Vocal tract sagittal slices estimation from MRI midsagittal slices during speech production of CV

Ioannis K. Douros[1,2], Yu Xie[3] , Chrysanthi Dourou[4], Jacques Felblinger[5], Karyna Isaieva[2], Pierre-André Vuissoz[2], Yves Laprie[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,
[2]Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France
[3]Department of Neurology, Zhongnan Hospital of Wuhan University, Wuhan, 430071, China
[4]School of ECE, National Technical University of Athens, Athens 15773, Greece
[5]Université de Lorraine, INSERM 1433, CIC-IT, CHRU de Nancy, Nancy, F-54000, France

ioannis.douros@loria.fr, xieyuyy@163.com, chrysanthi.dourou@gmail.com, jacques.felblinger@univ-lorraine.fr, karyna.isaieva@univ-lorraine.fr, pa.vuissoz@chru-nancy.fr, yves.laprie@loria.fr

Objectives

Even though midsagittal slices are informative, there are many cases where they do not give enough information about what is happening to the non-midsagittal planes which are important from an acoustic point of view like in the case of /u/ or /i/. In this work we propose an algorithm for estimating vocal tract para sagittal slices in order to have a better overview of the behaviour of the articulators during speech production. A transformation that connects the midsagittal dynamic slices of the vocal tract with the neighbouring frames was acquired at the image level. The transformation then was adapted to the midsagittal frames of a new speaker to give the synthesized neighbouring frames. To evaluate the results, image cross-correlation between the original and the estimated frames was used. Results show good agreement between the original and the estimated frames.

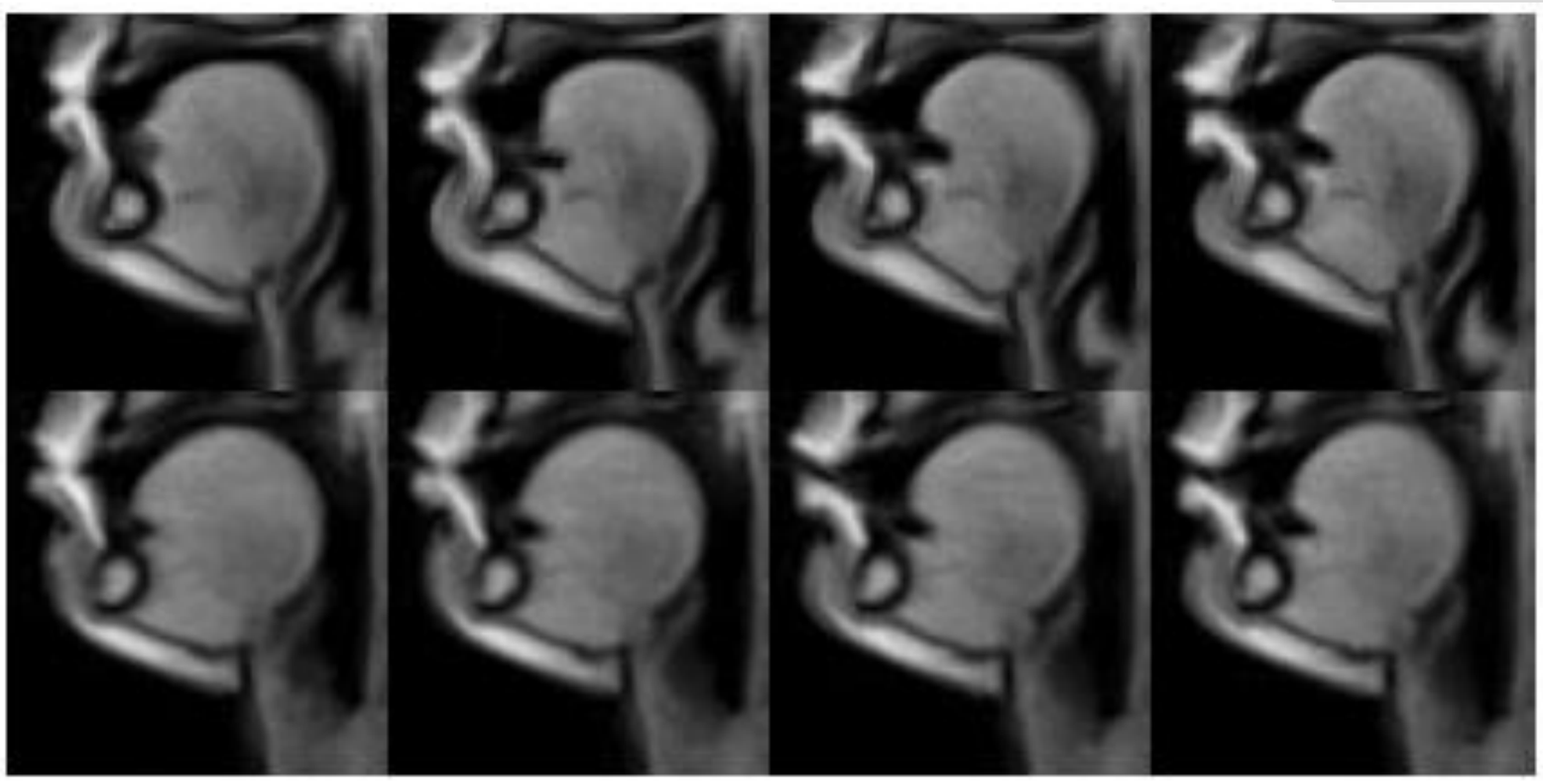

Fig. 1: Selected right frames of /pu/. Top: synthesized images and bottom: corresponding original ones.

Materials and Methods

Recordings of two French subjects (one male, one female) in the supine position were performed in a 3T MRI (Prisma, Siemens, Erlangen, Germany) with a 20-channel head and neck antenna.

An echo-dispersed, Cartesian T1-weighted echo gradient sequence VIBE (Volumetric Interpolated Breath-hold Examination) was used for the 3D recordings while for the 2D real time recordings we used radial RF-spoiled FLASH sequence.

12 CV syllables (combination of C={/f/,/p/,/s/,/t/} with V={/i/,/a/,/u/}) were used. The chosen planes were the midsagittal (M) its left (L) and right (R) adjacent planes. Data in each plane were acquired in different acquisition on the same session.

A non-rigid image transformation method [1] was used, based on the displacement field between the images. To measure the image similarity, histogram matching between the images is applied and then the mean square error of the pixels intensity is computed. To validate the results, cross-correlation between the synthesized and the original images, normalized by the autocorrelation of the original images, was used.

Algorithm Description

The algorithm's input is a 2D MRI CV from dynamic speech and gives as an output the corresponding L and R planes of every frame of the CV.

To create the single speaker model of a target CV, frame alignment between M frames of target speaker and M, L, R frames of the train speaker was applied using target speaker's frames as reference. The output was the frame aligned sequences of all planes of the train speaker (Ma, La, Ra). Two sets of image transformations TL, TR between Ma-La and Ma-Ra were computed that transformed the Ma frames to the corresponding La or Ra frames respectively. Additionally, another set of image transformations A was calculated that transforms each frame of Ma to the corresponding midsagittal frame of the test speaker. Since TL and TR were derived from the train speaker they can not be directly used on the test speaker. A was applied to TL and TR in order to adapt them to the test speaker space. The output of this step is TLa and TRa transformations. Finally, these transformations are applied to the dynamic images of the test speaker on the M plane to get the final estimations on Left and Right planes.

Results

In Fig. 1 we can see chosen images of the right side during /pu/ in both synthesized and the corresponding original form. Synthesized images have average correlation of 93.54% (± 1.06%) with the original ones, using normalized image cross-correlation. By visually inspecting the images, we can see that a small difference appears sometimes at the front part of the hard palate and some small differences in the eccentricity of the tongue. Apart from this point, images look quite similar in terms of vocal tract shape with an exception is a few cases were a small artifact may appear mainly at the front region of the tongue due to the existence of a similar artifact to the corresponding training images

Discussion and conclusion

One point is that since training and test speakers were of different gender thus with bigger anatomical differences than same gender subjects it is expected one to notice minor differences in the original and the synthesized images. A second point is that one can use standard automatic procedures to label the images at the preprocessing step using simultaneous audio recordings in which case our algorithm will run fully automatically. Further research could include exploring other ways to create the model or methods to better adapt it to a new speaker.

References

[1] T. Vercauteren et al. "Diffeomorphic demons: Efficient non-parametric image registration," NeuroImage, 2009.
[2] Y. Lim et al. "3D dynamic mri of the vocal tract during natural speech," MRM, 2019.
[3] I. Douros, et al. "Towards a method of dynamic vocal tract shapes generation by combining static 3d and dynamic 2d mri speech data," INTERSPEECH, 2019

N° ANR-15-CE23-0024-02