# Machine Learning Applied in Speech Science

1st Réka Trencsényi
*Department of Electrical and Electronic Engineering*
*University of Debrecen*
Debrecen, Hungary
trencsenyi.reka@science.unideb.hu

2nd László Czap
*Institute of Automation and Infocommunication*
*University of Miskolc*
Miskolc, Hungary
czap@uni-miskolc.hu

***Abstract*—We study the qualitative and quantitative information acquired from dynamic audiovisual sources storing image and sound signals recorded during human speech. Our main tool is machine learning, which connects data arising from records made by ultrasound and magnetic resonance imaging techniques.**

***Index Terms*—machine learning, neural networks, tongue contour tracking, ultrasound and MRI records, human speech**

## I. Introduction

Machine learning has become a more and more popular apparatus in the treatment of complex systems having a large number of variables and parameters. Nowadays, the technique has been coming increasingly into the limelight in the field of speech research, as well, as it is commonly used for both speech recognition and speech synthesis.

For instance, the hidden Markov model is often used for speech and speaker recognition [1], and the Gaussian mixture model is also favorable in speech emotion recognition [2]. Classification of speech patterns is a crucial task during speech recognition, which can be well realised by support vector machine models [3].

The conventional trend of speech synthesis is based on text-to-speech systems that can be implemented by the application of neural networks [4]. Moreover, also visual information can be used for generating acoustic signals. The visual and articulatory movement data describing the vocal organs can be gained from different sources to train the given network [5]. The simplest technique is electromagnetic articulography (EMA) which employs different sensors to measure the position and movement of the vocal organs in the oral cavity. Applying EMA, the acoustic-articulatory conversion can be carried out in both directions, by the reconstruction of speech from sensor data [6], or by the estimation of articulatory trajectories from acoustic data [7].

Besides sensor-based techniques, also imaging methods can be strong and supporting grounds of the study of human speech. Accordingly, such neural networks can be constructed that are trained by data acquired by ultrasound (US) procedure [8], [9]. In addition, also magnetic resonance imaging (MRI) can be highly helpful in speech research, however, it is not so spread in connection with machine learning.

There are a few publications that combine US and MRI sources by different methods [10], [11], but, as far as we know, there is a gap in the literature for the joining of the two sources by machine learning. Even this was the motivation of the present article.

## II. Subjects of machine learning

The fundamental framework of our present work is given by audiovisual sources that record dynamic image and sound information of human speech simultaneously. In the moving images, one can observe the continuous movement of the active vocal organs (e.g. lips, tongue, soft palate, epiglottis) and the static positions of the passive vocal organs (e.g. hard palate, glottis), as well. The sound as a speech signal is adjusted to the series of the corresponding images, resulting in synchronized sound and image packages belonging to the uttering.

The techniques we utilized produce two-dimensional records made by ultrasound (US) and magnetic resonance imaging (MRI). In the first case, mostly the region of the oral cavity can be monitored from outside by dint of a US transducer that can be fixed to the head of the speaker under the chin. In the US pictures, only the motion of the tongue and the epiglottis can be followed, but the other vocal organs are hidden for the viewer, as the glottis and lips are out of the scanning region, while the hard and soft palate can not be detected due to the special reflections of US waves in the oral cavity. Furthermore, the hyoid bone and the mandible are not transparent against US beams, therefore they shield the back part and the tip of the tongue, emerging as dark ranges in the image. MRI frames, however, can cover a more extended region of the human head since the complete voice box is visible and no screening effects disturb the imaging. Thus, the lips, the tongue, the hard and soft palate, the epiglottis, and the glottis can be easily identified. A great advantage of both methods is the good spatial and temporal resolution, although making MRI records always demands clinical conditions that are sometimes difficult to be ensured. The US package was recorded by the Micro system of the MTA-ELTE Lendület Lingual Articulation Research Group, and the MRI package was selected from the freely available database of the University of Southern California. In the US records, sentences containing vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) sound connections are uttered by a Hungarian female speaker, while, in the MRI records, only VCV structures are articulated by an American English male speaker.

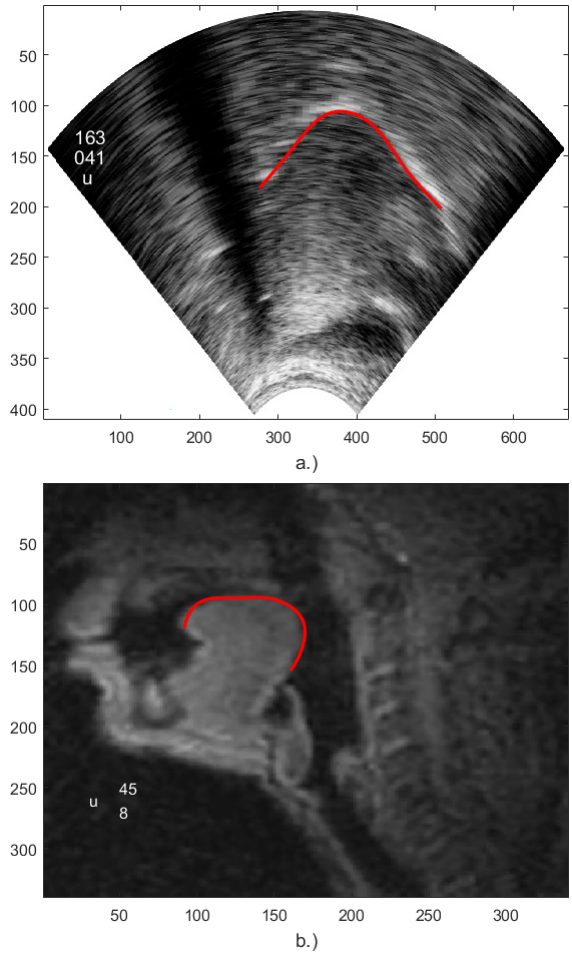


Fig. 1. A US (a.) and an MRI (b.) frame with the fitted tongue contours.

The basic tools for the implementation of machine learning were tongue contours that we fitted to the surface of the tongue displayed by the above mentioned US and MRI frames by dynamic contour tracking algorithms developed in MATLAB. The surface of the tongue appears as a bright zone in the US images and it can be determined as a contrast between the bright domain of the tissue of the tongue and the dark domain of the air above the tongue in the MRI frames. Consequently, contour tracking means searching for the pixels of maximal brightness designating the curve of the surface of the tongue in both cases. Parts a.) and b.) of Fig. 1 show a US and an MRI frame in the case of sound *u* together with the fitted tongue contours drawn by red curves. In the US frame, the tip of the tongue is on the right side of the image.

Using the dynamically computed tongue contours, we aimed to build a neural network that learns MRI tongue contours from the data of the appropriate US tongue contours.

## III. The Neural Network

We elaborated our machine learning algorithms in MATLAB constructing a neural network that contains one hidden layer with only 10 neurons because, as an initial attempt,

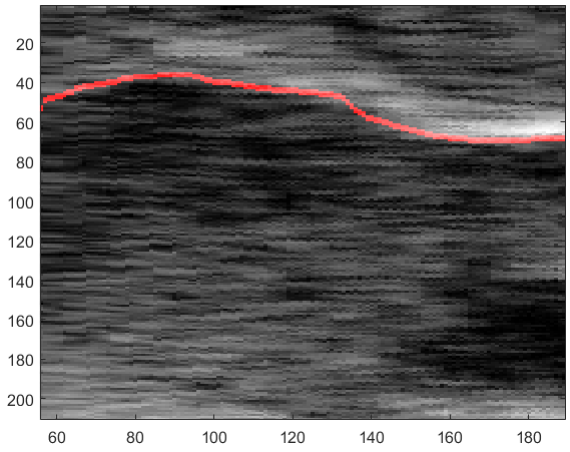

Fig. 2. The rectangular matrix structure of a US image after resampling. The tongue contour is drawn by red.

we tried to create a very simple network in order to test how efficiently it is capable of executing a relatively complex training task.

The input parameters of the network were derived from US tongue contours by selecting some feature points, which take part in the training. We specified five different cases according to the choice of 1, 2, 3, 4, and 5 points along the curves with specially fixed locations. Namely, the five feature points were positioned at 10%, 30%, 50%, 70%, and 90% of the total length of the tongue contour in each US frame. These percentage values are assigned to the tongue contours at an intermediate step of image processing and contour tracking, when, in the US frames (see Fig. 1.a), the algorithm creates radial sections of the image, and these sections are arranged into a rectangular matrix structure. After this resampling process, the tongue contour is formed as a series of linear sections joining one another in the matrix exemplified by Fig. 2. Taking the stretched tongue contours, the above mentioned five feature points are chosen in an equidistant manner, which is schematically illustrated by Fig. 3. For the training, we use only the vertical coordinates of the feature points. At the final step of contour tracking, the stretched curves are reshaped to the radial geometry of the US images, hence the distances between two adjacent feature points are no longer equal, as the stretched curve is deformed like a flexible rope. The relative positions of the five feature points in the radial geometry are demonstrated by Fig. 4, where the five different input cases of the neural network are separated. So, we train the same network by 1, 2, 3, 4, and 5 contour points, respectively, in such a way that, starting from the point at 10% of the total length of the tongue contour, always the next point is added to the previous points in the set of positions fixed by 10%, 30%, 50%, 70%, and 90%.

The output data set of the neural network was acquired from MRI tongue contours by designating the first 10 coefficients of the discrete cosine transforms of the tongue contours. Discrete cosine transform (DCT) is highly relevant in the shape of a

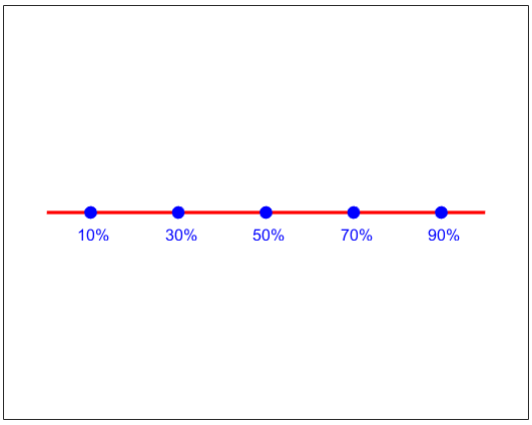


Fig. 3. The schematic arrangement of the five feature points of stretched tongue contours in the rectangular geometry.

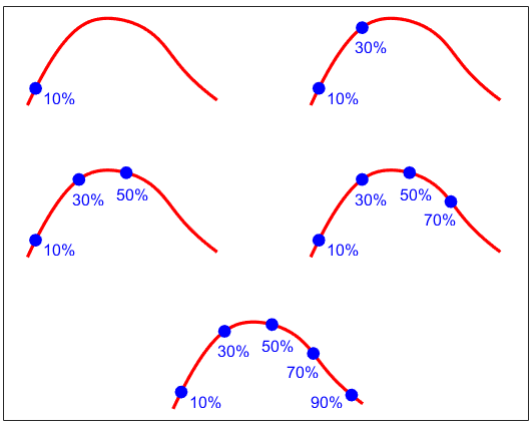


Fig. 4. The schematic arrangement of the feature points of the reshaped tongue contours in the radial geometry for the five different input cases of the neural network.

given tongue contour since it was applied for the smoothing of the final curve produced by the contour tracking algorithm, thus DCT coefficients carry important information about the local characteristics of the tongue contours.

The neural network was trained for a set of speech sounds including three vowels and three consonants by the assortment $\{a,\ e,\ o,\ k,\ s,\ t\}$. The main viewpoint in the selection of sounds was to have front and rear tongue positions, as well. Accordingly, sounds *a* and *o* are articulated with rear tongue position, while sound *e* is uttered with front tongue position. Sound *k* is velar, which means that the highest point of the tongue touches the soft palate, so it is formed in the rear part of the oral cavity. Sound *t* is alveolar, thus the highest point of the tongue touches the ridge behind the upper teeth in the front part of the oral cavity. Sound *s* is postalveolar, i.e. the place of articulation is near the alveolar region. All in all, the setting of the above group of speech sounds ensures the possibility to investigate the reproduction of the extremal positions of the tongue that can be available in our US and MRI records. In our US package, we have 1084 *a*-contours, 386 *e*-contours, 853 *o*-contours, 489 *k*-contours, 205 *s*-contours, and 1062 *t*-contours, so the total number of input parameters is 4079. In our MRI package, however, the number of frames belonging to the given sound is much smaller than that of the same sound of the US package. Therefore, in order to have the same number of output parameters as the inputs, we picked only one MRI tongue contour belonging to the middle frame of each sustained sound, then we repeated the computed DCT coefficients of the single curve as many times as desired. It can be an appropriate method because the shape of the MRI tongue contours locally changes slightly from frame to frame in the case of sustained sounds. As a result, the neural network is supplied by arrays of dimension $4079 \times n$, where $n$ stands for the number of training points by $n = 1, 2, 3, 4, 5$, and an array of dimension $4079 \times 10$ appears at the output according to the 10 DCT coefficients.

The machine learning was implemented by an algorithm that determines the weight factors and bias values of the neural network by the scaled conjugate gradient (SCG) method. It is a supervised learning algorithm for feedforward neural networks. During the optimization, the system of equations assigned to the given problem is solved iteratively knowing the input parameters, while the computed output parameters converge to the prescribed values. The benefit of the method is that a quite fast convergence can be guaranteed by minimizing the number of steps of the iteration algorithm, so the training can be realized in a relatively short time. The iteration steps occur 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.

## IV. Results and Conclusions

The constructed neural network detailed in section III connects data collected from US tongue contours and information acquired from MRI tongue contours, thus we followed the MRI-from-US direction during machine learning. It is more challenging than the reverse US-from-MRI way because, in the US images, the tongue can be seen only partially, as the rear part and the tip of the tongue are shadowed by dark zones (see section II). In the MRI frames, however, the tongue becomes apparent entirely. As a consequence, the neural network is fed by a narrower input data set, and a wider output parameter set is produced.

We evaluated our results qualitatively and quantitatively, as well. For qualitative judgement, we visually investigated the position and shape of the trained MRI tongue contours compared to the original fitted MRI tongue contour in the case of each sound of the set $\{a,\ e,\ o,\ k,\ s,\ t\}$. We found that the measure of agreement between the trained and fitted curves improves when the number of feature points of the US tongue contours is increased. Examples confirming our observations are presented by Fig. 5 and Fig. 6 in the case of sounds *a* and *t*, respectively, where the trained tongue contours drawn by red and the fitted tongue contours marked by green are depicted in the same frame. Parts a.) show the results

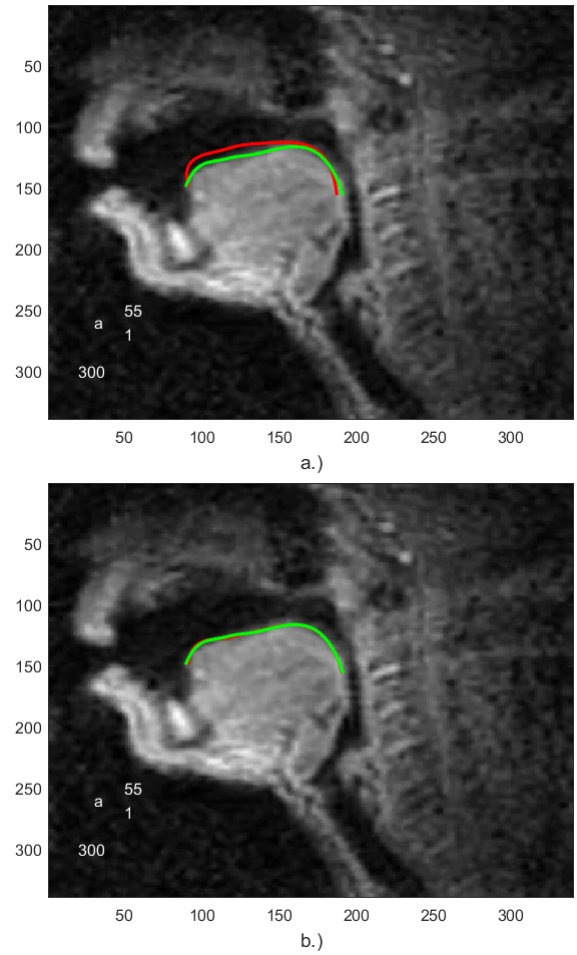


Fig. 5. The trained (red) and fitted (green) MRI tongue contours with 1 training point (a.) and 5 training points (b.) of the US tongue contour in the case of sound *a*.

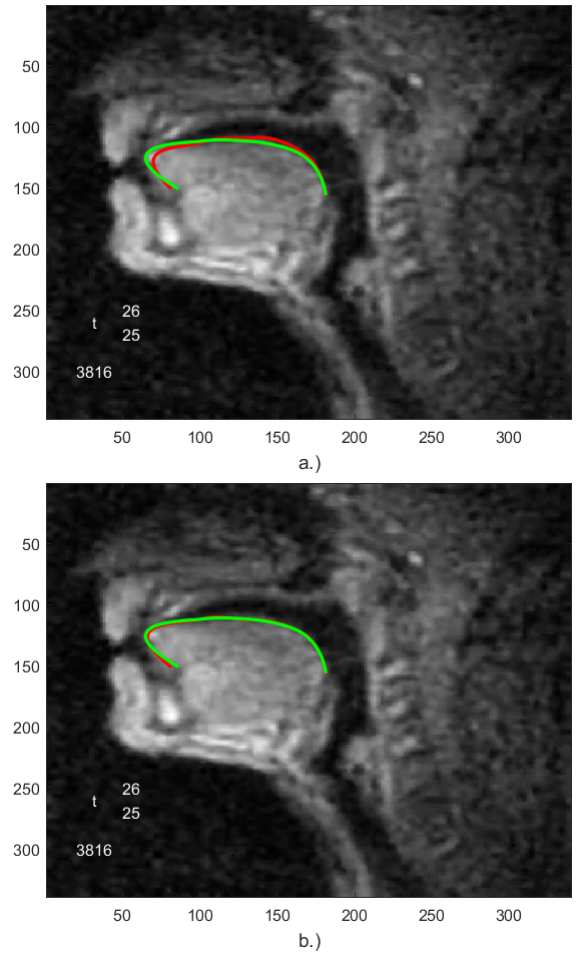


Fig. 6. The trained (red) and fitted (green) MRI tongue contours with 1 training point (a.) and 5 training points of the US tongue contour (b.) in the case of sound *t*.

of machine learning with 1 training point (1 feature point is selected on each US tongue contour at position 10% of the total length of the curve), while parts b.) represent the results of machine learning with 5 training points (5 feature points are selected on each US tongue contour at positions 10%, 30%, 50%, 70%, 90% of the total length of the curve).

For the quantitative description, we determined the distances between the trained and fitted MRI tongue contours by the application of the Nearest Neighbor Distance (NND) measure defined mathematically by

$$D_{F,G} = \frac{1}{n+m}\left(\sum_{i=1}^{n} \min_j |f_i - g_j| + \sum_{i=1}^{m} \min_j |g_i - f_j|\right). \quad (1)$$

Equation (1) is related to two different curves $F$ and $G$ with arbitrary numbers of points $n$ and $m$, which are not necessarily equal to each other generally. The first contribution of (1) expresses the sum of the minima of the distances measured between the given point of curve $G$ and all points of curve $F$ for all possible points of curve $G$. Similarly, the second contribution of (1) provides the sum of the minima of the distances measured between the given point of curve $F$ and all points of curve $G$ for all possible points of curve $F$. The total sum is normalized by the sum of the number of points of the two curves. In our case, curves $F$ and $G$ play the role of the fitted and trained MRI tongue contours with the condition $n = m$.

Applying (1), we calculated the NND measures for all trained tongue contours for each sound of set {*a, e, o, k, s, t*} in the case of all the five different training inputs, then we gave the averages of the NNDs belonging to all tongue contours of all sounds for all input settings. The results are summarized by Table I that quantitatively verifies our qualitative experiences arising from the visual study of the trained and fitted tongue contours. So, finally, we conclude that training with 5 five feature points of the US tongue contours produces the best match between the trained and fitted MRI tongue contours with the lowest NND.

## V. Summary

In this report, we analyzed two-dimensional dynamic US and MRI sources recording human speech signal together

TABLE I
THE AVERAGES OF THE NNDS

| | 1 point | 2 points | 3 points | 4 points | 5 points |
|---|---|---|---|---|---|
| NND | 5.6576 | 5.4410 | 4.8661 | 4.0224 | 3.7572 |

with the synchronized visual display of the vocal organs. We aimed to connect the two sources by machine learning. The basic tools for this were tongue contours fitted to the surface of the tongue of the US and MRI frames by our automatic contour tracking algorithms. The constructed neural network containing one hidden layer with 10 neurons was trained by 1, 2, 3, 4, and 5 feature points of the US tongue contours at the input, and the first 10 coefficients of the discrete cosine transforms of the corresponding MRI tongue contours were set as output parameters to be learned. The training was performed for a specific set including speech sounds $\{a,\ e,\ o,\ k,\ s,\ t\}$, applying the scaled conjugate gradient method. The obtained results were studied at both qualitative and quantitative levels. Qualitatively, we checked the visual match of the trained and fitted MRI tongue contours. We found that they are in a very good coincidence with each other, and the measure of agreement between the trained and fitted curves gets better when the number of feature points of the US tongue contours is increased. Quantitatively, we calculated the average nearest neighbor distances measured between the trained and fitted MRI tongue contours for all input settings. We stated that the closeness of the trained and fitted curves is the best when the number of feature points of the US tongue contours takes the greatest value of 5. Finally, we were led to the conclusion that our qualitative and quantitative results point in the same direction, as they certify each other, so the main importance is that complete MRI tongue contours can be perfectly trained by partial US tongue contours.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and speaker recognition system using artificial neural networks and hidden Markov model," In 2012 IEEE International Conference on Communication Systems and Network Technologies, 311–315, 2012.

[2] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network," IEEE access, 7, 26777–26787, 2019.

[3] B. A. Sonkamble, D. D. Doye, "An overview of speech recognition system based on the support vector machines," In 2008 IEEE International Conference on Computer and Communication Engineering, 768–771, 2008.

[4] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, ... , and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," In PMLR International Conference on Machine Learning, 195–204, 2017.

[5] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview—application of articulatory movements using machine learning algorithms," Acoustical Science and Technology, 36(6), 467–477, 2015.

[6] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, ... , and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 2362–2374, 2017.

[7] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," In NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning, 1–9, 2011.

[8] B. Denby, M. Stone, "Speech synthesis from real time ultrasound images of the tongue," In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, I-685, 2004.

[9] L. Tang, G. Hamarneh, and T. Bressmann, "A machine learning approach to tongue motion analysis in 2d ultrasound image sequences," In International Workshop on Machine Learning in Medical Imaging, 151–158, 2011.

[10] M. Aron, M.-O. Berger, E. Kerrien, "Multimodal fusion of electromagnetic, ultrasound, and MRI data for building an articulatory model," In 8th International Seminar On Speech Production - ISSP'08, ffinria-00326290f, 2008.

[11] R. Trencsényi, L. Czap, "Possible Methods for Combining Tongue Contours of Dynamic MRI and Ultrasound Records," Acta Polytechnica Hungarica, 18(4), 143–160, 2021.