one hand, I designated a discrete series of points consisting of a finite number of elements along the curves, bringing into existence the set of feature points. On the other hand, I derived DCT coefficients from the contours by the execution of discrete cosine transform (DCT) applied to the smoothing of curves. I excited the input of the neural networks by data gained from the US tongue contours, and at the output, I set parameters extracted from the MRI tongue contours, thus the system ultimately learns the shapes of the MRI tongue contours belonging to different speech sounds based on the US tongue contours applied as training patterns. Varying the type and the number of input and output parameters of the neural network and the number of hidden layers and neurons, respectively, I constructed different system configurations, and I analysed the obtained results in qualitative and quantitative manners, as well, choosing the system setting giving the best result. Initially, I fulfilled the machine learning by the usage of all US tongue contours serving as the source of input data, then I made a five-stage filtering on the family of US tongue contours, with which I excluded the false curves. As a result of this, I selected the contours belonging to sound transitions, the contours of negative slope, the contours overflowing on the palate contour, and the convex and the abnormal contours. The measure of qualitative and quantitative match between the MRI tongue contours trained and fitted by the automatic contour tracking algorithm improves if the con-

tours belonging to sound transitions, the contours of negative slope, the contours overflowing on the palate contour, and the convex and the abnormal curves are excluded from the set of US tongue contours serving as the source of the input data. The measure of qualitative and quantitative match between the MRI tongue contours trained and fitted by the automatic contour tracking algorithm improves if the number of feature points is increased at the input independently of the type of the output parameters. The measure of qualitative and quantitative match between the MRI tongue contours trained and fitted by the automatic contour tracking algorithm improves if the number of DCT coefficients is increased at the input independently of the type of the output parameters.

Thesis T_2: Tongue contours fitted to two-dimensional US and MRI frames can be harmonised by dint of machine learning algorithms so that the shapes of the MRI tongue contours can be trained based on parameters of the US tongue contours. The measure of qualitative and quantitative match between the MRI tongue contours trained and fitted by the automatic contour tracking algorithm improves if DCT coefficients are applied instead of the feature points at the output independently of the type of the input parameters. [3,6,7,8,10]

3. In the third phase, I dealt with production of machine speech, the starting point of which was formed by the US and MRI

records. Namely, I had the basic concept that I should implement speech synthesis relying on such visual information that can be gained from the mentioned two-dimensional image sources. I derived the necessary visual data, on one hand, by dint of tongue and palate contours fitted to the frames so that I developed two different algorithms, by the application of which the sagittal radial distances between the palate and the surface of the tongue can be measured dynamically in the vocal tract. On the other hand, I involved also the coefficients of the discrete cosine transform (DCT) used for smoothing the tongue contours fitted to the frames in the investigations. By means of the tools of machine learning, I tried to connect the distance data and DCT coefficients obtained to the different articulatory parameters characterising speech, which I interpreted in the framework of the acoustic tube model, as well as the principle of linear prediction (LPC). Accordingly, I excited the input of the neural network by radial distances or DCT coefficients, and, at the output, I set reflection coefficients extracted directly from the speech signal and defined by the acoustic tube model or cross sections obtained by the mediation of LPC coefficients originated from the speech signal. During my work, I wished to produce stand-alone, sustained speech sounds and continuous speech, respectively. The synthesis of speech sounds happened in US-MRI combination, which means that I effectuated the training of the speech sounds of the MRI records by visual data arising from the US records. I carried out the generation of continuous speech in US-US and MRI-MRI pairing, respectively, thus I interpreted parameters gained from the same source on both sides of the system for the training.

Thesis T_3: Starting from the two-dimensional US and MRI records, articulatory speech synthesis can be realised by utilising artificial intelligence. Based on sagittal radial distances, as well as DCT coefficients, reflection coefficients treated as articulatory parameters of the speech signal and the cross sections of the vocal tract can be trained by dint of neural networks, from which the original speech signal can be reconstructed by the acoustic tube model and the linear predictive coding, as well. During the synthesis, machine speech of better quality can be produced by the training of the reflection coefficients against the cross sections, and, at the level of training patterns, preference is given to the DCT coefficients against the radial distances. [12]

4. List of publications

a.) Articles connected to the topic of the doctoral dissertation:

1. R. Trencsényi, *MRI- és UH-felvételek geometriai elemzése a beszédszintézisben*, Acta Medicinae et Sociologica 11(31), 55-65, 2020.

2. R. Trencsényi, L. Czap, *UH-és MRI-nyelvkontúrok optimalizációja*, In Speech Research Conference, Hungarian Research Institute for Linguistics, Budapest, Hungary, 14-15th December 2020, 86-88, 2020.

3. R. Trencsényi, *A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata*, In XVI. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2020, Szeged, Magyarország, 2020. január 23–24., 233-244, 2020.

4. R. Trencsényi, L. Czap, *Possible methods for combining tongue contours of dynamic MRI and ultrasound records*, Acta Polytechnica Hungarica, 18(4), 143-160, 2021.

5. R. Trencsényi, L. Czap, *A possible optimisation procedure for US and MRI tongue contours*, In Proceedings of the 1st Conference on Information Technology and Data Science, CITDS 2020, CEUR Workshop Proceedings, Debrecen, Hungary, 6-8th November 2020, 259-269, 2021.

6. R. Trencsényi, L. Czap, *Machine learning applied in speech science*, In 23rd International Carpathian Control Conference, ICCC 2022, Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Sinaia, Romania, 29th May - 1st June 2022, 309-314, 2022.

7. R. Trencsényi, L. Czap, *Articulatory data of audiovisual records of speech connected by machine learning*, In 2nd Conference on Information Technology and Data Science, CITDS 2021, Proceedings Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Debrecen, Hungary, 16-18th May 2022, 297-301, 2022.

8. R. Trencsényi, L. Czap, *A neural network based approach for combining ultrasound and MRI data of 2-D dynamic records of human speech*, In 13th IEEE Internati-

onal Conference on Cognitive Infocommunications, CogInfoCom 2022, Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Budapest, Hungary, 22nd-23rd September, 47-52, 2022.

9. R. Trencsényi, L. Czap, *Optimisation techniques in speech processing*, In Doktoranduszok Fóruma 2021, Miskolc-Egyetemváros, Magyarország, 98-104, 2022.

10. R. Trencsényi, *A gépi tanulóalgoritmusok hatékonyságának vizsgálata kétdimenziós ultrahang- és MRI-felvételek adatainak összekapcsolásában*, In Doktoranduszok Fóruma 2022, Miskolc-Egyetemváros, Magyarország, 84-89, 2023.

11. R. Trencsényi, L. Czap, *Association of relevant anatomic contours of ultrasound and MRI images by optimisation via different geometric parameters*, In 25th International Carpathian Control Conference, ICCC 2024, Krynica Zdrój, Poland, 22nd-24th May 2024, pp. 1-6.

12. R. Trencsényi, L. Czap, *Ultrasound- and MRI-based speech synthesis applying neural networks*, In 25th International Carpathian Control Conference, ICCC 2024, Krynica Zdrój, Poland, 22nd-24th May 2024, pp. 1-6.

b.) Other article:

- R. Trencsényi, L. Czap, *Artikulációs fonetikai jellemzők verifikálása kvantitatív adatokkal*, Beszédtudomány/Speech Science, 2(1), 243-260, 2021.