**OPTIMISATON TECHNIQUES IN SPEECH PROCESSING**

***Réka Trencsényi1 and László Czap2***

1PhD, assistant professor

2PhD habil, professor

*1University of Debrecen, Department of Electrical and Electronic Engineering*

*2University of Miskolc, Institute of Automation and Infocommunication*

Abstract

*The topic of the present study is speech processing based on visual information acquired from dynamic records made by ultrasound (US) and magnetic resonance imaging (MRI) methods. Applying contour tracking algorithms based on dynamic programming, curves are fitted to the contour of the surface of the tongue and the palate. Our aim is to convert the different geometries of US and MRI sources to each other. With that goal in mind, special geometric transformations are performed on the anatomic contour curves by optimising some important parameters.*

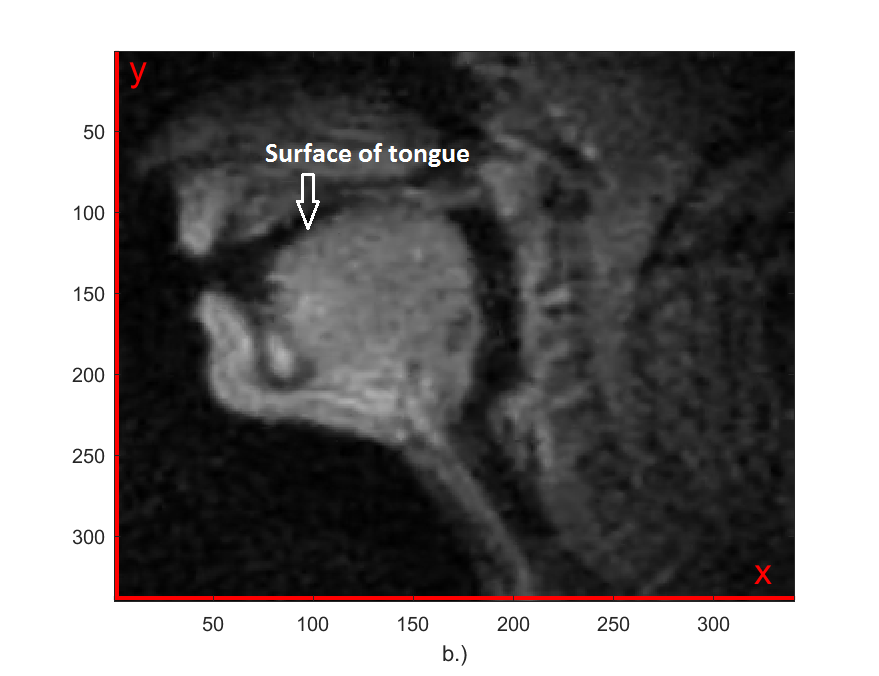
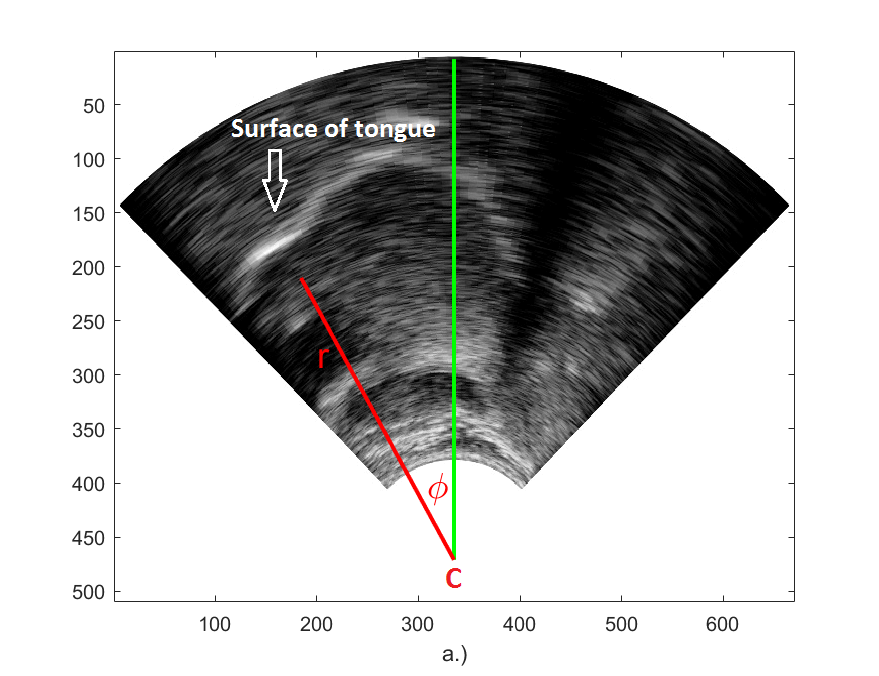
**1. INTRODUCTION**

Speech processing [1] covers an immensely wide range of speech research, which includes the complex study of speech signals and the multifold methods applicable for signal processing. The main aspects of speech processing are the acquisition, manipulation, storage, and transfer of the data that can be accessed from audio or audiovisual sources. In the case of audio sources, one can rely only on acoustic parameters arising from the speech signal without any visual representations. Along this line, the most crucial task is to find appropriate indexing techniques that can important e.g. in speech recognition, speaker segmentation and identification, topic classification, story segmentation, and information retrieval [2]. On the contrary, audiovisual-based speech processing is supported by also visual information that can be obtained from different imaging sources. The vocal organs can be monitored outside or inside the vocal tract by means of that the motion and the relative positions of the vocal organs can be studied.

IDE MÉG JÖN EGY KIS SHOWDER...

**2. ULTRASOUND AND MRI RECORDS**

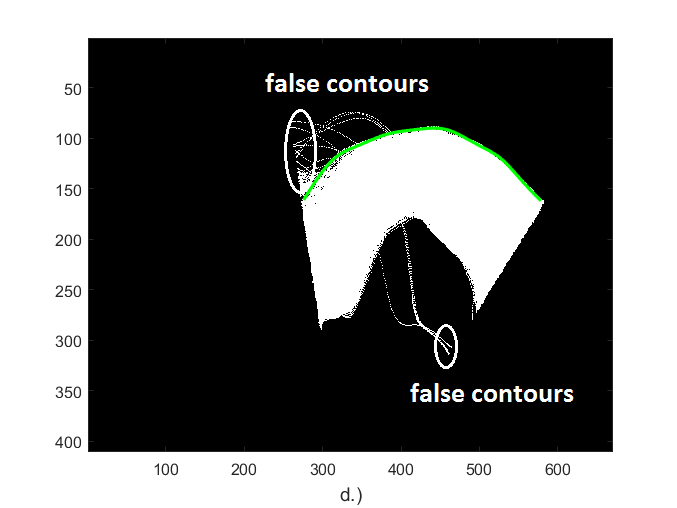
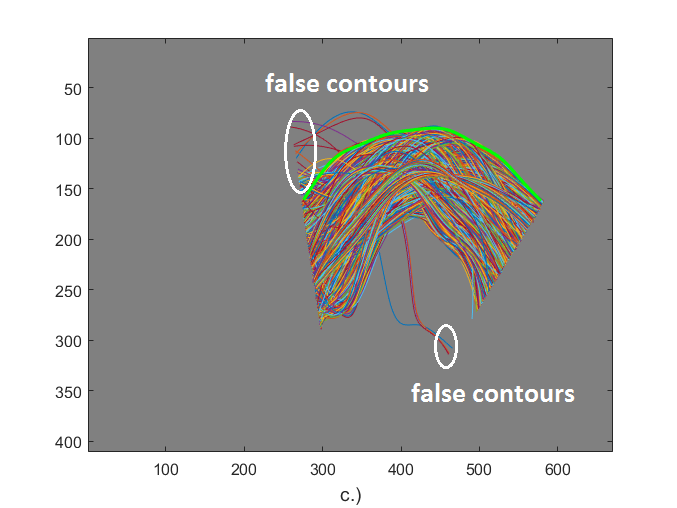
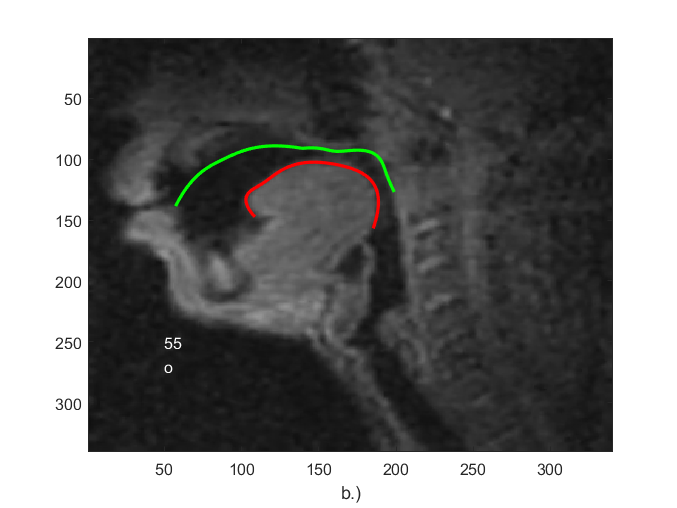
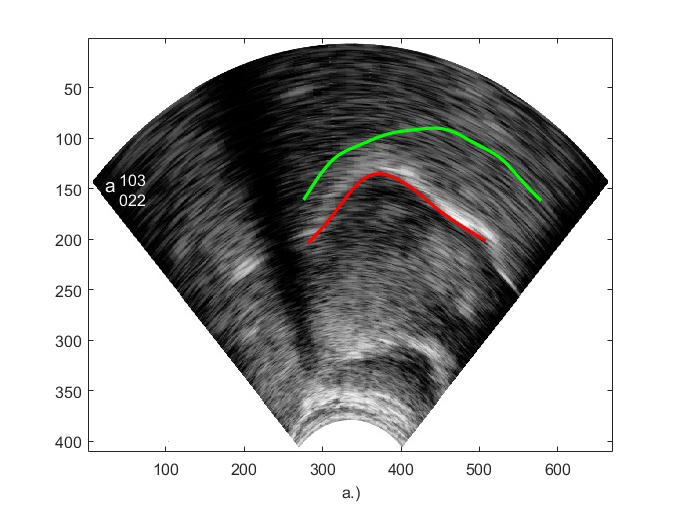
The basic tools of our study were dynamic records made by ultrasound (US) and magnetic resonance imaging (MRI) techniques, which create the opportunity to produce moving pictures during speech. Both methods present a side-view in the mid-sagittal plane, where the motion and the relative positions of the human vocal organs can be observed. A US and an MRI snaphot arising from the dynamic records are visualised in Figure 1.a and 1.b. As seen, in the US frames, the surface of the tongue appears as a bright range that is due to the reflection of ultrasound waves from the air above the tongue, while, in the MRI frames, the surface of the tongue can be identified as a contrast between the dark and bright regions formed by the air and the tongue. Conforming to the radial geometry of the US frames, the positions of the pixels of the images can be described by the polar coordinates (*r,φ*) determining the radial distance measured from point *C* and the angle between the vertical symmetry axis (green line) and the radius (red line). On the contrary, the MRI frames have a rectangular geometry, so the positions of the pixels can be given by the Cartesian coordinates (*x,y*).

Fig. 1.

US (a.) and an MRI (b.) snaphots arising from dynamic records

**3. TONGUE AND PALATE CONTOURS**

Using MATLAB environment, we developed contour tracking algorithms based on dynamic programming for following the US and MRI tongue contours, as well as the MRI palate contour. In the case of the US frames, however, we needed to elaborate another procedure for searching for the palate contour, since it can not be detected in the US records directly as a sharp boundary. Therefore, we found the US palate contour in the framework of a special extremum problem. US and MRI tongue contours are exemplified and drawn by red curves in Figure 2.a and 2.b in the case of sounds *a* (a.) and *o* (b.), respectively. Additionally, also the US and MRI palate contours are marked by green curves. Figure 2.c and 2.d show all tongue contours and the corresponding density map collected in one block that includes all sounds of all US records. The palate contour is indicated by green curve, which was determined as the envelope of the set of the correct tongue contours after the false contours bounded by white circles were ignored. The envelope of the family of curves was ascertained as the series of the points touched by the tongue at the highest possible positions during speech.

Fig. 2.

US (a.) and MRI (b.) tongue and palate contours drawn by red and green curves.

All tongue contours (c.) and the corresponding density map (d.) together with the determined palate contour.

**4. GEOMETRIC TRANSFORMATIONS**

Our aim was to harmonise the radial and rectangular geometries of the US and MRI sources by dint of the tongue and palate contours generated in the previous section. For this purpose, we created special transformations, with which a given US tongue contour or the US palate contour can be converted to the corresponding MRI frame, and vice versa by the inverse transformations. The transformations consist of three fundamental operations that are the scaling of the radial and angular ranges and the translation of the angular range. The scaling factor of the angular range was kept unique, which means that the transformations does not modify the width of the angular range, in other words, the tranformations are conformal. Thus, the scaling factor of the radial range and the translation term of the angular range remain as changing parameters. These two operations are demonstrated graphically by Figure 3, where a US and an MRI tongue contour appear with blue and red curves for the sake of clarification. The starting point *C* designated in Figure 3.a stands for the centre of all transformations. Figure 3.b illustrates the radial directions along which the blue curve is magnified by scaling factor *R*. Finally, Figure 3.c depicts the rotation of the blue curve in the plane of the figure by translation term *FI0*. Thus, parameters {*C,R,FI0*} are optimised such a way that the distance between the two curves must be minimal.

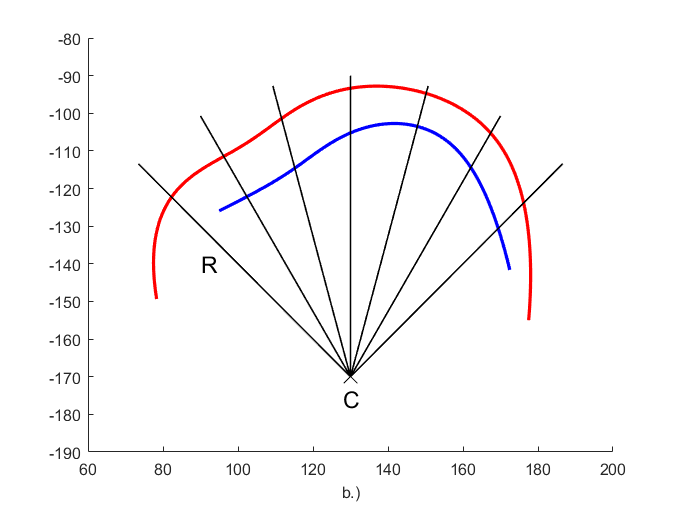
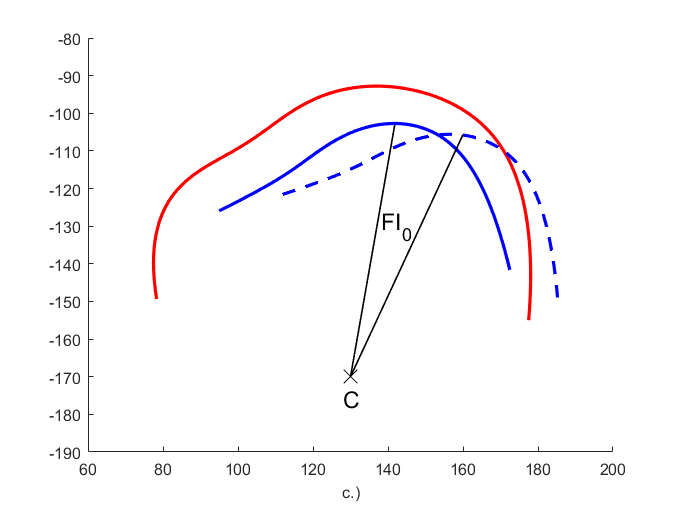
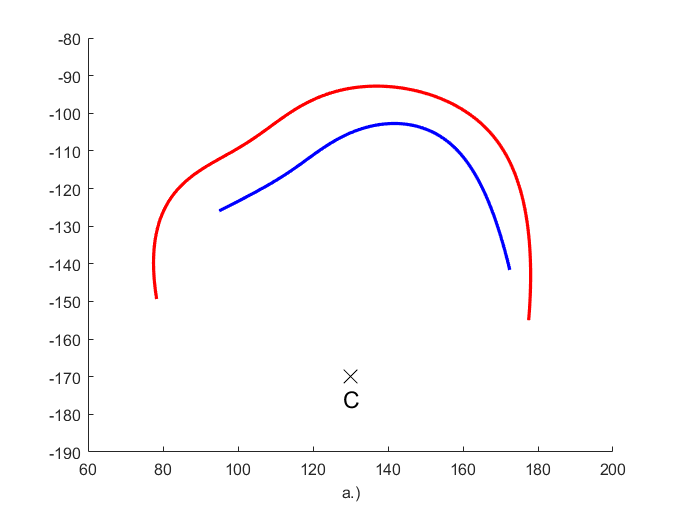


Fig. 3.

The graphical representation of the geometric tranformations: the centre of the transformations (a.), the scaling of the radial range (b.), the translation of the angular range (c.)

**5. OPTIMISATION PROCEDURE AND RESULTS**

The optimisation procedure was carried out for the set of tongue contours belonging to speech sounds *S*={*a,e,k,s,t*} supplemented by the palate contour. It means that parameters {*C=*[*y,x*]*,R,FI0*} from the previous section were optimised for the assembly of 6 contours simultaneously. In the first step, the average curve of all tongue contours belonging to the given sound was computed, then the tongue contour nearest the average was selected for all 5 sounds of set *S*. The optimisation algorithm was implemented for the selected tongue contours and the palate contour resulting in the parameter values contained by Table 1, where *C* denotes the centre of the transformations in the MRI images, and its Cartesian coordinates are given in pixel units. Magnification *R* is a dimensionless quantity, rotation *FI0*is expressed in radians, and also the minimal average distance between the linked US and MRI contours is displayed in the last column of Table 1.

Table 1.

The results of the optimisation for the first stage

|  |  |  |  |
| --- | --- | --- | --- |
| C | R | FI0 | MIN |
| [236,127] | 0.3612 | 0.25 | 58.8324 |

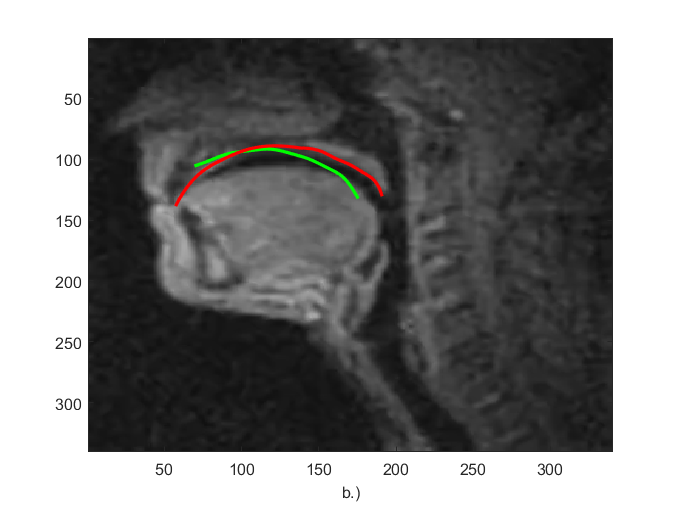
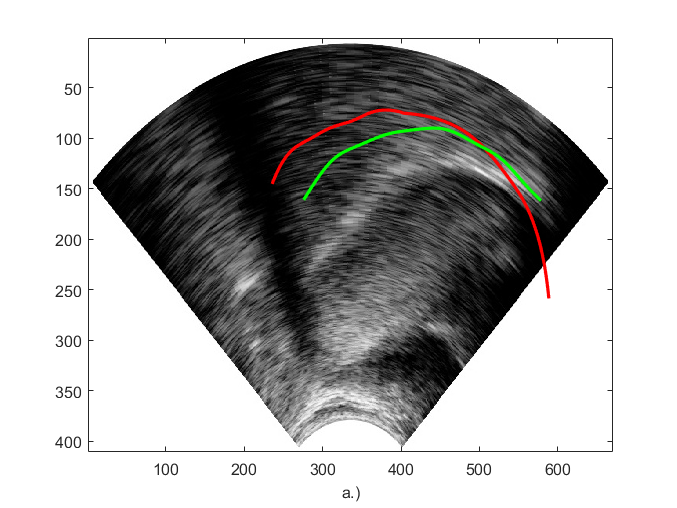
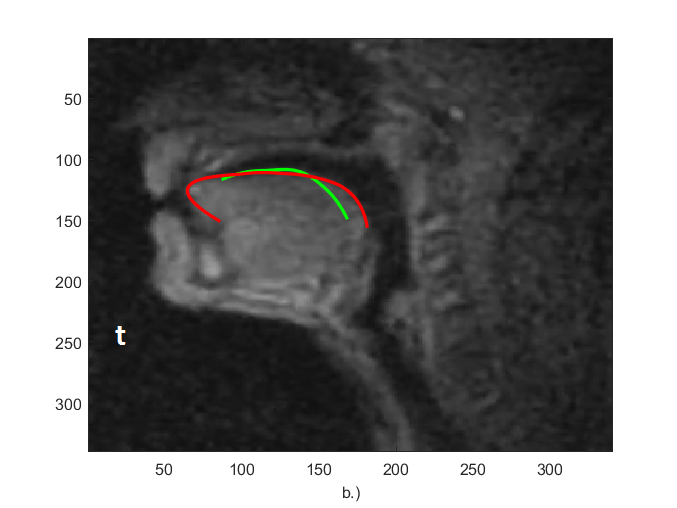
In the next step, by setting the parameters of Table 1, the US tongue contours of the sounds of set *S* nearest the corresponding MRI tongue contours were selected from all US tongue contours of the given sound, and after that, the optimisation procedure was repeated for this new set of the 5 tongue contours adding again the palate contour, as well. The results are summarised by Table 2, where it can be observed that, although parameters {*C,R,FI0*} keep their values compared to Table 1, the minimal average distance between the two curves is much smaller than in the previous case, which shows that the optimisation works more effectively in the second stage. As we obtained the same parameter set for {*C,R,FI0*} in the first two rounds, the process of the above type of iterations can not be continued because, by setting the parameters of Table 2, we would reobtain the same US tongue contours that were selected in the second stage.

Table 2.

The results of the optimisation for the second stage

|  |  |  |  |
| --- | --- | --- | --- |
| C | R | FI0 | MIN |
| [236,127] | 0.3612 | 0.25 | 37.2658 |

After the quantitative analysis, we investigated our results qualitatively, as well, by projecting the US tongue contours and the palate contour to the corresponding MRI frames, and vice versa. The results are presented by Figure 4 through the examples of sound *t* and the palate contour. The US contours are drawn by green curves, and the MRI contours are marked by red curves. Blocks a.) illustrate the transformations of the MRI contours to the US frame, while blocks b.) visualise the transformations of the US contours to the MRI frames. Based on the images, it can be stated that the visual match of the US and MRI contours is acceptable.

Fig. 4.

The bidirectional transformations of the US and MRI contours in the case of sound *t* and the palate contour

Besides set *S*, we studied the visual projections of such tongue contours, as well, that were not elements of set S, using the optimisation parameters obtained in Tables 1-2. We experienced that the visual comparison of the US and MRI contours shows qualitatively the same fitting as in Figure 4.

**6. SUMMARY**

In this study, we implemented special geometric transformations and optimisation procedures for tongue and palate contours gained from dynamic US and MRI records of human speakers by contour tracking algorithms or solving extremum problems. All programs were written in MATLAB environment. By the geometric transformations, the US and MRI contours can be mapped to each other in a bidirectional and biunique manner, so it leads to the result that the radial geometry of the US images and the rectangular geometry of the MRI images can be embedded into each other.

**REFERENCES**

[1] RABINER, L. R., SCHAFER, R. W.: **Introduction to digital speech**

**processing.** Now Publishers Inc, 2007.

[2] MAKHOUL, J., et al.: **Speech and language technologies for audio indexing**

**and retrieval.** Proceedings of the IEEE, 2000, 88.8: 1338-1353.