*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Using Silence MR Image to Synthesise Dynamic MRI Vocal Tract Data of CV

*Ioannis K. Douros* [1,2], *Ajinkya Kulkarni* [1], *Chrysanthi Dourou* [3], *Yu Xie* [4], *Jacques Felblinger* [2,5], *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]School of ECE, National Technical University of Athens, Athens 15773, Greece
[4]Department of Neurology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China
[5]Université de Lorraine, INSERM 1433, CIC-IT, CHRU de Nancy, F-54000 Nancy, France

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

## Abstract

In this work we present an algorithm for synthesising pseudo rtMRI data of the vocal tract. rtMRI data on the midsagittal plane were used to synthesise target consonant-vowel (CV) using only a silence frame of the target speaker. For this purpose, several single speaker models were created. 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 from multiple single speaker models are frame aligned and then averaged to create the final version of synthesised images. Synthesised images are compared with the original ones using image cross-correlation. Results show good agreement between the synthesised and the original images.

**Index Terms**: speech resources enrichment, pseudo rtMRI synthesis, image transformation, rtMRI data, vocal tract

## 1. Introduction

Regardless of the growth in real-time magnetic resonance imaging (MRI) techniques, research in speech production and modeling of the vocal tract faces limitations in building a complete articulatory synthesis model [1]. Articulatory data acquired using rtMRI techniques eased the in-depth analysis of human physiology and the movement of articulators during speech production. Such research activity made it possible to fill in the gap between speech production and its relationship to its linguistic aspects [2] like a better understanding of the existence of voiced fricatives.

Acquisition of articulatory data raises several issues such as the capability to extract precise in time and space speech dynamics, interpretation of acquired articulatory data, easiness and safety standards for the subjects. As the usage of MRI techniques provided detailed natural images of articulators without any known health hazard to the subject, they represent valuable techniques against others such as X-ray [3], electromagnetic articulography [4, 5], electropalatography [6] and ultrasound [7, 8].

Usually, in the current acquisition protocol 3D MRI images, the vocal tract position needs to be held motionless over the acquisition time. This way detailed images of the vocal tract can be recorded. However, those images correspond to frozen vocal tract configurations due to a long acquisition time (more than seven seconds). On the other hand, vocal tract images recorded with rt-MRI yield natural and complex information about articulatory spatiotemporal movements. The rtMRI protocol selects only one slice usually the midsagittal plane and captures tissues within the midsagittal slice at 50 Hz approximately in real time [9]. The major benefit of capturing rtMRI images is that it provides a considerable amount of data which suffices to analyze continuous rapid speech articulator movements [10, 11, 12].

The recent development in rtMRI imagining techniques provide tools to examine phonetic and phonological phenomena. There is a vast range of work but we can mention, for instance, vowel nasalization in Portuguese and French [13], coarticulation in VCV sequences [14], characterization of click consonants in African languages [15] ... Besides investigation in phonetics, rtMRI can have a big impact on automatic speech and speaker recognition to supplement the acoustic signal with the structure of the physical system and consequently increase the performance of recognition systems [12].

As discussed earlier, rtMRI acquisition of vocal tract data is a long process in terms of finding appropriate and available equipment, designing a recording protocol, selecting subjects, recording data and annotating the dynamics of speech articulators in films. Furthermore, the acquisition of articulatory data presents constraints in acquiring ”global” information like 3D dynamic rtMRI with high spatiotemporal resolution to capture vocal fold activity. Even though there are some attempts trying to address these issues [16, 17], it could be still interesting to be able to artificially synthesize articulatory data that could enlarge existing databases and make speech production studies easier.

In this work, we use a method [18, 19] 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.

http://dx.doi.org/10.21437/Interspeech.2020-1173

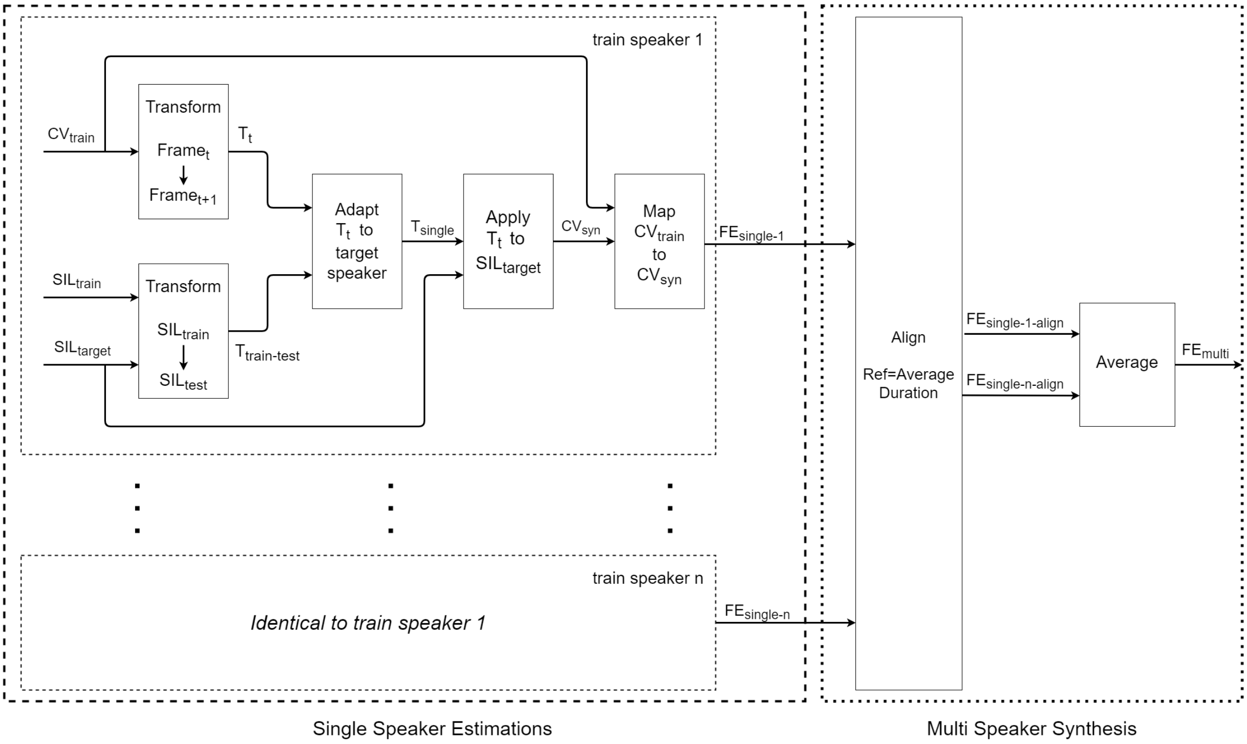


Figure 1: *Visual representation of the proposed algorithm*

## 2. Materials and Methods

The proposed algorithm can be divided into two main parts: a) estimation of the CVs of the target speaker using image transformations on the data of one train speaker to create single speaker estimation model b) combine single speaker models of all train speakers to create the final multi speaker based synthesized data. A visual representation of the algorithm can be seen in Fig 1.

### 2.1. Data Acquisition

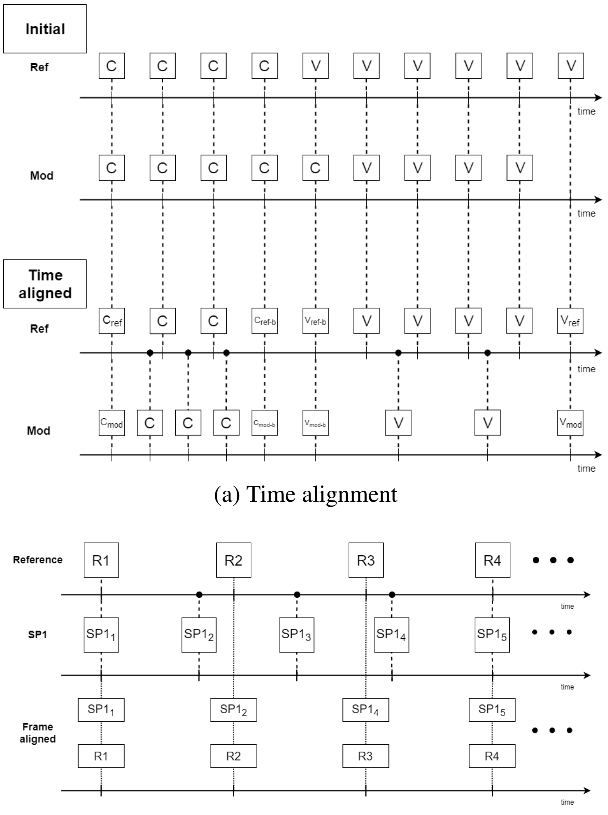
For this work, rtMRI recordings of eight (four male, four female) native speakers of French were used. The age of subjects was between 21 and 36 years old with average age of 27.25 Subjects had no previous speaking or hearing problems recorded. The data was acquired on Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) located in Nancy Central Regional University Hospital under the approved medical protocol “METHODO” (ClinicalTrials.gov Identifier: NCT02887053). For acquiring dynamic data, we used a 2D rtMRI sequence. In our approach, we used radial RF-spoiled FLASH sequence [9, 20] with TR $= 2.22$ ms, TE $= 1.47$ ms, FOV $= 19.2 \times 19.2$ $cm^2$, flip angle $= 5$ degrees, and slice thickness is 8 mm. Pixel bandwidth is 1670 Hz/pixel. The number of radial spokes is 9, and the resulting image resolution is $136 \times 136$. The acquisition time was 44 sec. Images were recorded at a frame rate of 50 frames per second using a 64 channel head-neck antenna.

### 2.2. Image transformation

Image transformations can be generally divided into two categories, the rigid and the non-rigid ones. One the one hand, rigid image transformations are faster and simpler and can capture well global differences between images like rotation or translation. On the other hand, non-rigid transformations are more complex and computationally heavier, but they can better describe differences locally in images. Since anatomical and articulation differences between speakers are local, we used a non-rigid image transformation method, based on an adaptation of demons algorithm for image registration [21]. To calculate the transformation between two images a displacement field is computed using the algorithm described in [22]. In order to measure the image similarity between the transformed and the target image, histogram matching between them is applied and then the mean square error of the pixels' intensity is computed.

### 2.3. Frame alignment

There are various ways that one can align dynamic sequences between them like applying different types of interpolation at one sequence to match the reference one. A disadvantage of such approaches is that they are quite complex and quite hard to be applied to rtMRI data of the vocal tract. Given sufficiently high acquisition frame rate, one can use piecewise linear alignment as we did in this work. The core idea is that some samples (let's call them boundaries) of the two sequences correspond at the same time points/events (like the beginning or the end of a phoneme) and the rest samples of the non-reference sequence are linearly compressed or extended in time. In order to align the frames, the frame of the modified sequence that is temporarily closer to each frame of the reference sequence is selected as the corresponding matching frame Fig. 2.



(a) Time alignment

(b) Frame alignment

Figure 2: *The two-step procedure of frame alignment. Linear piece wise alignment is applied between the reference (Ref) and the modified sequence to adjust the length and then the closest frame to each of the reference frames are selected from the modified sequence (SP1 in the example of the figure)*

# 3. Experiments

## 3.1. Speech Task

In this work we studied 12 CV syllables (/fi/, /fa/, /fu/, /pi/, /pa/, /pu/, /si/, /sa/, /su/, /ti/, /ta/, /tu/). Midsagittal rtMRI images of the vocal tract were acquired during the phonation of the CVs from the subjects. In order to remove coarticulation effects from previous CVs, subjects were instructed before each CV to breath from the nose with mouth and lips closed so that the vocal tract is returned back to the ”neutral” silence position. To constrain the vowel at the end of the CV, subjects were also instructed to pronounce /p/ at the end of each CV. Therefore in practice, subjects were pronouncing /sil//C//V//p/.Even though images could have been automatically labelled, we chose to manually label them to achieve better temporal accuracy.

## 3.2. Single Speaker Estimation

First, one train speaker is used in order to create the model for the target CV synthesis of the test (target) speaker. A silence frame was kept at the beginning of each CV labeling. For the rest of the work, when we refer to CV we mean the CV with the silence frame at the beginning Fig. 3. For all the image transformations in this work (both at this and at the later steps) we used MATLAB imregdemons function with 3 pyramid levels with values $100, 50, 25$ for the image resolution and accumulated field smoothing of $1.3$ for the smoothing of the deformation field. We also applied histogram matching before the image transformation to have a similar contrast between the images, in the cases that we transformed images of different subjects.

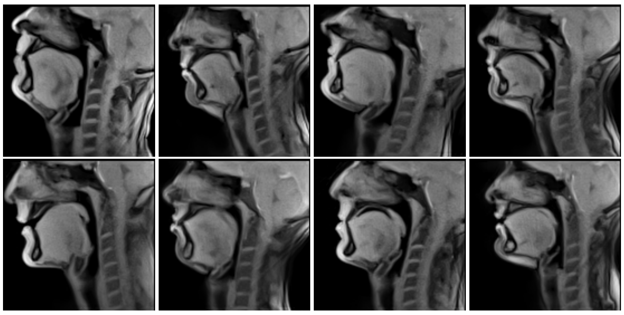


Figure 3: *Silence frames of eight speakers. One can notice differences in speakers' anatomy and heads' position*

A set of non-rigid image transformation $T_t$ ($t = 1, 2, .., \#CVframes - 1$) was computed that transforms every time frame of the CV of the train speaker to the next one. Another image transformation $T_{train-test}$ was also calculated that transforms the silence frame of the train speaker to the silence frame of the test speaker. The next step was to use $T_{train-test}$ to adapt the set of $T_t$ transformations from the domain of the train speaker to the domain of the test speaker. We call this set of transformations $T_{single}$. $T_{train-test}$ was applied to the silence frame of the train speaker to synthesise the silence of the test speaker. $T_{single}$ was then applied to the synthesised silence frame of the test speaker and propagated to every newly synthesised frame until all CV frames are synthesised. Synthesised images at this stage have some artifacts due to the adaptation of the transformation and various transformations that were applied. In order to suppress them, the training images are mapped to the synthesised ones to create the single speaker CV frame estimation ($FE_{single}$). This process removes some artifacts due to the small smoothing of the deformation field during the last transformation which suppresses some abnormalities at a very local level.

## 3.3. Multi Speaker Synthesis

At this step, the process described in subsection 3.2 is repeated for all the $N$ speakers in the training set, therefore $N$ $FE_{single}$ are acquired. Let's call them $FE_{single-n}$. The duration of $FE_{single-n}$ (in terms of image frames) is directly derived from the duration of the CV of the corresponding training speaker. Since every train speaker speaks at a different pace during data acquisition, every $FE_{single-n}$ has different length. In order to combine them, a specific duration is required to be used as a reference. Ideally, it would be the CV duration of the target speaker but since we consider this information unavailable for the purposes of this algorithm, the average duration of all $FE_{single-n}$ was selected. We note here that all $FE_{single-n}$ have kept the labels from the corresponding training speakers. This information was taken into account during the calculation of the reference duration which in fact is equal to the average frames of C plus average frames of V. Averages were rounded to the closest integer because they refer to image frames. Linear piecewise alignment was used to time align all $FE_{single-n}$ with the reference. C and V parts of $FE_{single-n}$ CV are in-

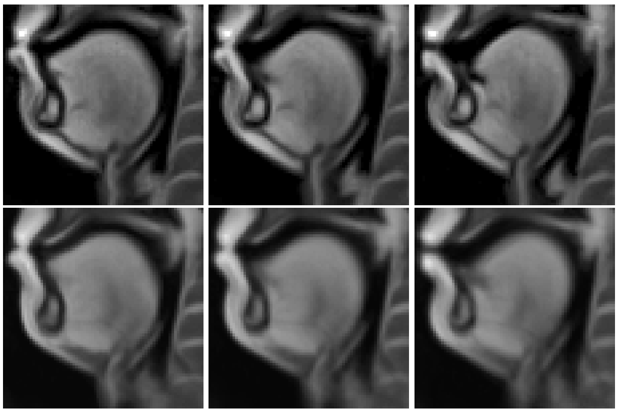


Figure 4: *Selected frames for /pu/ of speaker 6. Top: original images; Bottom: synthesised images*

dependently and linearly extended or compressed until the duration of both C and V of it matches with those from the reference CV. The resulting $FE_{single-n}$ are time aligned but not frame aligned, which means that they have the same duration with the reference (in terms of time) but the time between each frame is not stable as their amount varies. Since the output of the algorithm is the synthesis of pseudoMRI data, the resulting sequence should have stable time difference between neighbouring samples based on the rtMRI frame acquisition. Using the reference duration, the desired time points of the samples are computed and the frames of the time aligned $FE_{single-n}$ that are closer to the time points are kept. If a frame is closer to two time points it is copied and kept twice. The resulting frame estimations $FE_{single-n-align}$ have all the same number of frames and at the same time points. In order to get the final multi speaker frame estimation $FE_{multi}$ we average the the images of all $FE_{single-n-align}$ across each frame.

### 3.4. Evaluation

For the purpose of evaluation, we used 8-fold cross validation using seven speakers for train and one for test on the 12 studied CVs. Original frames $OF$ of CVs of the test speaker were compared with $FE_{multi}$. $FE_{multi}$ and $OF$ were aligned with the procedure described in subsection 3.3 using $OF$ as a reference. To each frame of the aligned sequence $FE_{multi-align}$, histogram matching was applied in order to match with the corresponding frame of $OF$. In order to validate the results, we used cross correlation between the corresponding images of the two sets, normalized by the autocorrelation of the corresponding $OF$ images. Total average after cross validation of correlation coefficient for all CVs is $0.9361$ with standard deviation (sd) of $0.0046$. Detailed results can be seen in Table 1.

Table 1: *Cross validated results of correlation coefficient per CV. Total average for all CVs is* $0.9361$ *with sd* $0.0046$

| CV | /fi/ | /pi/ | /si/ | /ti/ |
|---|---|---|---|---|
| mean | 0.9402 | 0.9384 | 0.9356 | 0.9373 |
| sd | 0.015 | 0.0149 | 0.0152 | 0.0153 |
| **CV** | **/fa/** | **/pa/** | **/sa/** | **/ta/** |
| mean | 0.9404 | 0.9317 | 0.931 | 0.9363 |
| sd | 0.0103 | 0.0158 | 0.0161 | 0.0145 |
| **CV** | **/fu/** | **/pu/** | **/su/** | **/tu/** |
| mean | 0.9388 | 0.9254 | 0.9374 | 0.9404 |
| sd | 0.0131 | 0.0162 | 0.0161 | 0.0107 |

## 4. Discussion

By visually examining the results in Fig 4, one can notice that synthesised images look quite similar to the original ones. However, synthesised images are more blurry because of the averaging of the training deformation fields of the training speakers at the multi speaker synthesis step of the algorithm. Additionally, there is something visually similar to the shadow effect that appears at the tongue in the synthesised images. There are two main reasons responsible for this behaviour. The first one is that since only a silence frame was used for synthesising the images, every small error on the transformations that starts from the first frames of the C is further propagated and stacked with further errors until the last frames of V. This is further supported by the fact that the ”fake” shadow effect is more obvious the further an image is from the beginning. The second reason is that every speaker has quite a different style of speaking which makes the single speaker synthesised images slightly different between them. When images are combined in the multi speaker part, these small single speaker estimation differences are also affecting the ”fake” shadow effect. This effect is more obvious at the front part of the region of the tongue as it is the articulator with the bigger movement in the examined examples, therefore it is more affected by stacking errors and speaking differences between subjects.

Another remark is that apart from the point that mentioned earlier, the regions of the vocal tract like the shape of the palate, the lips, the velum etc appear to visually be very similar between original and synthesised images, which is further supported by the numerical evaluation that gives an average similarity of $0.9361$ between them (maximum value could be 1 which would show identical images). Additionally, the proposed algorithm appears to be quite robust since the matching of the synthesised and the original images is quite high, even though training subjects have very different anatomies and very different head positions during the MRI acquisition as can be seen in Fig. 3 This visual conclusion is again supported by the numerical results of $0.0046$ standard deviation that the average similarity value has.

Finally, by using standard automatic techniques to label images in the beginning, for example by using the audio from simultaneous MRI and sound recordings, this algorithm is fully automated giving flexibility in synthesising CVs of target speaker using only its silence frame. Future directions of this work could be to further examine how the blurriness or the ”fake” shadow effects could be suppressed in order to synthesis better quality images. One could also think of extending this algorithm to synthesise VCV, CVC, whole words or phrases.

## 5. References

[1] Y. Laprie, B. Elie, A. Tsukanova, and P.-A. Vuissoz, “Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech,” in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2110–2114.

[2] B. Elie and Y. Laprie, “Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink,” *Speech Communication*, vol. 82, pp. 85–96, 2016.

[3] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, no. S1, pp. S56–S56, 1990.

[4] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992.

[5] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research." *Phonus.*, 2000.

[6] W. J. Hardcastle, "The use of electropalatography in phonetic research," *Phonetica*, vol. 25, no. 4, pp. 197–215, 1972.

[7] M. Stone and E. P. Davis, "A head and transducer support system for making ultrasound images of tongue/jaw movement," *The Journal of The Acoustical Society of America*, vol. 98, no. 6, pp. 3107–3112, 1995.

[8] D. H. Whalen, K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-LeHouillier, E. Vatikiotis-Bateson, and D. S. Hailey, "The haskins optically corrected ultrasound system (hocus)," *Journal of Speech, Language, and Hearing Research*, 2005.

[9] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time mri at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[10] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[11] A. Toutios and S. S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, 2016.

[12] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time mri," *Computer Speech & Language*, vol. 52, pp. 1–22, 2018.

[13] C. Carignan, R. K. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, "A real-time mri investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of french," *Journal of phonetics*, vol. 50, pp. 34–51, 2015.

[14] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time mri and articulatory coordination in speech," *Comptes rendus biologies*, vol. 325, no. 4, pp. 547–556, 2002.

[15] M. Proctor, Y. Zhu, A. Lammert, A. Toutios, B. Sands, U. Hummel, and S. Narayanan, "Click consonant production in khoekhoe: A real-time mri study," in *Khoisan Languages and Linguistics. Proc. 5th Intl. Symposium*, 2014, pp. 337–366.

[16] Y. Lim, Y. Zhu, S. G. Lingala, D. Byrd, S. Narayanan, and K. S. Nayak, "3d dynamic mri of the vocal tract during natural speech," *Magnetic resonance in medicine*, vol. 81, no. 3, pp. 1511–1520, 2019.

[17] M. Ahmad, J. Dargaud, A. Morin, and F. Cotton, "Dynamic mri of larynx and vocal fold vibrations in normal phonation," *Journal of Voice*, vol. 23, no. 2, pp. 235–239, 2009.

[18] I. K. Douros, A. Tsukanova, K. Isaieva, P.-A. Vuissoz, and Y. Laprie, "Towards a method of dynamic vocal tract shapes generation by combining static 3d and dynamic 2d mri speech data," in *INTERSPEECH*, 2019.

[19] I. K. Douros, A. Kulkarni, Y. Xie, C. Dourou, J. Felblinger, K. Isaieva, P.-A. Vuissoz, and Y. Laprie, "Mri vocal tract sagittal slices estimation during speech production of cv," in *28th European Signal Processing Conference (EUSIPCO 2020)*, 2020.

[20] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time mri of speaking at a resolution of 33 ms: undersampled radial flash with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, 2013.

[21] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[22] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.