MRI vocal tract data during CV production

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

The main purpose of the presented algorithm is to enlarge MRI speech corpus by synthesising data. In this work, we use a method [2] that captures the dynamics of speech during CV production by using non-rigid image transformations. This information is adapted and then applied to a target speaker in order to synthesise its CV production data using only his silence frame. We evaluated the performance of the proposed method using image cross-correlation in which we compared the original images of the target speaker pronouncing the same CVs with synthesized images.

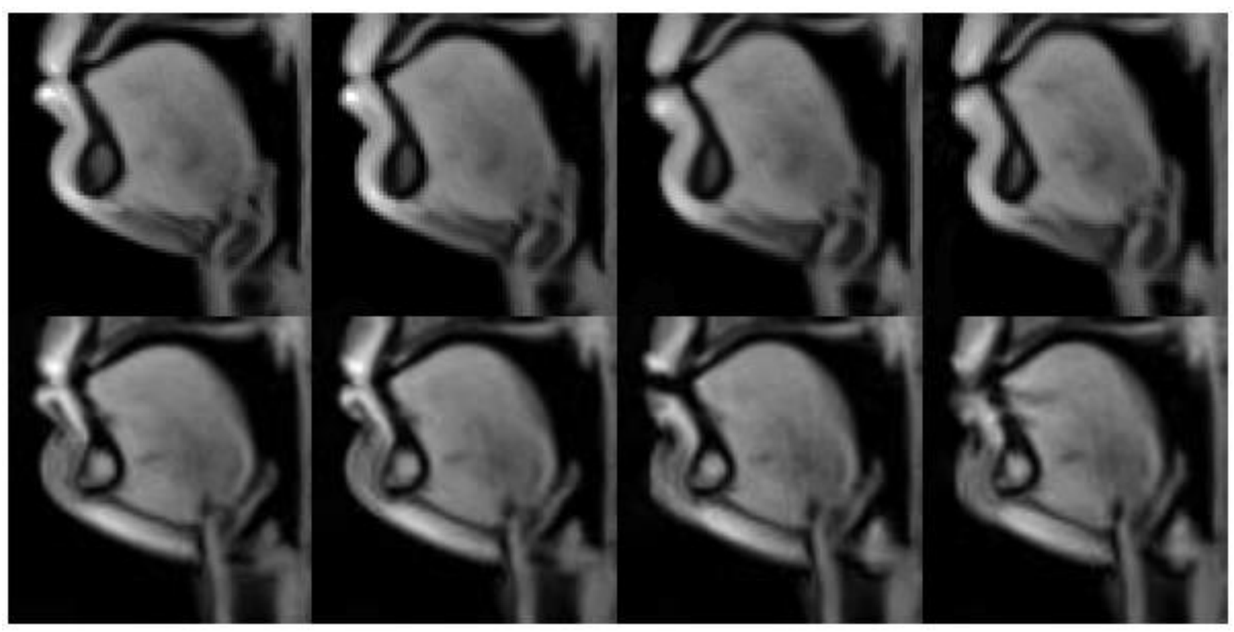

Fig. 1: Selected frames of /pi/. Top: synthesized images and bottom: corresponding original ones.

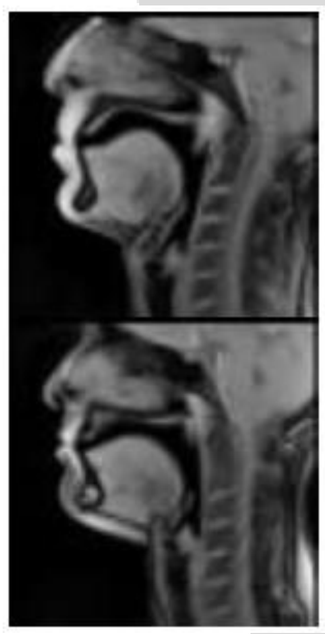

Fig. 2: Silence frame. Top: train frames and bottom: test frame

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 input of the algorithm is a silence frame of both train and target speaker and the rtMRI data of the target CV.

An image transformation is computed from each CV frame to the next one, creating a set of transformations that describe the dynamics of the CV production. Another image transformation is computed from the silence frame of train speaker to the silence frame of the target speaker and is used to adapt the set of transformations computed previously to the target speaker. The adapted set of transformations is applied to the silence of the target speaker to synthesise his/her CV pseudo rtMRI data. Synthesised images are compared with the original ones using image cross-correlation. Results show good agreement between the synthesised and the original images.

Results

In Fig. 1 we can see chosen images during /pi/ in both synthesized and the corresponding original form. Synthesized images have average match of 94.37% (± 0.96%) with the original ones over the set of syllables studied, using normalized image cross-correlation. By visually inspecting them, we can see that some difference appears at the back part of the tongue, which is a little flatter. Additionally, lip protrusion is weaker on the upper lip and some artifacts appear sometimes at the level of the epiglottis. Apart from this, 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 region of the tongue due to the existence of a similar artifact in the corresponding training images. However, there are also cases that synthesized images had less artifacts and were smoother compared to the original ones.

Discussion and conclusion

Some differences can be observed in the images mainly due to different articulation and vocal tract shapes. Additionally, the algorithm uses the same phoneme duration as the train speaker since it does not take into account any known information about the speaking style of the test speaker. Further research could include learning this registration more globally, or by using DNN learning techniques or other ways to implement the duration model.

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