Synthesize MRI vocal tract data during CV production

Ioannis Douros, Chrysanthi Dourou, Yu Xie, Jacques Felblinger, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie

▶ To cite this version:

Ioannis Douros, Chrysanthi Dourou, Yu Xie, Jacques Felblinger, Karyna Isaieva, et al.. Synthesize MRI vocal tract data during CV production. ISSP 2020 - 12th International Seminar on Speech Production, Dec 2020, Providence / Virtual, United States. hal-03090873

HAL Id: hal-03090873

https://hal.inria.fr/hal-03090873

Submitted on 30 Dec 2020

Synthesize MRI vocal tract data during CV production

Ioannis K. Douros[1,2], Chrysanthi Dourou[4], Yu Xie[3], 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

1 Synopsis

A set of rtMR image transformations across time is computed during the production of CV that is afterwards applied to a new speaker in order to synthesize his/her CV pseudo rtMRI data. Synthesized images are compared with the original ones using image cross-correlation.

2 Purpose

To be able to enlarge MRI speech corpus by synthesizing data.

3 Introduction

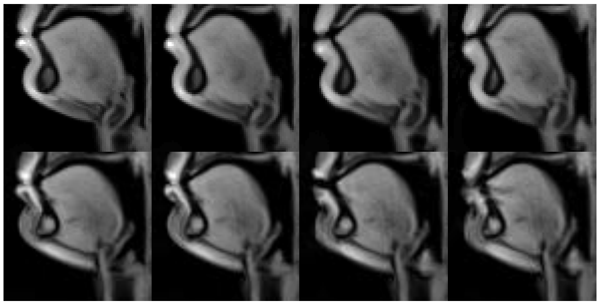
Magnetic resonance imaging (MRI) is extensively used in many speech related fields like speech production, articulatory speech synthesis etc since it provides significant advantages over other competitive modalities like x-ray, ultrasound or EMA. Its non-invasive nature, its ability to cover the whole vocal tract with good spatiotemporal resolution along with the fact that there are no known health hazards, has made MRI one of the preferred ways for speech scientists to collect data [1]. However, as with almost all methods for collecting articulatory data, it is quite hard to acquire a lot of hours of articulatory data compared to speech only data (that can have even more than 400 hours of speech like in case of WSJ), which can put limitations on how complex the potential models can be and their performance. Under this scope, it could be interesting to be able to artificially synthesize articulatory data that will enlarge the existing databases and offer new potential to speech studies. In this work, we propose a method that captures the dynamics of CVs by using some image transformations and adapts these transformations to a target speaker in order to synthesize the data of the target speaker pronouncing the training CVs. Synthesized images were compared to the original images of the target speaker pronouncing the same CVs using image cross-correlation.

4 Materials and Methods

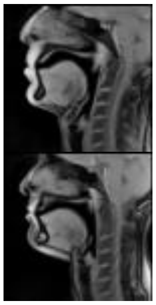
In order to evaluate the interest of the method we chose a difficult case, with a male and a female speaker, thus with bigger anatomical differences than same gender subjects. One male and one female subjects were asked to repeat /pi/,/pa/,/pu/,/si/,/sa/,/su/. Before the pronunciation of every CV, subjects were instructed to breath from the nose with closed mouth. The midsagittal plane was chosen for the acquisition. The data were acquired on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) in CHRU of Nancy under the approved ethics protocol (ClinicalTrials.gov NCT02887053). The sequence used is real-time (50 fps) MRI Flash sequence [2]. In total 2000 images (including silence) were acquired. The first step of the algorithm is to calculate a non-rigid image transformation T_t, that transforms every time frame of the train speaker to the next one. To compute the non-rigid image transformation a MATLAB function which is based on the algorithm described in [3] was used. The next step is to compute a non-rigid transformation $T_{train-test}$ which transforms the first frame of the train speaker to first frame of the test speaker. The next step is to adapt T_t transforms to the test speaker by transforming them using $T_{train\text{-}test}$ transform. The newly created transformation is then applied to the first frame of the test speaker and then propagated to every newly synthesized frame. The last step is to map the corresponding training time frames to the synthesized ones to suppress some artifacts that are created due to the transformations. In order to validate the results, cross-correlation between the synthesized and the original images [4], normalized by the autocorrelation of the original images, was used.

5 Results

Below 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.



Selected frames of /pi/. Top: synthesized images and Bottom: corresponding original images.



Silence frames. Top: train frame and Bottom: test frame

6 Discussion

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.

7 References

[1] Douros, Ioannis K., et al. "A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research." (2019).

[2] Uecker, Martin, et al. "Real-time MRI at a resolution of 20 ms." *NMR in Biomedicine* 23.8 (2010): 986-994.

[3] Vercauteren, Tom, et al. "Diffeomorphic demons: Efficient non-parametric image registration." *NeuroImage* 45.1 (2009): S61-S72.

[4] Woo, Jonghye, et al. "A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.5 (2018): 520-531.