Towards a method of dynamic vocal tract shapes generation by combining static 3D and dynamic 2D MRI speech data

Ioannis K Douros, Anastasiia Tsukanova, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie

Towards a method of dynamic vocal tract shapes generation by combining static 3D and dynamic 2D MRI speech data

Ioannis K. Douros[1,2]*, Anastasiia Tsukanova*[1]*, 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

ioannis.douros@loria.fr, anastasiia.tsukanova@loria.fr, karyna.isaieva@univ-lorraine.fr, pa.vuissoz@chru-nancy.fr, yves.laprie@loria.fr

Abstract

We present an algorithm for augmenting the shape of the vocal tract using 3D static and 2D dynamic speech MRI data. While static 3D images have better resolution and provide spatial information, 2D dynamic images capture the transitions. The aim of this work is to combine strong points of these two types of data to obtain better image quality of 2D dynamic images and extend the 2D dynamic images to the 3D domain.

To produce a 3D dynamic consonant-vowel (CV) sequence, our algorithm takes as input the 2D CV transition and the static 3D targets for C and V. To obtain the enhanced sequence of images, the first step is to find a transformation between the 2D images and the mid-sagittal slice of the acoustically corresponding 3D image stack, and then find a transformation between neighbouring sagittal slices in the 3D static image stack. Combination of these transformations allows producing the final set of images. In the present study we first examined the transformation from the 3D mid-sagittal frame to the 2D video in order to improve image quality and then we examined the extension of the 2D video to the 3rd dimension with the aim to enrich spatial information.

Index Terms: image transformation, modality transformation, MRI data, speech resources enrichment, vocal tract

1. Introduction

Despite having a rich history of research, speech production and its different aspects still persist as an open question in science. Such factors as novel technologies, the access to computationally advanced equipment and more detailed, versatile and voluminous data have all made it possible to obtain a more nuanced understanding of the way we speak. One of the most revealing sources is articulatory data as a bridge between the biomechanics of the speech as a mechanical process and its linguistic aspects. However, acquiring such data has always been a plight, facing a trade-off between the richness, completeness and interpretability of the information, the technique's ability to capture speech dynamics and the comfort and safety for the subject (X-ray [1]; electromagnetic articulography [2, 3]; ultrasound [4, 5]). With this regard, magnetic resonance imaging (MRI) as well as real-time MRI (rtMRI) can be a valuable resource, since it provides a much more complete picture of the articulators than within its competitors, is non-invasive, does not temper with the natural speech production and poses no known health hazards for the subject.

At present, MRI can capture a position of the vocal tract that was held stable over the acquisition time (typically, a dozen of seconds). The three-dimensional space is represented as a number of images, each collapsing together the information of its respective slice. This way we can obtain a comprehensive picture of the vocal tract, but due to the extended acquisition time, this picture is frozen. Such data can be successfully used for modeling vocal tract geometry [6, 7] and for volumetric studies [8, 9], both of which can subsequently find their uses (articulatory models [10], area functions [11]) with an implementation of temporal modeling and control that have to stem from elsewhere other than the MRI sequences. There are attempts to incorporate the temporal influence — coarticulatory effects — into such static data [12, 13], but the evidence is that attaining and maintaining a given static position for a period of time can be an insurmountable challenge for the speaker, resulting in unrealistic images, especially for producing liquids [14] and imposing control over nasalization.

The protocol of rtMRI, on the contrary, selects only one slice — for speech production research, typically the mid-sagittal one — and captures the tissues within that slice in real time [15, 16, 17]. The speech observed with such a method is unrestricted and therefore highly natural, allowing for a deep understanding of the dynamics of the articulators. As studies such as [18] show, having access to the mid-sagittal slice can be sufficient for applications given an estimate for the absent third-dimension information. Such methods as area function estimation [19] are commonly applied.

The biggest advantage of rtMRI over regular MRI is, naturally, its acquisition rate, which is considered to be sufficient to analyze the rapid-paced speech movements [20, 21, 22]. It cannot be denied, though, that in the attempt to gain enough temporal coverage we lose a lot of image sharpness and clarity. If the slice is not thin enough, the intricate geometry of the articulators gets projected on a single plane (there are phonemes with quite a complex three-dimensional behavior, such as the lateral /l/); whenever the speaker moves too fast, no position will be manifested for long enough to be captured by the machine. Both of these points can result in ghost effects (for example, the presence of two outlines of the tongue tip, which is an especially rapid articulator), image blurring or other artifacts, subsequently affecting the analysis and rendering image segmentation especially difficult.

Hence the motivation for this paper: the need to combine the strong points of MRI and rtMRI. MRI has good image quality and volume information; rtMRI the temporal resolution. To attain such a goal, we need to learn how to overcome their flaws, that is, how to enhance rtMRI images with the knowledge we have from static, well controlled MRI acquisitions, in order to fix their artifacts, and how to augment the image that is restricted to two dimensions so that it becomes volumetric.

The objective is to address the two issues for consonant-vowel (*CV*) syllables, taking the 2D rtMRI *CV* sequences as well as the corresponding 3D MRI *C(V), V* captures.

2. Materials and Methods

2.1. Data acquisition

The acquisition was carried out in two parts: the 2D real-time MRI data (rtMRI) were recorded at Max Planck Institute in Göttingen, Germany, while 3D static data (3D MRI) was recorded at Nancy Hospital, France.

2.1.1. Subjects

The selected subjects are 2 adult male native speakers of French speaking French. Subject 1 (S_1) is male, 32 years old, 180 cm tall and 65 kg, while subject 2 (S_2) is male, 35 years old, 182 cm tall and 74 kg.

2.1.2. 2D data

Our rtMRI dataset was recorded on a Siemens Prisma-fit 3T scanner (Siemens, Erlangen, Germany). We used radial RF-spoiled FLASH sequence [23] with TR $= 2.02$ ms, TE $= 1.28$ ms, FOV $= 19.2 \times 19.2$ cm, flip angle $= 5$ degrees, and slice thickness is 8 mm. Pixel bandwidth is 1600 Hz/pixel. Image resolution is 136×136. The acquisition time varied from 34 sec to 90 sec. We followed the protocol described in [24]. Images were recorded at a frame rate of 55 frames per second with the algorithm presented in [23].

2.1.3. 3D data

The 3D MRI data was recorded at Nancy Central Regional University Hospital under the approved medical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053).

Subject S_1's data was recorded on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). We used 3D VIBE (TR $= 3.57$ ms, TE $= 1.43$, flip angle $= 9$ degrees) for the acquisition. Acceleration factor is iPAT $= 3$. Scan slice thickness is 1.2 mm, FOV $= 22 \times 20$ cm and pixel bandwidth is 445 Hz/pixel. Duration of the acquisition was 15 seconds and the image resolution is 256×232 with 120 slices. Audio was recorded simultaneously with the MRI acquisition.

2.1.4. Audio recordings

In every session, audio was recorded simultaneously with the MRI acquisition. Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner by using a FOMRI III optoacoustics fibre-optic microphone. The subject wears earplugs to be protected from the noise of the scanner, but is still able to communicate orally with the experimenters via an in-scanner intercom system. Since the sound is recorded at the same time with the MRI acquisition, there is additional noise in the audio signal. In order to de-noise it, we used the de-noising algorithm proposed in [25].

2.2. Phonetic alignment

The transcription of the continuous speech corpus was phonetized by eLite HTS [26], and those phonetic labels were force aligned with HTK [27] using Merlin as frontend [28]. We manually checked the alignment results and manually corrected them in case of errors.

2.3. Image transformation

Although rigid transformations are simpler and less costly computationally, they will not be able to catch the differences in anatomy and articulation between the speakers and between the sagittal frames because these differences are more complex than rotations and translations. In our approach, we used a non-rigid image transformation method, based on an adaptation of demon's algorithm for image registration [29] To find the transformation between the images the algorithm described in [30] was applied. It calculates the displacement field between two images which shows how much and in which direction every pixel of the images should move in order to match the two 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.

2.4. Denoising procedure

To denoise the images, first we applied standard background subtraction (the background intensity being 10% of the maximum). The resulting images had cloud-like noise, to eliminate which we used thresholding, cutting off all the pixels with values less than defined (10% of maximum intensity). Such manipulation does not lose any essential information in the vicinity of the vocal tract, while the level of noise reduces strongly, leaving just some speckles. These point-like outliers can be smoothed out with a median filter.

The only problem that remained then was to treat the artifacts that were mostly caused by the speaker's movement. These artifacts had fog-like appearance and did not impact strongly the sharpness of the edges. Therefore, the natural idea was to use edges as boundaries of application of a Gaussian filter and to smooth out the artifacts, keeping the edges sharp. We used Canny edge detection procedure, and then we connected edges close to each other in order to remove small holes with the help of the [31] toolbox. After this, we applied a Gaussian filter (size 5×5, $\sigma = 5$) so that the image pixel convolved with the kernel only in the case when, according to the breadth first search algorithm, there was a way to go from this pixel to the central point of the filter without crossing a single edge.

3. Experiments

In our study we focused on 12 CV syllables (/fi/, /fa/, /fu/, /pi/, /pa/, /pu/, /si/, /sa/, /su/, /ti/, /ta/, /tu/) selected from bigger words occurring in non-spontaneous sentences. Our aim was to create a semi-automated procedure that would be able to transform a 2D video to 3D. We used the two speakers (S_1 and S_2) for our experiments and we examined the case of using static 3D and dynamic 2D data of a) same speaker: S_1 3D and S_1 2D; b) different speakers: S_1 3D and S_2 2D. This means that we used 3D static images of S_1 to 3D-transform 2D videos of both S_1 and S_2. The experiments were carried out in Matlab. Our algorithm is comprises the following steps: image pre-processing, C and V 2D mapping, space and time extension, image combination and denoising.

3.1. Images pre-processing

We select the images corresponding to the studied syllables by using the phonetic alignment information. One of the problems that we had to face was the fact that the images were acquired using different MRI sequences and even different MRI machines. This resulted in images with different contrast levels, different resolutions and different head position within the image.

The first thing that we did was upsampling the dynamic images in order to match the resolution of the static ones. This is necessary for computational purposes since the algorithm works

at a pixel level. We increased the images 0.55 times using cubic interpolation.

Since our algorithm is mainly designed for use in speech studies, our main priority was to create a realistic vocal tract (VT) shape in the transformed images and therefore less care was given to capturing details that do not affect the VT shape, like the internal anatomical details of the brain or tongue. For this reason we used a window to keep only the VT part of the head. The window was manually designed for one of the images and then applied to the rest of the images.

In total, 3 initial windows were designed, one for the static images and two for the dynamic (one per speaker). We visually checked that the VT stays within the window in every image. We tried to keep the minimum window possible, especially close to the palate because in the dynamic images it is sometimes unclear where the palate ends due to blurring; therefore we tried to cut the palate out by the window placement. However, due to small movements of the head during the acquisition it was not possible to cut it in every case. Another thing to consider while placing the windows is that since we had cross-speaker experiments, the window had to be sometimes larger than the VT so that its dimension would be enough to fit the VT of the other speaker with a different anatomy. In the end, a window with dimensions 146×131 was selected.

3.2. Double 2D mapping

We automatically selected the mid-sagittal frame of the 3D images both for the C and for the V part of the studied CV syllable. We then transformed the previously selected static images to the first and last frame of the given dynamic images respectively. 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 every case, both for histogram matching and for image transformation, the dynamic image served as a reference. The output at this step is two images, one for the C (I_c) and one for the V (I_v), that have the articulation of the corresponding dynamic image but with improved contrast and more visible details. We also save the deformation field T_n and T_i respectively.

3.3. Space and time extension

This stage can be further divided into three steps: deformation in time, deformation in space and combination of the two.

The first step is to use the static images in order to find the deformation field that would transform the sagittal slice i to the sagittal slice $i + 1$. In our case the starting frame was the mid-sagittal and we calculated the deformation fields both going to the left and right; therefore, i is negative for the first half of the slices, since we go from slice i to slice $i - 1$ for this half. We call this transformation $T_s(i)$.

In the second step we extract the deformation field that would transform frame i to the frame $i + 1$ using the dynamic sequence. Let's call this transformation $T_{d-n}(i)$. We do the same calculations for the inverse sequence (from frame $i + 1$ to i, the reason will be explained later on) and we call it $T_{d-i}(i)$.

The last step is to use the previously computed transformations to produce the first version of the complete 3D dynamic images. To do so, we first have to synthesize the sagittal frames of the transformed image I_c. This can be done by transforming the transformation between the static and the dynamic data (T_n) using the transformation for sagittal transformation (T_s). The resulting transformation ($T_{s-r}(i) = T_s(T_n(i))$) is then applied to I_c. Note that to obtain the sagittal slices, we first apply the transformation to I_c to synthesize, let us say, its left neighbor frame, and then this left frame is used to synthesize the next left frame, and so on. The same procedure applies to the right frames as well.

At this point we have the synthetic 3D image corresponding to the first frame of the dynamic images (the C part). We now apply the $T_{d-i}(i)$ transformation to all the 3D synthetic slices in order to get a 3D dynamic video.

We apply exactly the same procedure using the inverse transformations and as a starting reference point the I_v image.

What we eventually have at the end of this step is two first versions of the 3D dynamic image transformation, one starting from the C and propagating forward to V (S_f) and one start from V and propagating backwards to C (S_b).

3.4. Image combination

Since we have two versions of 3D dynamic images (S_f, S_b), we need to combine them. The approach that we chose to follow was to keep the images from the backwards transformation S_b which was created based on the static data of the vowel, for the whole duration of the vowel based on the phonetic annotations from the audio file. For the images that correspond to the consonant we used the images created by the backwards transformation S_b in combination with the images created from the forward transformation S_f. In order to achieve similarity between the images, we transformed the images corresponding to consonant from S_f to the corresponding images of S_b to obtain the S_{f-t}. The reference images both for image transformation and for histogram matching was selected to be the images from S_b since it is the vowel what is the syllable nucleus. This way, relying on S_b increases robustness. Finally, we cropped the left part of the transformed images S_{f-t} for the duration of the consonant to cover the place of articulation [32] (the lips for /p/ and /f/, the alveolar ridge for /t/ and /s/) and we paste it to the rest of the part from the S_b. We applied this to all synthesized sagittal frames.

The output of this stage S_{comb} is a 3D dynamic video; the V part of it corresponds to the output of S_b, and the C part is a combination: the left side of the image is the output of S_{f-t}, the right of S_b.

3.5. Denoising

The denoising procedure that we used was same as described in Section 2. The reasons why some image processing techniques were necessary for our combined set of images S_{comb} is because of: artifacts in the MR images; motion blurring; noise induced by MRI; noise created due to image transformation from one modality to another (midsaggital 3D to 2D), which then propagates and gets worse with the subsequent image transformations. Finally, despite histogram matching and image transformation, sometimes there are some differences at the gluing points, mainly regarding contrast. Since our main purpose is to produce the VT shape, we created two versions of the final images: one that mainly focuses on improving the general quality of the image and a second one which aims at eliminating information not directly related to the shape of the VT-like texture, in order to facilitate the use of these images in experiments around the VT shape, like segmentation. Examples of denoised images can be seen in Figure 1.

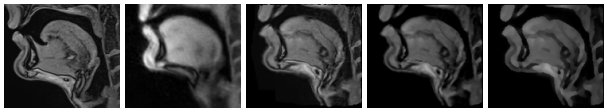

Figure 1: *From left to right: /f(a)/ by S_1 in 3D; 2D; the raw generated midsagittal sequence when transforming to S_1; its denoising; its denoising and smoothing*

3.6. Evaluation

Another issue that we faced was how to evaluate the synthesized 3D dynamic images, especially the non-midsagittal frames since there is no reference for them. Therefore, we asked two linguists to give us their qualitative opinion of how well the synthesized midsagittal sequences represented the produced sound and how closely the synthesized midsagittal images matched the reference dynamic ones. Their opinion was that in the majority of the situations the transformations described the original 2D data adequately well, both in the same- and cross-speaker cases (Figure 2).

For the non-midsagittal slices, we visually inspect them to check if we have smooth transitions based on the what we would expect from the static 3D MRI. Although there were some problematic cases, either to do with an insufficiently good image combination or due to filtering, in most of the cases the non-midsagittal slices looked more or less as we would expect (Figure 3).

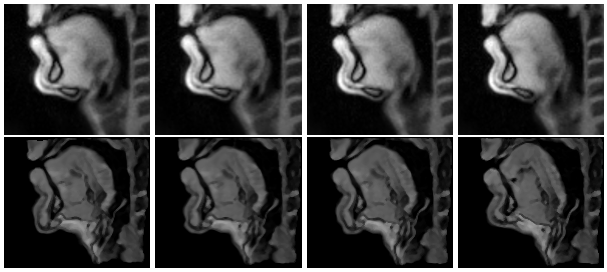

Figure 2: *Every 3rd image in the original dynamic 2D articulation of /si/ by S_2 (above, left to right) and the generated midsagittal slice sequence when transforming to the 3D data of S_1 (below, left to right)*

4. Conclusions

One of the challenges that we had to face was the fact that static articulation is different from natural speech. This was especially obvious for the consonants, since the position of the articulators is highly dependent on the anticipated vowel. This problem persisted even despite the coarticulation-aware protocol of the 3D MRI acquisitions (Section 2.1). Especially in the plosives /t/ and /p/, just maintaining a stable articulation is a challenge for the speaker. However, in the case of the vowels, the situation is better since they are more similar to the corresponding dynamic phonemes. This is a major point to consider, given that the proposed algorithm is based on the initial 3D midsagittal slice to the 2D mapping. An aspect that the core transformations S_b, S_f struggled at was to capture the articulators' contact (the lip closure, the tongue touching the palate) due to the calculations of the deformation field, even though the constriction narrows in the C part. That is why we extract this type of information from S_f which capture well the closures (and struggle to open them) but as we approach V, in many cases other parts of the VT manifest dismorphing because of the critical differences between the static and the dynamic C captures. According to the evaluators, this is especially the case of /tu/: they approved /t/ and /u/ parts, but noticed a jerk between them.However, in case of /sV/ or /fV/, S_b managed to capture all the midsagittal transitions well on its own. There was no significant difference between the S_b and the S_{comb} versions.

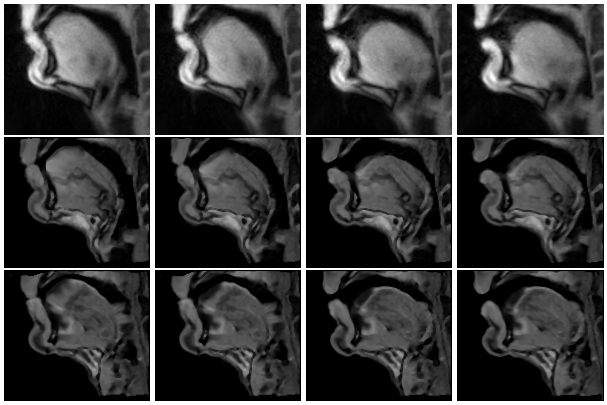

Figure 3: *Every 2nd image in the original dynamic 2D articulation of /pa/ by S_1 (above, left to right), the generated midsagittal slice sequence when transforming to the 3D data of S_1 (middle) and sagittal sequence (slice 67 out of 120; below)*

An important remark is that we used data from different machines, acquired with different sequences resulting in a very different image quality. Moreover, we dealt with two types of articulation (static with dynamic) and even used different speakers with anatomical differences , the resulting images were adequately robust. Their behaviour in same-speaker and cross-speaker 3D dynamic transformation was consistent.

Finally, we can see that the synthesized image quality improves for the midsagittal slices as they have increased contrast and resolution while preserving the vocal tract shape from the 2D images and in the majority of the cases the anatomical information as well. For the non-midsagittal frames, we can still synthesize reasonable images, but with more noise/artifacts, and in some cases problems at the gluing points.

A future direction would be to study more syllables and eventually create a fully automated method to directly apply it to 2D dynamic MRI databases (existing or new) for data enrichment. Finally one can think of using additional information of other modalities in order to further improve the transformation results.

5. Acknowledgement

This work was financed by Lorraine Université d'Excellence (LUE) grant and ANR ArtSpeech. We would like to thank Anastasia Shimorina and Chrysanthi Dourou for their help and comments.

6. References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] J. Dang, "Estimation of vocal tract shape from speech sounds via a physiological articulatory model," in *In Proceedings of 5th Seminar on Speech Production: Models and Data*. Citeseer, 2000.

[7] O. Engwall, "Tongue talking: studies in intraoral speech synthesis," Ph.D. dissertation, KTH, 2002.

[8] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter, 1970, no. 2.

[9] Z. I. Skordilis, A. Toutios, J. Töger, and S. Narayanan, "Estimation of vocal tract area function from volumetric magnetic resonance imaging," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 924–928.

[10] P. Birkholz and D. Jackel, "A three-dimensional model of the vocal tract for speech synthesis," in *Proceedings of the 15th international congress of phonetic sciences*. Barcelona, Spain, 2003, pp. 2597–2600.

[11] B. H. Story, "Phrase-level speech simulation with an airway modulation model of speech production," *Computer speech & language*, vol. 27, no. 4, pp. 989–1010, 2013.

[12] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[13] A. Tsukanova, B. Elie, and Y. Laprie, "Articulatory speech synthesis from static context-aware articulatory targets," in *International Seminar on Speech Production*. Springer, 2017, pp. 37–47.

[14] 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.

[15] 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.

[16] M. Ruthven, A. C. Freitas, R. Boubertakh, and M. E. Miquel, "Application of radial grappa techniques to single-and multislice dynamic speech mri using a 16-channel neurovascular coil," *Magnetic resonance in medicine*, vol. 82, no. 3, pp. 948–958, 2019.

[17] M. Fu, M. S. Barlaz, J. L. Holtrop, J. L. Perry, D. P. Kuehn, R. K. Shosted, Z.-P. Liang, and B. P. Sutton, "High-frame-rate full-vocal-tract 3d dynamic speech imaging," *Magnetic resonance in medicine*, vol. 77, no. 4, pp. 1619–1629, 2017.

[18] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[19] S. Maeda and Y. Laprie, "Vowel and prosodic factor dependent variations of vocal-tract length," in *InterSpeech-14th Annual Conference of the International Speech Communication Association-2013*, 2013.

[20] 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.

[21] 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.

[22] 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*, 2018.

[23] 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.

[24] 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.

[25] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[26] S. Roekhaut, S. Brognaux, R. Beaufort, and T. Dutoit, "eLite-HTS: Un outil TAL pour la génération de synthèse hmm en français," in *Démonstration aux Journées d'étude de la parole (JEP)*, 2014.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[28] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[29] 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.

[30] 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.

[31] [Online]. Available: https://www.peterkovesi.com/matlabfns/index.html#edgelink

[32] Z. Fagyal, D. Kibbee, and F. Jenkins, *French: A linguistic introduction*. Cambridge University Press, 2006.