Towards a 3 dimensional dynamic generic speaker model to study geometry simplifications of the vocal tract using magnetic resonance imaging data

Ioannis K Douros

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Towards a 3 dimensional dynamic generic speaker model to study geometry simplifications of the vocal tract using magnetic resonance imaging data

THÈSE

présentée et soutenue publiquement le 02 September 2020

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Ioannis K. Douros

Composition du jury

Président :	Salvatore-Antoine Tabbone	Université de Lorraine, LORIA
Rapporteurs :	Michel Desvignes	Université Grenoble Alpes, GIPSA-lab
	Didier Demolin	Sorbonne Nouvelle (Paris III), LPP
Examinateurs :	Dan Dediu	Université Lumière Lyon 2, DDL
	Corinne Fredouille	Avignon Université, LIA-CERI
Encadrants :	Yves Laprie	Université de Lorraine, CNRS-LORIA
	Pierre-André Vuissoz	Université de Lorraine, IADI

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Résumé

Dans cette thèse, nous avons utilisé les données de l'IRM du conduit vocal pour étudier la production de la parole. La première partie consiste en l'étude de l'impact que le vélum, l'épiglotte et la position de la tête a sur la phonation de cinq voyelles françaises. Des simulations acoustiques ont été utilisées pour comparer les formants des cas étudiés avec la référence afin de mesurer leur impact. Pour cette partie du travail, nous avons utilisé des IRM statiques en 3D. Comme la parole est généralement une phénomène dynamique une question s'est posée, à savoir s'il serait possible de traiter les données 3D afin d'incorporer des informations temporelles de la parole continue. Par conséquent, la deuxième partie présente quelques algorithmes que l'on peut utiliser pour améliorer les données de production de la parole. Plusieurs transformations d'images ont été combinées afin de générer des estimations des formes du conduit vocal qui sont plus informatives que les originales. À ce stade, nous avons envisagé, outre l'amélioration des données de production de la parole, de créer un modèle de référence générique qui pourrait fournir des informations améliorées non pas pour un sujet spécifique, mais globalement pour la parole. C'est pourquoi nous avons consacré la troisième partie l'étude d'un algorithme permettant de créer un atlas spatio-temporel de l'appareil vocal qui peut être utilisé comme référence ou standard pour l'étude de la parole car il est indépendant du locuteur. Enfin, la dernière partie de la thèse, fait référence à une sélection de questions ouvertes du domaine qui restent encore sans réponse, quelques pistes intéressantes que l'on peut développer à partir de cette thèse et quelques approches potentielles qui pourraient être envisager afin de répondre à ces questions.

Mots-clés: IRM, production de la parole, conduit vocal, simulation acoustique, transformation des images, amélioration des données articulatoires, atlas spatio-temporel

Abstract

In this thesis we used MRI (Magnetic Resonance Imaging) data of the vocal tract to study speech production. The first part consist of the study of the impact that the velum, the epiglottis and the head position has on the phonation of five french vowels. Acoustic simulations were used to compare the formants of the studied cases with the reference in order to measure their impact. For this part of the work, we used 3D static MR (Magnetic Resonance) images. As speech is usually a dynamic phenomenon, a question arose, whether it would be possible to process the 3D data in order to incorporate dynamic information of continuous speech. Therefore the second part presents some algorithms that one can use in order to enhance speech production data. Several image transformations were combined in order to generate estimations of vocal tract shapes which are more informative than the original ones. At this point, we envisaged apart from enhancing speech production data, to create a generic speaker model that could provide enhanced information not

for a specific subject, but globally for speech. As a result, we devoted the third part in the investigation of an algorithm that one can use to create a spatio-temporal atlas of the vocal tract which can be used as a reference or standard speaker for speech studies as it is speaker independent. Finally, the last part of the thesis, refers to a selection of open questions of the field that are still left unanswered, some interesting directions that one can expand this thesis and some potential approaches that could help someone move forward towards these directions.

Keywords: MRI, speech production, vocal tract, acoustic simulation, image transformation, articulatory data enhancement, spatio-temporal atlas

Περίληψη

Σε αυτή τη διατριβή χρησιμοποιήσαμε δεδομένα μαγνητικής τομογραφίας της φωνητικής οδού για να μελετήσουμε την παραγωγή ομιλίας. Το πρώτο μέρος αποτελείται από τη μελέτη της επίδρασης της σταφυλής, της επιγλωττίδας και της θέσης του κεφαλιού στη φώνηση πέντε γαλλικών φωνηέντων. Ακουστικές προσομοιώσεις χρησιμοποιήθηκαν για σύγκριση των φωνοσυντονισμών μεταξύ των περιπτώσεων που μελετήθηκαν αυτών της αναφοράς προκειμένου να μετρηθεί ο αντίκτυπός τους. Για αυτό το μέρος της εργασίας, χρησιμοποιήσαμε τρισδιάστατες στατικές μαγνητικές τομογραφιές. Καθώς η ομιλία είναι συνήθως ένα δυναμικό φαινόμενο, ένα απ΄ τα ερωτήματα που προέκυψε είναι εάν θα ήταν δυνατή η επεξεργασία των τρισδιάστατων δεδομένων ώστε να ενσωματωθούν πληροφορίες συνεχούς ομιλίας. Ως εκ τούτου το δεύτερο μέρος παρουσιάζει μερικούς αλγόριθμους που μπορεί κανείς να χρησιμοποιήσει για να εμπλουτίσει τα δεδομένα παραγωγής ομιλίας. Διάφοροι μετασχηματισμοί εικόνας συνδυάστηκαν για να δημιουργηθούν εκτιμήσεις σχημάτων της φωνητικής οδού που να είναι πιο ενημερωτικά από τα αρχικά όσον αφορά την πληροφορία που εμπεριέχουν σχτικά με τα φαινόμενα της παραγογής ομίας. Σε αυτό το σημείο, οραματιστήκαμε, εκτός από την εμπλούτιση των δεδομένων παραγωγής ομιλίας, να δημιουργήσουμε ένα γενικό μοντέλο ομιλητή που θα μπορούσε να παρέχει βελτιωμένες πληροφορίες όχι για ένα συγκεκριμένο υποκείμενο, αλλά γενικότερα για κάθε ομιλητή. Ως αποτέλεσμα, αφιερώσαμε το τρίτο μέρος στην ανάπτυξη ενός αλγορίθμου για τη δημιουργία ενός τρισδιάστατου δυναμικού (χωροχρονικού) άτλαντα της φωνητικής οδού που μπορεί να χρησιμοποιηθεί ως αναφορά ή αντιπροσωπευτικός ομιλητής για μελέτες ομιλίας καθώς θα είναι ανεξάρτητος από κάποιο συγκεκριμένο υποκείμενο. Τέλος, στο τελευταίο μέρος της διατριβής, παρουσιάζεται μια επιλογή από ανοιχτές ερωτήσεις στο πεδίο της ομιλίας οι οποίες παραμένουν ακόμα αναπάντητες, μερικές ενδιαφέρουσες κατευθύνσεις που θα μπορούσε κανείς να επεκτείνει αυτήν τη διατριβή καθώς και μερικές πιθανές προσεγγίσεις που θα μπορούσαν να βοηθήσουν κάποιον να προχωρήσει προς αυτές τις κατευθύνσεις.

Λέξεις-κλειδιά: μαγνητικές τομογραφίες, παραγωγή ομιλίας, φωνητική οδός, ακουστικές προσομοίωσεις, μετατροπές εικόνων, εμπλουτισμός δεδομένων άρθρωσης φωνής, χωροχρονικός άτλαντας

Acknowledgments

PhD is a personal dream, a destination that you are trying to reach using as a boat your hard work and as a compass your vision for a better world. This journey became possible thanks to some people who guided me, helped me and supported me during all this time. First of all I would like to thank my supervisors, Yves Laprie and Pierre-André Vuissoz for believing in me and choosing me for this PhD position. The guidance and the freedom to explore that they gave me were key elements for this work. Moreover I would like to thank the team directors, Denis Jouvet from Multispeech team at INRIA-LORIA lab and Jacques Felblinger from IADI team at CHU of Nancy for the resources that they provided. Also my colleagues Marine, Gabriela, Yu, Emmanuel, Nicolas, Adrien, Anastasiia and Karyna for their useful discussions.

A big thank you to my martial arts trainer Kostas G. for passing on me the "fighting spirit" that kept me focused on my target during the hard times, to my friends in France Ajinkya, Sandipana, Theo, Lou, Denis, Dominique, Mao, Yuan, Pan, Meng and Ivana, to my friends in Greece Nance, Babis, Vasilis, Mahi, Dimitris K., Dimitris M., Giorgos, Makis, Valeria, Antonis, Sotiris, Ioannis, Fotini P. and Iakovos as well as to my friend Lilly from China for the joyful and funny moments that we shared together and their support and help during my difficult times. Without them everything would have been much harder. I also want to thank every "random" person that I met in France, Greece and every other country that I visited, that made my days happier and easier with his/her help, advice, wish or with just a simple smile. A special thank you to Fotini X.Y. for the beautiful moments, her useful lessons and for being by my side throughout my PhD.

My PhD journey wouldn't have been possible if I hadn't met two people: Nasos Katsamanis and Vali Despotopoulou. Whatever I have achieved and I may achieve in the future will be owed to them. I thank them from my heart for the vision, inspiration, knowledge and the way of thinking that they have passed on to me. Additionally, I would like to say a very big thank you to my French family, Nicole and Serge, that I had the chance to meet and stay with them from the first day of my PhD until the end of it. The hospitality, care and energy that they spend on all my issues is unimaginable. They were always there for me, supporting me to all my difficult times as well as to all my "strange" ideas and plans. The biggest luck of my entire PhD is having these two people by my side. Finally, I would like to say a very big thank you to my parents Kostas and Areti and my sister Chrysa for supporting all my dreams and giving their everything so that I can reach as close to the sky as I could. Words are not enough to describe how grateful I am to them.

Contents

List of Figures

List of Tables

List of Abbreviations

C ..... Consonant

CT ..... Computed Tomography

CV ..... Consonant Vowel

DTW ..... Dynamic Time Warping

EBCT ..... Electron Beam Computed Tomography

EMA ..... ElectroMagnetic Articulography

EMD ..... Earth Mover's Distance

EPG ..... ElectroPalatoGraphy

FFT ..... Fast Fourier Transform

FOV ..... Field Of View

fps ..... frames per second

GRU ..... Gated Recurrent Units

HMM ..... Hidden Markov Model

HNM ..... Harmonic Noise Model

LPC ..... Linear Prediction Coding

LSTM ..... Long Short Term Memory

MFCC ..... Mel Frequency Cepstral Coefficient

MR ..... Magnetic Resonance

MRI ..... Magnetic Resonance Imaging

PCA ..... Principal Component Analysis

PML ..... Perfect Match Layer

PSOLA ..... Pitch Synchronous OverLapping Add

RF ..... Radio Frequency

RNN ..... Recurrent Neural Network

rtMRI ..... real time Magnetic Resonance Imaging

SIFT ..... Scale Invariant Feature Transform

SSIM ..... Structural SIMilarity

TE ..... Temps d'Echo

TLCA ..... Transmission Line Circuit Analog

TR ..... Temps de Répétition

V ..... Vowel

VCV ..... Vowel Consonant Vowel

VT ..... Vocal Tract

VV ..... Vowel Vowel

Chapter 1

Introduction

Speech production is a very broad field that studies how speech is created, starting from how thoughts are translated into concepts (conceptualization), how concepts into words and words are assembled into phrases (formulation), how the various articulators and the various muscles are coordinated together in order to produce speech (articulation) and finally checking if the produced output is correct (self-monitoring) [Lev92, SNA13, KS02].

Since speech is one of the main "tools" that people use every day in order to communicate and express their thoughts and ideas [Har14], it is very important to study it. In Sections 1.1-1.3, we are going to present the use of speech production in people's lives, the purpose and the motivation behind this work and describe the background knowledge required for this thesis.

1.1 Applications of speech production knowledge

The first interest is probably medical applications. People with speech pathologies that need some kind of operation to remove part of the vocal tract or to add an artificial implant can greatly benefit from articulatory modeling. It can guide the surgeon on how to perform the operation and where and how to "cut" the section to be removed so that the speaker will keep the biggest part possible of his ability to speak. In case of implants, they can be tuned and adjusted offline for the specific speaker in an optimal way.

Speech synthesis and voice conversion is another field that could potentially profit from advances in speech production research [FEG+08]. People with laryngectomy for example can have a realistic sound artificial voice thanks to speech synthesis techniques [WTK+20]. Speech production knowledge can further improve the results of synthetic voice at various stages like using articulatory information of the target speaker to give him back his original voice in an artificial way, or by guiding the modeling of the articulators by providing information about their importance given a pronounced sound [FKJ06] for instance. Speech synthesis is also playing an important role in the future world with smart houses and devices that communicate with the user through speech.

Speech production can enhance speech recognition systems by providing information about the nature of speech during the training phase [KFL+07, RSS96, FK01]. This is especially useful in situations with extreme noise conditions as in most real life applications

like giving commands to the phone during driving or to smart devices at home. Another example is during piloting a fighting aircraft where the noise is very big, especially in helicopters, and high accuracy of speech recognition is necessary for efficient communication and fatigue detection [GFW+06].

Knowledge of speech production can also help people who are trying to learn a new language by providing precise information about how the phonemes are supposed to be articulated and how one should place the articulators in order to achieve a good oral production. Speech therapists can also benefit from such knowledge by being able to visualise and better guide people who stutter so that they can improve their communication faster [KGBL08].

Here we presented the main fields that could benefit from speech production knowledge. Usually speech production should be combined with knowledge from other fields like speech synthesis or speech recognition in order to fully exploit its potential. Having presented its main applications we are going to describe the purpose and the motivation for this thesis in Section 1.2.

1.2 Purpose and motivation

This thesis was conducted under a collaboration between the Multispeech team at LORIA laboratory and the IADI team at the Hospital of Nancy. They provided access to high speed computers and advanced MRI facilities that facilitated this study. Some images from the MRI facilities and equipment can be seen in Fig. 1.1.

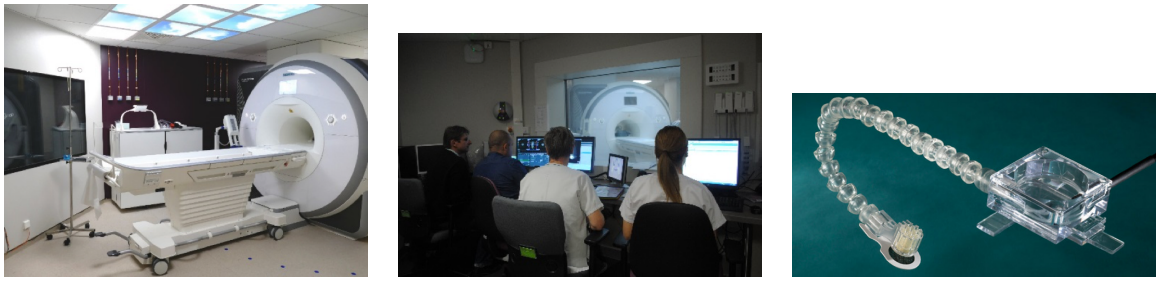


Figure 1.1: Examples of the facilities provided by the hospital. From left to right: the MRI scanner, the control room and the optical microphone

The main purpose of this PhD subject is to study speech production using MRI images. More specifically, the targets of this thesis can be summarized in the following:

- Use static MRI data of the vocal tract and acoustic simulations to explore the impact of speech articulators, geometric simplifications and head positions on speech phonation.
- Work on algorithms for spatial and temporal vocal tract's shape extension, i.e. transform the dynamic 2D mid-sagittal MRI images of the vocal tract into dynamic 3D MRI images of the vocal tract or estimate the temporal evolution of the vocal tract shape during CV (Consonant Vowel) production by using silence frames.

- Work on an algorithm to create a 3D dynamic atlas of the vocal tract of CVs that will serve as a generic speaker model to describe the speech dynamics in spatial and temporal domains.

During the investigation of these topics, the following publications were made, where this thesis played a major role and directly contributed to them:

1. **Ioannis K. Douros**, Pierre-André Vuissoz, Yves Laprie, "Comparison between 2D and 3D models for speech production: a study of French vowels", International Congress on Phonetic Sciences (ICPhS), August, Melbourne, Australia, 2019

   (Presented in Chapter 3)

2. **Ioannis K. Douros**, Pierre-André Vuissoz, Yves Laprie, "Acoustic impact of geometric approximation at the level of velum and epiglottis on French vowels", International Congress on Phonetic Sciences (ICPhS), August, Melbourne, Australia, 2019

   (Presented in Chapter 3)

3. **Ioannis K. Douros**, Pierre-André Vuissoz, Yves Laprie, "Effect of head posture on phonation of French vowels", International Congress on Phonetic Sciences (ICPhS), August, Melbourne, Australia, 2019

   (Presented in Chapter 3)

4. **Ioannis K. Douros**,Yves Laprie, Pierre-André Vuissoz, Benjamin Ellie, "Acoustic Evaluation of Simplifying Hypotheses Used in Articulatory Synthesis", International Congress on Acoustics (ICA), Aachen, Germany 2019

   (Presented in Chapter 3)

5. **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", INTERSPEECH, Graz, Austria, September 2019

   (Presented in Chapter 4)

6. **Ioannis K. Douros**, Chrysanthi Dourou, Yu Xie, Jacques Felblinger, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie, "Synthesize MRI vocal tract data during CV production", International Seminar on Speech Production (ISSP), December, Providence, USA, 2020

   (Presented in Chapter 4)

7. **Ioannis K. Douros**, Yu Xie, Chrysanthi Dourou, Jacques Felblinger, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie, "Vocal tract sagittal slices estimation from MRI midsagittal slices during speech production of CV", International Seminar on Speech Production (ISSP), December, Providence, USA, 2020

   (Presented in Chapter 4)

8. **Ioannis K. Douros**, Ajinkya Kulkarni, Yu Xie, Chrysanthi Dourou, Jacques Felblinger, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie, "MRI vocal tract sagittal slices estimation during speech production of CV", European Signal Processing Conference (EUSIPCO), January, Amsterdam, Netherlands, 2021

   (Presented in Chapter 4)

9. **Ioannis K. Douros**, Ajinkya Kulkarni, Chrysanthi Dourou, Yu Xie, Jacques Felblinger, Karyna Isaieva, Pierre-André Vuissoz, Yves Laprie, "Using Silence to Synthesise Dynamic MRI Vocal Tract Data of CV", submitted to INTERSPEECH, October, Shanghai , China, 2020

   (Presented in Chapter 4)

10. **Ioannis K. Douros**, Yu Xie, Chrysanthi Dourou, Karyna Isaieva, Pierre-André Vuissoz, Jacques Felblinger, Yves Laprie, "A 3D dynamic spatio-temporal atlas of the vocal tract during consonant-vowel production from 2D real time MRI", submitted to Journal of the Acoustical Society of America

    (Presented in Chapter 5)

Furthermore, a few more papers were published during this PhD period where this thesis was not the major contribution. These are the following:

1. Anastasiia Tsukanova, **Ioannis K. Douros**, Anastasia Shimorina, Yves Laprie, "Can static vocal tract positions represent articulatory targets in continuous speech? Matching static MRI captures against real-timeMRI for the French language", International Congress on Phonetic Sciences (ICPhS), August, Melbourne, Australia, 2019

2. **Ioannis K. Douros**, Jacques Felblinger, Jens Frahm, Karyna Isaieva, Arun A. Joseph, Yves Laprie, Freddy Odille, Anastasiia Tsukanova, Dirk Voit, Pierre-André Vuissoz, "A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research", Interspeech, Graz, Austria, September 2019

3. Anastasiia Tsukanova, **Ioannis K. Douros**, Yves Laprie, "DNN-based parametric speech synthesis enhanced with articulatory information", International Seminar on Speech Production (ISSP), December, Providence, USA, 2020

4. Ajinkya Kulkarni, **Ioannis K. Douros**, Vincent Colotte, Denis Jouvet, "Emotion recognition from phoneme-duration information", International Seminar on Speech Production (ISSP), December, Providence, USA, 2020

5. Karyna Isaieva, Yves Laprie, Freddy Odille, **Ioannis K. Douros**, Jacques Felblinger, Pierre-André Vuissoz, "Measurement of tongue tip velocity from real-time MRI and phase-contrast cine-MRI in consonant production", Journal of Imaging

6. Yu Xie, Julien Oster, Emilien Micard, Bailiang Chen, **Ioannis K. Douros**, Liang Liao, François Zhu, Marc Soudant, Jacques Felblinger, Francis Guillemin, Gabriela Hossu, Serge Bracard, "Impact of Pretreatment Ischemic Location on Functional Outcome After Thrombectomy", submitted to European Journal of Neurology

A long term goal following this thesis would be the application of acoustic simulations to dynamic sounds like CV, VCV (Vowel Consonant Vowel) or the extensions of all the work done with CVs to whole words, towards the direction of creating a global, realistic and speaker independent dynamic articulatory model that would be fully parameterised with direct control of speech articulators and anatomical characteristics.

Our work towards this direction was mainly motivated by the following factors. First there are very few studies that explore the effect of smaller articulators or the head position on phonation and we believe that articulatory synthesis should be able to take into account all these aspects as well. Secondly, in order to create a fully controlable model we need to understand how vocal tract shapes transform over time and space during speech production. Even though there are studies on this field, there is still room for improvement (see Section 1.3 for more details). Another motivation is the technological limitations regarding speech imaging that prevent us from acquiring the complete articulatory information as we would like. Moreover, the potential of using the results and the methods explored in this study to the fields of phonetics, speech production pathologies, vocal tract surgery, image segmentation of vocal tract further inspired us. Finally, another motivation is the application of our results to the field of articulatory speech synthesis. Since this is an effective approach to understand speech production and vice versa we hope that it will bring us one step closer to the final goal of a complete articulatory synthesis system.

1.3 Global overview

In this Section we are going to give an overview of the main advancements in the fields related to this study. We start by presenting several approaches to acquire articulatory information, we continue with a description of popular speech synthesis techniques and finally we demonstrate some articulatory data synthesis and atlas creation approaches.

1.3.1 Techniques to capture articulatory information

Ultrasound is an imaging technique based on the transmission of high frequency sound pulses through tissues. When the wave moves from one type of tissue to another, there is a reflection back to the wave sensor/source, which is used to construct the images. Even though ultrasound imaging is easy and cheap, the image quality is quite noisy and only a part of the tongue contour is visible which creates problems.

In the past, people used X-rays films to acquire videos of the dynamics of the whole vocal tract. Such approaches allowed the acquisition of videos with frame rate up to 100 fps. However, X-ray imaging uses an ionizing type of radiations which is hazardous for the subject if he/she is exposed for a long period of time and this raises ethical issues which led to a stop to those acquisitions. Additionally, since X-ray is a projection of the whole vocal tract onto a 2D plane, there is a considerable overlap between the tissue which makes it hard to distinguish all the articulators. Moreover, potential fillings can pose a big difficulty by hiding the tongue contour. Nowadays this technique is not allowed anymore due to the health risks.

A similar approach based on X-ray imaging, is X-ray microbeam [WMWK90]. The advantage is that it greatly reduces the amount of radiation that a subject absorbs. This is achieved by using a pinhole which allows only a small focused line of X-ray to pass through. Pellets are places on the surface of the articulators like the tongue or the lips and these pellets are then tracked. The reduction of radiation comes at the cost of precision since only the pellets are tracked without imaging the whole vocal tract. However, even with the lower precision it is possible to approximate the contours and study speech production [JLL93]. Like X-ray, this technique is also not allowed anymore due to health risks.

Another technique that utilises the concept of tracking sensors is EMA (ElectroMagnetic Articulography) [Hix71, HVG+96]. The idea of EMA is that a number of sensors (usually 24 in the most recent machines) are glued on the articulators inside and outside of the subject's vocal tract. Sensors are connected with cables directly to the EMA acquisition device (in case of the sensors inside the vocal tract, a cable enters the mouth for every sensor placed inside). Around the head of the subject, three variable low intensity magnetic fields are generated and calibrated. Additional sensors glued into the nose and ears allow the compensation of the head movement. When subject speaks, the movements of the articulators creates the movement of the attached sensors inside the magnetic field which creates currents that are measured by the EMA devices and after processing gives sensors' positions. The advantages of EMA is that it provides a very high frame rate (around 1250Hz but normally is downsampled to 250Hz to compensate for measurement errors), is acceptably invasive, not hazardous (providing that appropriate sterilization has been used) and not expensive. However, it does not provide information about the whole vocal tract since sensors cannot be glued in the pharyngeal region. Furthermore it provides only the position of the sensors and this technique can affect slightly articulation. Additionally, sensors have to be glued to the tongue, and wires are inside the mouth and it takes time to prepare the subject for acquisition.

One can consider using a different type of sensors in order to acquire articulatory information. An approach like this is EPG (ElectroPalatoGraphy) [Har72]. Instead of placing the sensors on the tongue the idea of EPG is to place the sensors on the palate and indirectly measure the tongue position by checking which sensors on the palate are activated when the tongue touches the palate. This is done by placing a subject specific artificial palate with electrodes connecting to sensors. There are only two wires entering the mouth (one per side), one connecting to half of the sensors and the other connecting to the other half. When the tongue touches the palate, the electric circuit between the sensors that are touched is closed, making current passing through the tongue as it activates them. The frame rate of EPG is usually around 100 Hz. Advantages of EPG is that it gives good precision of the tongue parts that are touching the palate not only in the midsagittal plane but also in parasagittal planes. Usually it samples the palate region at a 8×8 square. Additionally, there are only two cables entering the mouth from the sides making it easier for the subject (compared to EMA) to speak naturally. However, EPG presents two big drawbacks. First, it generally requires a subject specific artificial palate to be created, usually from plaster printing and second it cannot be used to study sounds that require no contact between the tongue and the palate like vowels. Despite these limits, it can be very useful to study the effects of the previous and the next phoneme on the production of the current one during continuous speech. This is thus a

good tool for studying coarticulation and generation of palatal consonants for instance.

In the last years, Magnetic Resonance Imaging (MRI) [GGM92] is becoming increasingly popular. There are several MRI subcategories, mainly depending on the place, the function and the object that one wants to image. MRI is usually based on Hydrogen (H) atoms since the molecules of fat and water (which contain hydrogen) exist at a great extend in the human body. Other molecules like Phosphorus could also be used for certain cases like teeth imaging [SBN+16]. The main idea is that a strong magnetic field is applied which forces the spin of protons of (usually) H to align to the magnetic field. Protons are precessing spinning around the axis of the magnetic lines at Larmor frequency. RF (Radio Frequency) coils of the scanner send radio frequency signals (with Larmor frequence) which make the atoms go to a higher state since they absorb energy. When the radio frequency is stopped the atoms are returning to their original state and they emit a weak signal at the Larmor frequency which decays during time. During the whole process, gradient coils are creating small distortions to the strong magnetic field which locally changes the magnetic field. As a result, the resonance frequency of proton is position dependant. By using RF pulses of specific frequencies, it is possible to control which regions will be excited. These signals are captured by the sensors (antennas) in the coil. These signals do not correspond directly to the image but to another representation of it, called k-space. For this reason the data acquired at this step is usually called raw data. The k-space is equivalent to the Fourier transform of the image. Therefore in order to pass from the raw data to the image, Inverse Fourier Transform is used. The process that uses the samples from the k-space (acquired by the antennas) in order to create the final image is called reconstruction.

Even though there are a few disadvantages in the use of MRI for imaging (like high cost, noise during the acquisition, supine position of the subject), the advantages that it offers are numerous. The first is that this is a non invasive method without known hazards. Additionally, it captures detailed images (compared to the previous described techniques) of the whole imaged region (vocal tract in our case) and not just some markers. For these reasons we choose to work with MRI data for the rest of this work.

1.3.2 Speech synthesis approaches

Speech synthesis is the production of acoustic (or even audiovisual) signals that resemble human speech using a computer-based system [Rab78]. It has a great field of applications, from talking smartphones to systems able to produce synthetic voice for blind people or for people unable to speak [BHG+14].

A speech synthesizer works by taking a text as an input and producing sound as an output. A typical speech synthesis system consists of two parts: the first which transforms the text into linguistic information and the second part which transforms the linguistic information to a sound wave.

The most popular methods used by a speech synthesiser are mainly divided into the following categories: 1) Concatenative, 2) Deep learning, 3) Articulatory. Below we are going to describe them in more detail and explain the strong and weak points of each of them.

Concatenative approach

In the concatenative approach, the speech database is divided into units (for example phonemes) that are properly selected and concatenated to produce the target synthesised speech signal. Some of the major problems of this technique are the construction of the corpus (corpus size, number of sentences etc) how to choose the appropriate units from the database and how to glue them to create the final speech signal. Additionally, since huge databases are required to include the possible units, computational costs of processing the corpus and its exploitation when synthesising speech should be taken into account as well.

In [HB96]a method is proposed for selecting appropriate units for synthesis by treating the database as a HMM (Hidden Markov Model) network, where states represent the phonemes in the database and their connections represent the potential concatenations. State occupancy cost is the estimated difference between the target unit and database unit and the transition cost is the estimation of the quality of the concatenation between potential consecutive units. By using Viterbi algorithm with pruning, the units (phonemes in this work) of the synthesised speech signal are selected. Speech is synthesised by concatenating the units.

Another approach would be to use units of varied length instead of fixed length like in [STI04] where context-dependent phoneme sequences were used as units. Units could be single phonemes or phoneme sequences that appear frequently in the training data. In this study, 86 hours of broadcast news of Japanese were used for training. Target and concatenation costs are computed and Viterbi algorithm is used to synthesise the final waveform. Their proposed method improved the results, using the mean opinion score and paired comparison test as evaluation criteria. As they state in their work, the improvements are due to two factors. The first one is that the database used was not recorded for specific speech tasks but was natural everyday speech, which improved the naturalness of synthesised speech. The second is that by using longer units the total number of concatenated chunks of speech is reduced. As a result, the computational cost is reduced and therefore bigger databases can be used. Additionally, pruning of the Viterbi algorithm can be avoided which provides a better maximum-score path.

There are also a few recent works that try to apply concatenative speech synthesis to limited resources languages like in [GS] where they study a local Indian language (Marathi). Since the main purpose is the use of speech synthesis in low resourced languages (2 hours in their case) their aim is to reduce the number of units as much as possible. Therefore they have variable length of units starting from word, moving down to syllable and further down to phoneme. Digits are treated separately. To synthesise a target text, words are used as units. If a given word does not exist in the database, syllables are considered as units for the synthesis of this specific word and again if there are missing syllables, the process is repeated using phones as units. As explained earlier using bigger units reduces the number of concatenations which results in a reduced amount of data required for training. However, prosody modeling or gluing issues at the concatenation points are some of the issues that appear.

A popular technique that is being used for modifying signal duration and/or prosodic features is PSOLA (Pitch Synchronous OverLapping Add) [CS86]. The main idea is that

speech is divided into small overlapping segments (using pitch synchronous windows). By moving them further or closer, or by copying or eliminating some windows one can synthesize the new signal with the desired properties by adding the windowed signal in the time domain [MC90]. This idea is further extended in [Kaw06] by addressing several issues like spectral distortions caused by the estimation of F_0. Another more advanced approach that gives better results and greater flexibility (at the cost of complexity) is the use of HNM (Harmonic Noise Model) [Sty01] where the speech spectrum is divided in low and high bands. The limit is defined by the maximum voiced frequency which is computed based on the fundamental frequency of the speech signal. The low band part is considered to include voiced speech and is modelled using harmonic sinusoid signals. The high band part can be represented by a modulated noise component and it is stochastic. Using these parametric models one can synthesise speech with the desired characteristics. Even though the assumption of this technique is not valid as noise can exist in the whole spectrum, synthesised speech has a high quality in terms of intelligibility, naturalness and pleasantness.

To sum up, concatenative speech synthesisers can produce high quality synthetic speech by choosing the appropriate units from a database and properly modify and glue them. However, this approach raises several issues. Some of them are the design, creation and processing of the corpus, the supervision of the concatenation (like cost issues and the number of chunks to be used) and of course the concatenation itself. Moreover, concatenative approaches do not provide any knowledge regarding speech production or the contribution of the articulators to the produced sounds. Therefore they do not provide any bridge between articulatory gestures and the speech signal produced. Furthermore, they require a huge corpus covering all expressions to be able to generate speech appropriate to any situation. Even though these approaches may give good results for applications that mainly focus on the resulting synthesised sound, we consider them not suitable for the purpose of this thesis.

Deep Learning approach

Deep architecture models have the ability to encode all the levels of speech, from the sentence to the synthetic signal. For synthesising speech, the neural network predicts speech parameters for the given input sentence, uses them to generate the speech waveform.

In order to synthesise speech with this approach, a dictionary is used to transform text into phonemes. Phonemes are then mapped to input contextual features (conveying phonetic, linguistic and prosodic information). The input contextual features are transformed into a set of answers about linguistic information such as whether the current phoneme is a vowel or not, whether the position of the current phoneme is at the beginning of a sentence, etc. Subsequently, the neural network predicts the speech parameters for a given input contextual features for each frame. Finally, from the predicted speech parameters, the speech waveform is synthesized using a vocoder.

To better handle the dependencies between speech frames, RNN (Recurrent Neural Network) based architectures such as LSTM (Long Short Term Memory) or GRU (Gated Recurrent Units), were employed to address synthesis as a sequence to sequence learning problem [FQXS14].

With advances in deep neural network based architectures, several end to end speech synthesis systems have recently been proposed, namely Tacotron [WSRS+17], Char2wav [SMK+17], Wavenet [ODZ+16].

The Tacotron speech synthesis system adopts a multi-stage encoder-decoder architecture with multiple RNNs and blocks called CBHG (convolutional block highway gated linear units), where each CBHG contains multiple convolutional layers, a highway network and a bidirectional GRU. The speech waveform is generated by the Griffin-Lim method from the output synthesized spectrogram, where acoustic features are directly conditioned on input text instead of linguistic features. This makes Tacotron a complete end to end system without explicit knowledge of a language. WaveNet architecture is based on stacks of dilated convolutions, which are termed "casual" because they do not use the knowledge of the future. The Char2Wav architecture uses RNNs for both the reader and the generator and an attention mechanism, i.e the Graves positional attention mechanism, is employed.

The deep learning approach gives very natural synthesised sound but it requires a lot of training data and time. Additionally, it uses only speech signals and it does not provide any insight about the speech production mechanisms. Therefore, we consider it not suitable for our study.

Articulatory speech synthesis

Articulatory synthesis is an important tool to study speech production because it provides better understanding of the acoustic impact of speech articulators' movements since it bridges the acoustic and the articulatory domains of speech [EL16] (articulators of the vocal tract can be seen in Fig. 1.2).

However, appropriate data techniques are required [NAH97, ANH97] for imaging the vocal tract (1.3.1) in order to create vocal tract models. There are three main categories of models

1. Geometric models: Geometric models use geometrical primitives (lines, curves, planes etc) in order to model the shape of the vocal tract. The advantage is their low computational cost, since the shape of the vocal tract is simplified unlike other methods. The main disadvantage is that the computed vocal tract does not necessarily fit the real vocal tract geometry well. Another difficulty is that moving from one shape to another is not guarranteed to be natural when using artificial primitives [BJK06].

2. Biomechanical models: The main advantage of biomechanical models is that they can model how speech movements of the articulators are constrained and therefore produce a very realistic model of how the vocal tract behaves when someone speaks. However, they have many degrees of freedom, which makes the control difficult and they require a great amount of data and computational power to be exploited. In addition the adjustment of the underlying physical model of muscles is not easy and often not complete [FVVDD+06].

3. Statistical models: Statistical models are derived from images of the vocal tract via some data analysis method, usually from PCA (Principal Component Analysis). They require a small number of parameters in order to model the vocal tract, which makes them computationally cheap, at the cost of tuning. Since the statistical models follow the data a-priori, attention should be paid to the tuning so that the output would be physically possible [Mae90], [HM11].

In the following subsections 1.3.2-1.3.2 we are going to present some details regarding the advances of each model based on the presentation of Birkholz et al. [BJK06].

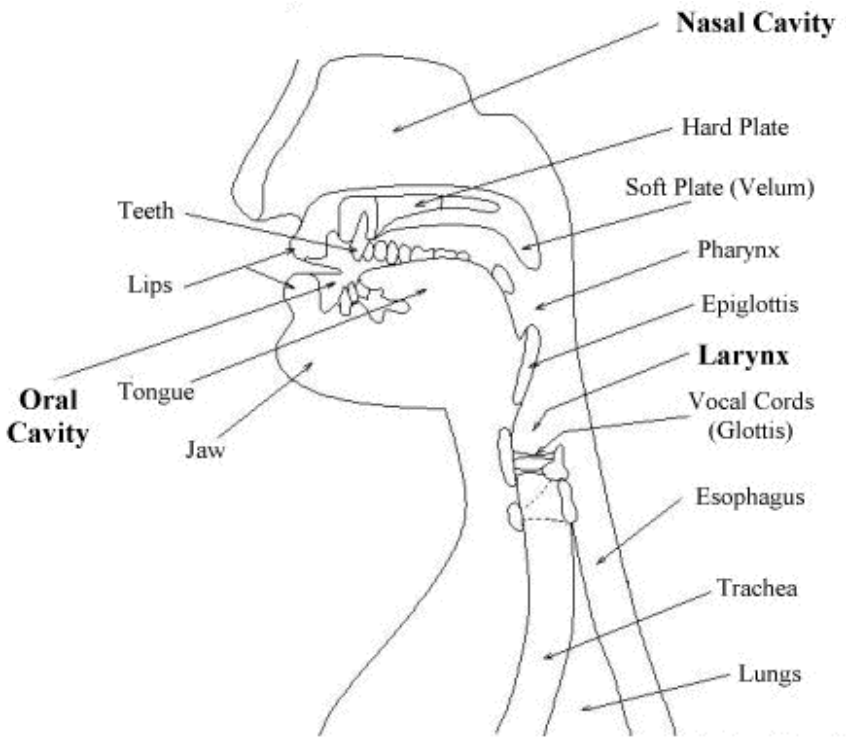


Figure 1.2: Speech production system focusing on speech articulators. Image from: http://athena.ecs.csus.edu/ changw/Sounds/SpeechRecog/acoustics.html

Geometric models of the vocal tract shape

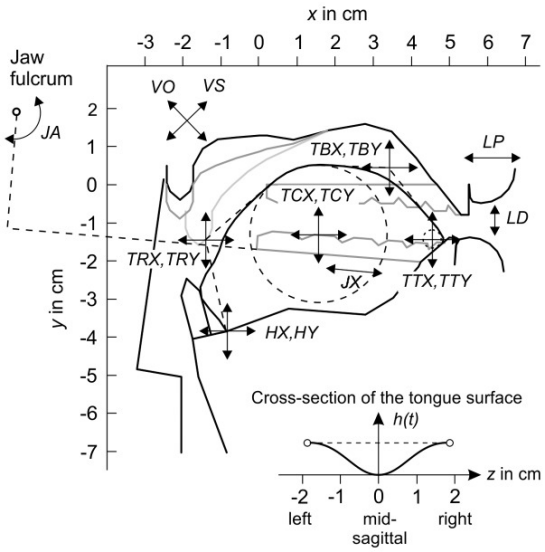
Geometric models were the first type of models employed for modeling the vocal tract by using geometric primitives. One of the first models is presented in [LS71]. This model mainly focuses on the shape of the lips, the jaw opening, the tongue shape and the larynx height. X-ray data of one male speaker were used. Four vocal tract shapes were used to describe the tongue, corresponding to neutral configuration, palatal hump, velar hump and pharyngeal hump. All the other configurations are described by using interpolation between these primitives. The model is evaluated by synthesising formants of several phonemes and explore the effect of each part of the model on formants. The main contribution of this work concerned the treatment of the jaw opening separately

from other articulators, which enables a simpler description of vowel production. This model was further extended in [Mer73] where velum and hyoid bone were also taken into account to build a more accurate model. Apart from all isolated English phonemes, this extended model was able to intelligibly synthesise all VCV of English. However, both models were created using male subjects. In [DB10] an attempt is made to adapt this model to a female speaker and study the effect of the larynx length in articulation.

These models are 2-dimensional and newer attempts tried to create 3D geometric models of the vocal tract like in [BJK06] where a 3D model is created using X-ray images of a male speaker. It is based on seven 3D meshes that describe the articulators and the wall of the vocal tract. The articulators are modelled using 23 parameters in total, 15 for the tongue, 3 for the jaw, 1 for the velum, 2 for the lips and 2 for the hyoid. For the tongue modeling, two circles, two Bezier curves and three line segments were used. Three parameters were used for each circle (2 for the center and 1 for the radius), three parameters for each Bezier curve and one parameter per line segment. The two circles are used to model the tongue body and the tongue tip. The back of the tongue and the tongue blade are formed by the two Bezier curves. Finally, the anterior edge of the tongue contour is described by three line segments. Regarding the rest of the vocal tract, one parameter is used to describe the velopharyngeal port, two parameters to describe the lips, one for the lip opening and one for protrusion, three parameters to describe the jaw motion, two for the position of the root of the jaw and one for the jaw opening and finally two parameters to describe the position of the hyoid bone. This model can synthesise realistic vocal tract shapes and natural sounding speech. This model has also been modified to be used in articulatory singing synthesis [Bir07]. Further extensions of this model are described in [Bir13] where MRI data were used to create a similar geometric model in order to study CV coarticulation (Fig. 1.3).

The model uses again 23 parameters but in this version 2 parameters are used for the jaw (instead of 3), 3 parameters for the lips (instead of 2), 2 parameters for the velum (instead of 1) and 14 parameters for the tongue (instead of 15). To evaluate the model 56 CVs (7 consonants, 8 vowels) were studied. 20 listeners were employed to test the intelligibility of the synthsised CVs. Results varied per speaker but globally the vowels were more intelligible than the consonants.

Overall, geometric models offer the advantage that they are computationally cheap. However, in order to achieve better synthesis results the number of parameters should be increased. Their main disadvantage is the a-priori definition of the parameters to be used, which is quite an intuitive process and may not necessarily represent reality well. As stated in [BJK06, LS71] some of the parameters for the geometric models are probably correlated, which first shows the intuitive way of choosing parameters and seconds shows that models are more complex than needed. Fitting geometric models to real vocal tract shapes is highly dependent on the geometric simplifications used because these models are based on geometrical primitives and not derived from real data. Therefore their use is probably not the best choice to study speech production.



Control parameters of the vocal tract model.

Name	Description	Min.	Max	Unit
HX	Horiz. hyoid position	0.0	1.0	
HY	Vert. hyoid position	−6.0	−3.5	cm
JX	Horiz. jaw displacement	−0.5	0.0	cm
JA	Jaw angle	−7.0	0.0	deg
LP	Lip protrusion	−1.0	1.0	
LD	Vert. lip distance	−2.0	4.0	cm
VS	Velum shape	0.0	1.0	
VO	Velic opening	−0.1	1.0	
TCX	Tongue body center X	−3.0	4.0	cm
TCY	Tongue body center Y	−3.0	1.0	cm
TTX	Tongue tip X	1.5	5.5	cm
TTY	Tongue tip Y	−3.0	2.5	cm
TBX	Tongue blade X	−3.0	4.0	cm
TBY	Tongue blade Y	−3.0	5.0	cm
TRX	Tongue root X	−4.0	2.0	cm
TRY	Tongue root Y	−6.0	0.0	cm
TS1	Tongue side elevation 1	−1.4	1.4	cm
TS2	Tongue side elevation 2	−1.4	1.4	cm
TS3	Tongue side elevation 3	−1.4	1.4	cm
TS4	Tongue side elevation 4	−1.4	1.4	cm
MA1	Min. area tongue back region	0.0	0.3	cm^2
MA2	Min. area tongue tip region	0.0	0.3	cm^2
MA3	Min. area lip region	0.0	0.3	cm^2

Figure 1.3: Example of a geometric model of the vocal tract. On the top, the geometric primitives, on the bottom the parameters' explanation. Image from [Bir13].

Biomechanical models

Biomechanical models are trying to describe the physiology of the articulators by describing their tissue properties and behaviour based on the activation of the various controlling muscles. In [PP97] EMA and X-ray data were used to create a 2D biomechanical model of the tongue. 7 muscles were considered for the model and the finite element method was used to simulate the muscles' behaviour, using 48 elements and 63 nodes. 24 combinations of VV (Vowel Vowel) of one speaker were studied by simulating data and comparing them with the original in the kinematic and acoustic domains. In cite [PPZP03] an improved 2D version of the previous biomechanical model of the tongue is presented, using 192 elements and 221 nodes. Both models are 2D and focus on the tongue. In [SP16] a 2D tongue model is combined with lips and jaw biomechanical models and evaluated on 10 French vowels. The model was able to produce both rounded and unrounded types of vowels, but within each type, vowels were not significantly different. 9 parameters were used, 6 for the tongue (superior longitudinal was not treated separately as in the previous models), 2 for the lips and 1 for the jaw.

There are also attempts to create 3D biomechanical models of the tongue. In [HPP17] MRI and CT (Computed Tomography) images were used to create a 3D biomechanical model. Its parameters were tuned based on in vivo and ex vivo (of fresh cadaver) measurements of the mechanical properties of the tongue. The model consists of 8679 nodes and 6534 hexahedral elements and describes the behaviour of 9 muscles (5 extrinsic and 4 intrinsic). Simulation results showed that it describes what is reported in the literature well regarding the deformations of the tongue during speech production.

As seen from the above description, biomechanical models look a good choice for articulatory synthesis and speech production studies as they provide a good description of the kinematics of articulators and the relation between muscle activations. However, they have some serious disadvantages that make their use limited from a practical point of view. First, they have many degrees of freedom which makes them difficult to control [BJK06]. As a result, a huge amount of data is needed and a huge computational power is required to run realistic simulations. Acquiring big amount of data is also a significant drawback because acquisition is invasive and difficult. An additional reason for the high computational demands is the use of the finite element method in the majority of the biomechanical models, which is a method with a high computational cost [RLN$^+$17]. Furthermore, due to their complexity, they cannot model all the muscles of the tongue and they focus on the major ones which can be problematic since there is some speaker variability in the description of the non major muscles. Even though this may not look a big issue it may have some important effects on speech synthesis. Such an example is the case of palatoglossus muscle which is a small muscle of the tongue [TV06] with no significant impact on the tongue contour during speech production and which is therefore often not considered in most tongue models [PP97, HPP17]. However, this muscle plays an important role in the velar opening which is a crucial articulator for nasal sounds [KF$^+$82, PP97]. Therefore despite their potential interest, biomechanical models fail to model all the details of the articulators. To sum up, although biomechanical models can give some insight about speech production, they do not seem to be the best choice from a practical point of view.

Statistical models of the vocal tract shape

Statistical models are using data analysis methods applied on contours of speech articulators in medical images to model their behaviour. In [Mae90, Mae91]an articulatory model created from cineradiographic and labiofilm data was used to study effects of articulation on acoustics. This model consists of 7 parameters to describe the whole vocal tract and was used to explore the impact of articulators' movement on the vowels by checking the F1-F2 formant space. Additionally, intra-speaker variability was studied for two speakers during the production of /i/ and /a/ in several phonetic contexts. In [Mae82] a numerical scheme to solve the acoustic equations of the vocal tract is presented based on time-domain simulations (using an electric-acoustic analogy) The resulting synthesiser can produce good quality synthesised phonemes in terms of naturalness and intelligibility. This model was applied for articulatory speech synthesis of VCV from X-ray films [LLM+13b] and from static MRI [TEL17].

There have also been attempts to create 3D models like in [BBR+02] where MRI and video data were fused for articulatory modelling of the tongue and lips. 3D MRI data of the vocal tract was acquired in 3 stacks of parallel slices, perpendicular to the midsagittal line. The first stack was coronal, the second oblique tilted 45 degrees and the last one was axial. In total 25 static positions were used for the model construction. Subject's face was video recorded during a separate session using 32 face markers. Data modalities are fused and linear and principal component analysis are applied to create the final model with 9 parameters. This model was an extension of the model proposed in [BBRS98]. A 3D model of the velum is presented in [SB05] using MRI and CT data. Data from one speaker was used to build the model using PCA. This model uses 1 parameter for the control of the velum but it does not consider the cases where there is contact between the velum and other vocal tract structures. This model was extended to also describe the cases where there is contact between the velum and the tongue or the pharyngeal wall. Additional EMA data of the velum with MRI was used to build a model with 2 degrees of freedom [SB08]. Similar approaches were used in [BBR+02] to create models for most of the speech articulators using PCA.

These models were created using data from several acquisitions of target phonemes from one subject, which implies that inter-subject variability is not addressed. More recent approaches are attempting to create articulatory models from direct acquisitions. In [LLM+13b] a 2D model of the velum is presented and evaluated by synthesising sentences that include nasal phonemes. X-ray data of continuous speech from one speaker were used. Velum was manually delineated and PCA was applied to them. Since the proportion of nasal sounds was significantly smaller, some data augmentation technique was required so that PCA will describe the movement of the velum and not treat is as noise. The final model consists of 2 parameters and describes around 70% of the variance. In [LETV18] a new model of the velum and the epiglottis is presented. This model was derived mainly using static MRI data (and very few instances of rtMRI). Its main idea is the representation of the velum and epiglottis contour from their centerlines. Then PCA is applied on them to build the statistical model and regenerate the contours from the centerlines. This model is more general because it can describe special cases of velum configurations (like situations when the velum rolls up to itself) that the previous model

was unable to approximate and is also more robust to delineation errors that may happen during data preprocessing.

In general, statistical models are the most promising type of articulatory models because they can fully describe and completely control the geometry of the vocal tract with a good realism since models are derived directly from images. They need a limited number of parameters compared to other articulatory types of models but they offer a good flexibility because it is possible to change static parts easily (for instance the palate) and they involve articulatory compensatory possibilities which are important to adapt them to other speakers. Additionally, since they use a small number of parameters they do not require a big amount of computational power that biomechanical models do. A weak point is that delineated contours are required for building them which can make the construction of statistical models quite hard sometimes since delineation usually has to be made manually to ensure high precision. However, with advancements in automatic segmentation algorithms [TGH+19] this issue could be efficiently solved.

1.3.3 Articulatory data augmentation

Regardless of the improvements in rtMRI techniques, research in speech production and modeling of vocal tract faces limitations in building a complete articulatory synthesis model [LETV18]. 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, like better understanding of the existence of voiced fricatives [EL16] for instance.

Acquisition of articulatory data raises several issues such as the capability to extract precise speech dynamics in time and space, 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 [WMWK90], electromagnetic articulography [PCS+92, Wre00a], electropalatography [Har72] and ultrasound [SD95, WIT+05].

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 (between seven and fourteen seconds in our case). On the other hand, vocal tract images recorded with rtMRI yield natural and complex information about articulatory spatiotemporal movements. The rtMRI protocol selects only one slice usually the midsagittal plane and actually captures tissues within the midsagittal slice at 50 Hz approximately in real time. The major benefit of capturing rtMRI images is that it provides considerable amount of data which suffices to analyze continuous speech articulator movements [NTR+14, TN16, RTP+18].

The recent development in rtMRI imaging techniques provides 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 [CSF+15], coarticulation in VCV sequences [DHMS02], characterization of click consonants in African languages [PZL+14].

Besides investigation in phonetics, rtMRI can have a big impact in automatic speech and speaker recognition to supplement acoustic signal with the structure of the physical system and concequently increase the performance of recognition systems [RTP+18].

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 [LZL+19, ADMC09], it could be still interesting to be able to artificially synthesize articulatory data that could enlarge existing databases and make speech production studies easier.

1.3.4 Generic speaker modeling

Understanding speech gestures (independently of the subject) is of primary importance, therefore the objective of constructing a generic speaker model is an important avenue of research. A generic speaker model could help in advancing speech production studies by allowing us to study speech dynamics, synthesise articulatory gestures, adapting models etc. The field of neuroimaging provides excellent resources about techniques and methods that one can use in order to create such a model. Generic models for the brain, usually called atlas have been proven to be a powerful tool for brain related studies. Several types of atlases exist, each of them focusing on a different aspects. For example, there are atlases that focus on the static description of the brain anatomy [SDM+04] while others on the dynamic evolution of the brain with respect to age [CBWJ13]. There are also studies focusing on the registration methods [XRLH18] or automatic segmentation techniques for adults [KMAS+11] or children [GRH+08].

There are some attempts to use similar approaches in the speech production field mainly for describing and analysing the tongue motion [SDD+01, XSG+19] by using tagged cine-MRI. Dynamic atlases of the whole vocal tract could provide valuable assistance in the study of speech production when the vocal tract geometry changes rapidly, and consequently when changes of the area function have a strong acoustic impact [STTN17, THM+06]. Even though the use of atlas is not widespread in speech research, there are some attempts to describe the temporal evolution of the vocal tract [WLM+15, WXL+15, WXL+18]. However there is still room for improvements since these approaches lack flexibility (see Chapter 5 for more details).

1.4 Thesis organization

The current thesis is organised as follows

In Chapter 2 we present the various existing MRI databases for speech research, data included, their technical specificities, strong points and limitations. We argue about the need of new databases and we describe the design and the creation of such a database that we used for the work of this thesis.

In Chapter 3 we used 3D static MRI images of the vocal tract during vowel production to acquire the three dimensional vocal tract shape. We use acoustic simulations in order to study the impact of simplifying acoustic hypotheses by applying simplifications on the geometry of the vocal tract at the level of the velum and the epiglottis. We also explored the effect that the head position has on phonation.

In Chapter 4 we explore the combination of 3D static MRI with 2D rtMRI data in order to reconstruct 3D dynamic data, as a data acquisition post processing step. Additionally, we present some methods based on image deformation fields that synthesise volumetric data and midsagittal 2D dynamic data of CV production based on 2D rtMRI acquisitions.

In Chapter 5 we describe an algorithm to create a spatio-temporal CV atlas of the vocal tract. In this Chapter, we describe the creation process starting from the MRI acquisition protocol, subject selection and the atlas creation process using non-rigid image transformations and an adaptive kernel method in order to create the final generic speaker model.

In Chapter 6 we discuss the contribution of this thesis and we list some directions that look interesting and promising for future research.

Chapter 2

Databases

2.1 Requirements for a database

Database is a collection of data, properly organised so that one can easily search and find the items that we are looking for. They play an important role in speech production research as they provide the data many studies are built on. The data that a MRI database for such studies could include vary from static or dynamic MR images of the vocal tract, audio recordings, languages, number of speakers, other modalities, several emotions etc (see Section 2.2 for examples). Depending on the main question addressed by a study some existing databases may be more appropriate than others. It is quite clear that the choice of a database for a certain study is based on the type of data included in the database (like static images or dynamic ones). However, it also depends on the way that the data were acquired.

To illustrate this clearly we will use the real-time MRI data as example. One aspect is that different articulatory phenomena have different requirements in terms of temporal and spatial resolution so that one could be able to properly study them. In Fig. 2.1 spatial and temporal resolution requirements are shown for several speech tasks.

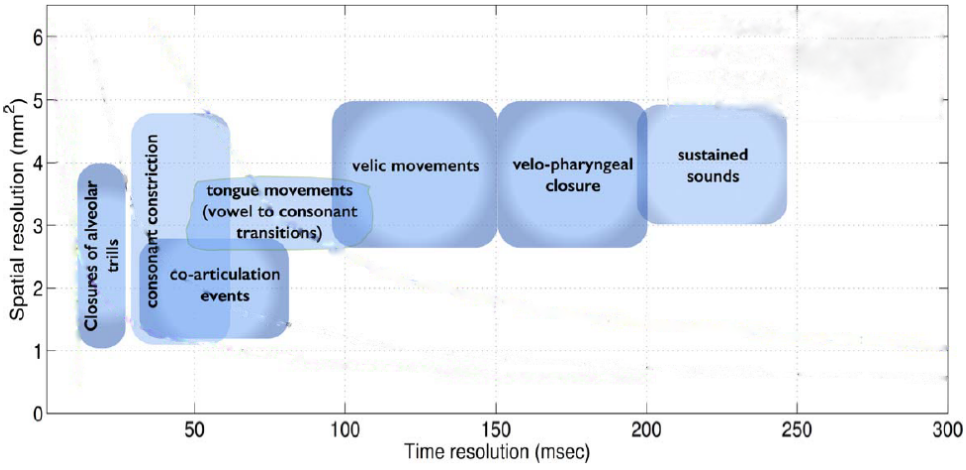


Figure 2.1: Spatio-temporal resolution requirements for several speech tasks [LSMN16]

Another important aspect is the k-space sampling technique used to reconstruct the image as presented in section 1.3.1. This is one of the steps in the MRI acquisition process that limits the frame rate of real-time MR images and several approaches have been explored in order to speed it up [Hen99]. Two of the most common approaches is spiral and radial sampling but there are more attempts based on heuristic approaches [CCW13]. Each approach offers different image quality and different artifacts can appear in the reconstructed images. Fig. 2.2 shows some examples of image artifacts based on the sampling technique.

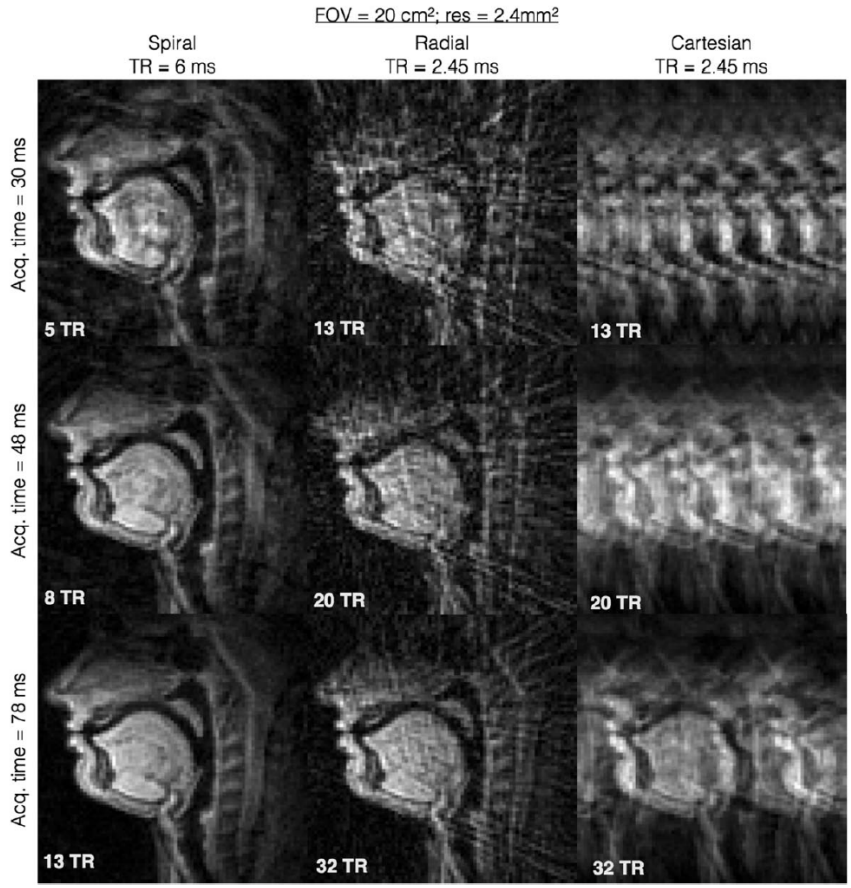


Figure 2.2: Aliasing artifacts for spiral, radial and Cartesian acquisitions for different acquisition times [LSMN16].

Of course there are many more data acquisition aspects one should take care of when choosing the appropriate database for the target purpose for example whether the audio was recorded simultaneously with the MRI acquisition or during a separate session, the denoising algorithm etc. Everytime the most appropriate database should be chosen based on the research question and some compromise.

One of the first studies that uses MRI data is the one of Baer et al. in [BGGN91] were the volume of the vocal tract was acquired for two speakers during the production of four vowels and their area functions were computed and studied. Later, in the work of Story et al. [SHT96] MRI and EBCT (Electron Beam Computed Tomography) volumetric

data of one speaker were used during the production of two vowels in order to study and compare the area functions acquired with these two techniques. Over the years and the technology advancements, many more studies have been conducted in the field of speech research using MRI data. Some examples are the work of Badin et al. [BBR+02] where MRI coupled with video data are used to create 3 dimensional articulatory models of the tongue, the lips and the face, Silva and Teixeira in [ST17] used rtMRI to define the critical articulators during speech production, Badin et al. [BTL19] used MRI data of one multilingual speaker to study the effect of coarticulation in three languages etc. Therefore it becomes clear that MRI databases can play an important role in advancing speech production research by providing appropriate data to the research community in order to address various research questions.

2.2 MRI databases for speech production research

As described in Chapter 1 our plan for the first part of the work is to conduct some speech production experiments by using acoustic simulations. Below we are reporting the currently available MRI databases, describe their limitations and explain the need for the creation of a new one suitable for our study. A list of available databases can be seen in Fig. 2.3. This is not an exhaustive list but a list which represents the range of possible types of MRI databases.

Database name	Language	Subjects	Data Type	Spatial Resolution	Frames per second	Audio	Data
Vowels MRI	English	4 (2m-2f)	Static 3D	256*256	-	Yes	9 vowels
ATR MRI	Japanese	1 (m)	Static 3D	512*512	-	Yes	5 vowels
rtMRI-TIMIT	English	10 (5m-5f)	Real time 2D	68*68	23.18	Yes	460 sentences
rtMRI for Portuguese	Portuguese	1 (f)	Real time 2D	128*128 & 64*64	14	Yes	Vowels isolated & in VCV
USC-EMO-MRI	English	10 (5m-5f)	Real time 2D	68*68	23.18	Yes	Grandfather passage & 7 senntences (in 4 emotions)
USC speech and VT morphology	English	17 (8m-9f)	Real time 2D & static 3D	68*68 & 150*180*60	23.18	Yes	24 CVC & 54 VCV & passage reading & continuous speech
Seeing speech	English	1 (f)	Real time 2D	90*86	7	Yes	103 phonemes in several context

Figure 2.3: List of MRI databases with data information

2.2.1 Vowels MRI database

Vowels MRI database [vow] includes 3D scans (stack of 2D slices) of nine American English vowels from four speakers (2 male - 2 female). The vowels included are /aa, ae, ah, eh, er, ey, ih, iy, ow, uh, uw/. For all phonemes there are 36 axial slices and for most of them there are additionally 38 coronal slices. Image resolution is 256×256. Two types of clean audio recording are provided at 8 KHz, one with the subject lying on a sofa with sustained phonation and one with the subject standing and speaking naturally. Potential uses of this database is examining intersubject variability of oral area during speech production of several vowels [HJPA+03]. Sample images of this database can be seen in Fig. 2.4

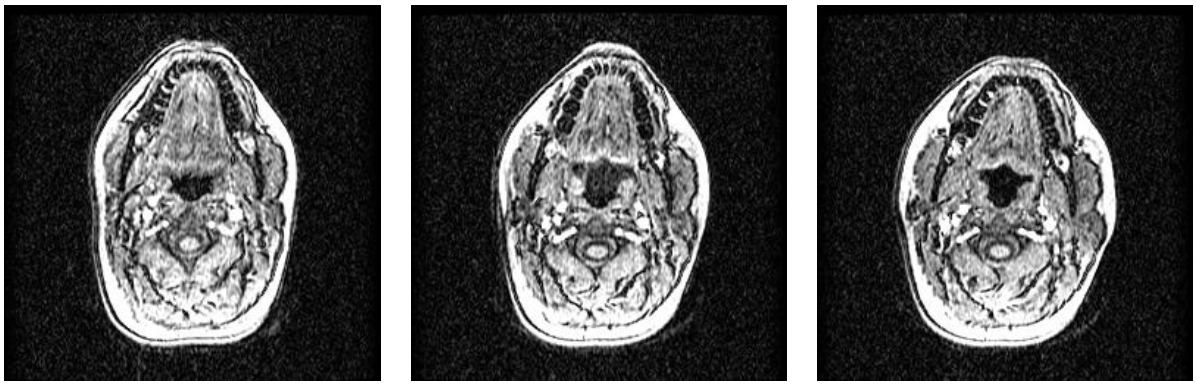


Figure 2.4: Examples of axial slices at the level of the tongue of f1 speaker during the production of /ah/, /ey/, /iy/ vowels (from left to right). Images taken from [vow].

2.2.2 ATR MRI database

The ATR MRI database [KTAH09] includes 3D static MRI scans of the five Japanese vowels of one male subject. Every vowel was repeated 64 times. 51 slices were acquired at a resolution of 512×512. cineMRI data are also included at a resolution of 256×256 at 30 fps (frames per second). Database also includes simultaneous and separate audio recordings at a sampling frequency of 48 KHz. Examples of studies with this database include the investigation of the correlation between the body size, vocal tract length and formant, pitch frequencies [HKT+12], vocal tract length estimations based on vowels [KKT+14], temporal changes in area function [THM+06] and others. Examples of images from this database can be seen in Fig. 2.5.

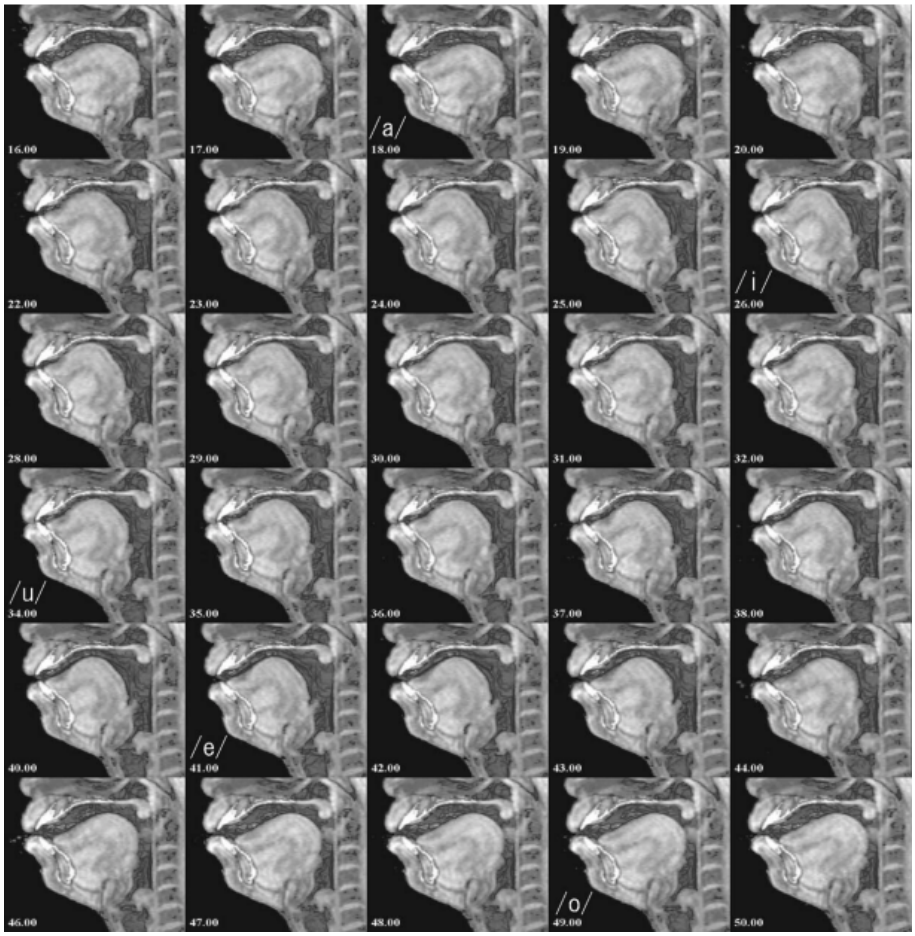


Figure 2.5: Example of cineMR images in ATR MRI database. Image shows selected frames from the production of five Japanese vowels /a/, /i/, /u/, /e/, /o/. Image from [THM+06].

2.2.3 rtMRI-TIMIT database

rtMRI-TIMIT database [NBG+11] includes 2D real-time scans of the midsagittal plane from ten subjects (5 males - 5 females). Resolution of the images is 68×68 and the frame rate is 23.18 fps. Subjects are uttering 460 sentences from the MOCHA-TIMIT corpus [Wre00b]. The database includes a wide range of phonemes of American English in several contexts. Additionally it includes simultaneously acquired audio recordings aligned with the video. Audio was acquired at a sampling frequency of 20 KHz and was denoised with the algorithm presented in [BNNN06]. Phonetic transcriptions are also provided for all the data using the algorithm presented in [KBG+11]. For four of the subjects (2 males - 2 females) EMA recordings with sound of the same 460 sentences are also provided. This database has several applications like estimating articulatory dynamics [PLK+11], dynamic articulatory modeling [BKGN10], articulatory-acoustic mapping [GN10], articulatory recognition [KBRN11], phoneme recognition [DKM18] etc. Examples of images from rtMRI-TIMIT database can be seen in Fig. 2.6.

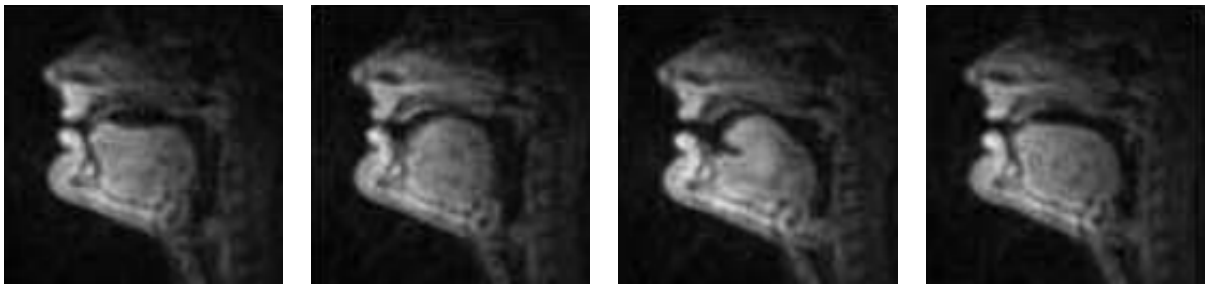


Figure 2.6: Examples of rtMRI from rtMRI-TIMIT database. Images show m1 speaker pronouncing /s/, /iy/, /er/, /ah/ (from left to right). Image from [NBG+11].

2.2.4 rtMRI database for Portuguese

rtMRI for Portuguese [TMO+12] includes 2D real-time scans of the midsagittal plane from one female subject. The frame rate is 14 fps and the resolution of the images is in some cases 64×64 and in the rest 128×128. This database is mainly focused on the nasal vowels of Portuguese. It includes rtMRI scans of the nasal vowels in several positions within a word and the isolated nasal and oral vowels with the aim of studying gestural dynamics. It also includes scans of nasal consonants in the VCV context. The corpus was designed so that it can be compared with EMA data presented in [OMT09]. Simultaneous audio recording were acquired at a sampling frequency of 16 KHz and synchronised with the MRI videos. Audio was denoised by OptiMRI software. Potential uses of this database could be automatic segmentation algorithms of the vocal tract [ST15], study of the velar movement of nasal vowels [MOST12], study of the oral articulation of nasal vowels [OMST12] etc. Images from rtMRI for Portuguese database can be seen in Fig. 2.7.

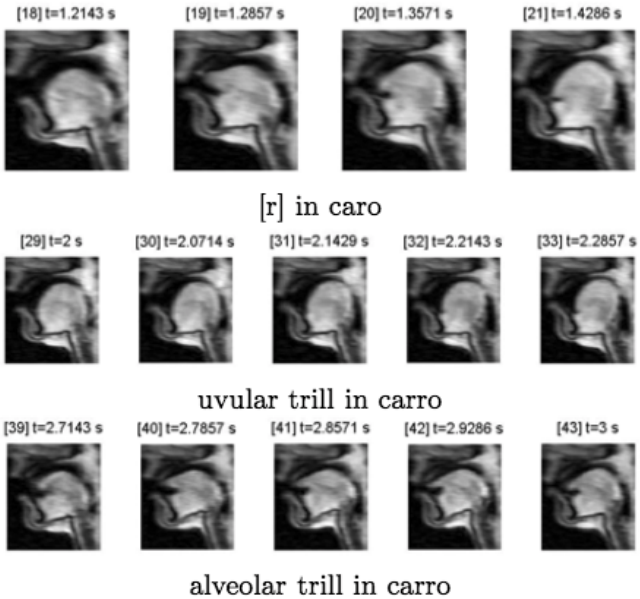


Figure 2.7: Example of rtMR images in rtMRI for Portuguese. Image from [TMO+12]

2.2.5 USC-EMO-MRI corpus

USC-EMO-MRI corpus [KTK+14] includes 2D rtMRI midsagittal scans of emotional speech (anger, sadness, happiness, neutral) from ten actors (5 males - 5 females). The corpus is the grandfather passage and seven custom made additional sentences. Speakers pronounced the corpus in American English with four different emotions several times. More specifically, the grandfather passage was repeated once per each emotion with normal pace and twice for neutral, one at a normal and one at a fast pace. For the seven sentences, subjects uttered each of them seven times per each emotion. Image resolution is 68×68 and the frame rate is 23.18 fps. Synchronised audio from simultaneous recordings is also provided. Audio was sampled at 20 KHz and denoised using the algorithm presented in [BNNN06]. Applications of this database could include articulatory emotion correlation [KLN11] or vocal tract segmentation [KKLN14]. Images from USC-EMO-MRI corpus can be seen in Fig. 2.8

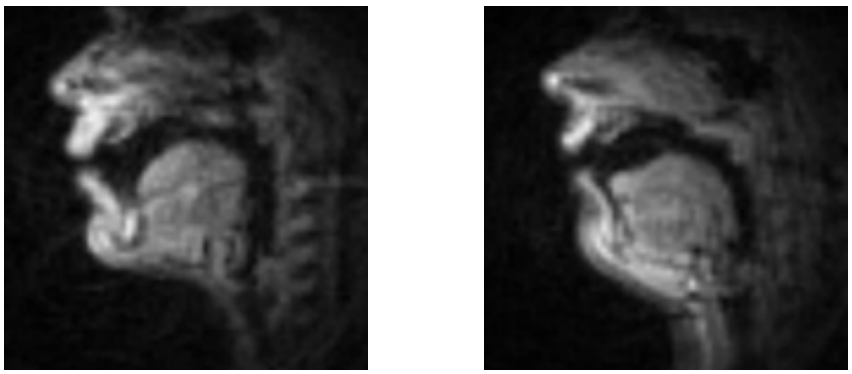


Figure 2.8: Examples of real-time MRI from USC-EMO-MRI corpus. On the left is m1 speaker and on the right f1 speaker. Image from [KTK+14].

2.2.6 USC Speech and Vocal Tract Morphology MRI database

This database includes 2D rtMRI scans of the midsagittal plane and 3D volumetric MRI scans [SST+17]. 17 subjects (8 males - 9 females) were included. 3D acquisitions include, apart from some morphological indicators, scans of 13 sustained vowels and 14 sustained consonants within one phonetic context each. The 3D volume resolution is $150 \times 180 \times 60$. 2D corpus mainly includes several repetitions of 24 CVC, 54 VCV, passage reading, sentence reading, and spontaneous speech. Resolution of the images is 68×68 and the frame rate is 23.18 fps. Aligned, simultaneous audio recordings are also included in the database. Audio was recorded at a sampling frequency of 100 KHz, downsampled to 20 KHz and denoised using the algorithm described in [BNNN06]. USC Speech and Vocal Tract Morphology MRI database could be useful for implementing airway segmentation algorithms of the vocal tract [EL] or speech identification algorithms [SSF18]. Sample images from dynamic and static images of this database can be seen in Fig. 2.9 and Fig. 2.10 respectively.

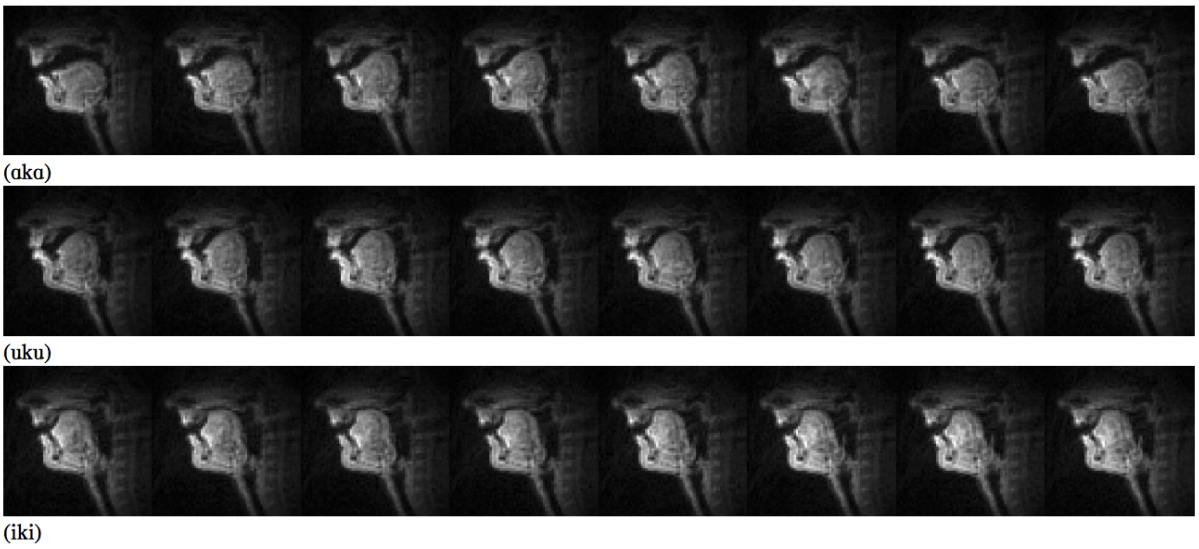


Figure 2.9: Example real-time images in USC Speech and Vocal Tract Morphology MRI database. Images show the midsagittal rtMR images of m3 speaker pronouncing /k/ in different contexts. Differences in articulation are clear due to coarticulation. Image from [SST+17].

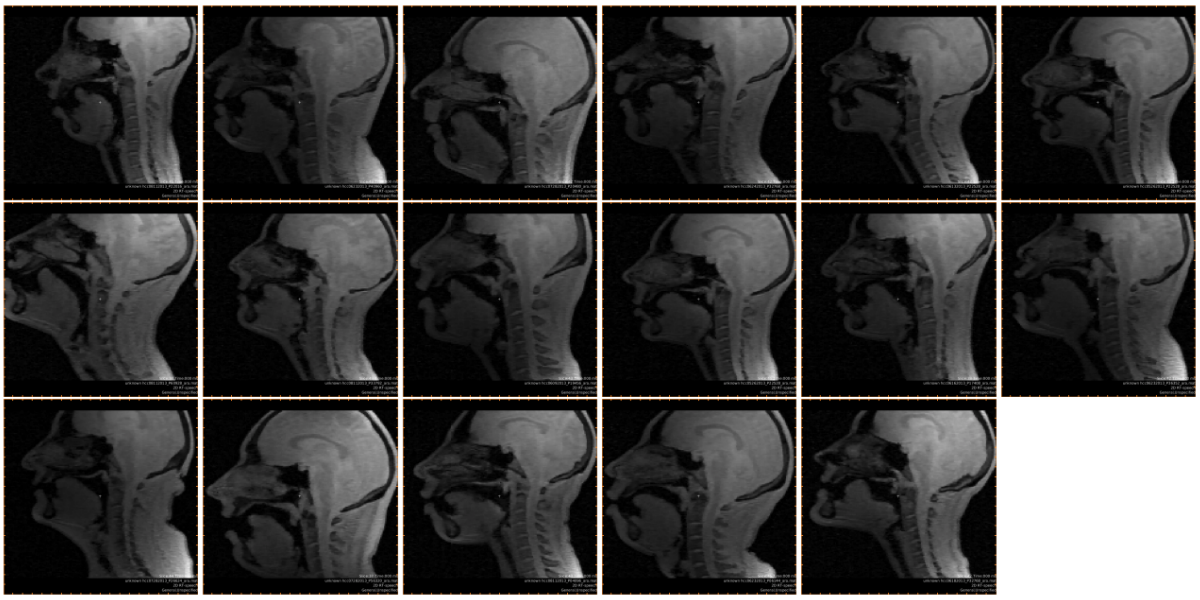


Figure 2.10: Example static images in USC Speech and Vocal Tract Morphology MRI database. Images show the midsagittal slice of volumetric images of all 17 subjects included in this database. Image from [SST+17].

2.2.7 "Seeing speech" database

"Seeing speech" [LSSS+15] is an online resource for speech production studies. It includes rtMRI videos (with the aligned acoustic signal) of the midsagittal plane of 103 sounds in several phonetic context. For the complete dataset, one female speaker was used. Image resolution is 90×86 and frame rate is 7 fps. Ultrasound images are also provided for the same female speaker. Part of the ultrasound data are also available from a male speaker. The ultrasound image resolution is 512×276 and the frame rate is 30 fps. This database can be used for velopharyngeal activity studies [SVP17]. Example images of seeing speech database can be seen in Fig. 2.11.

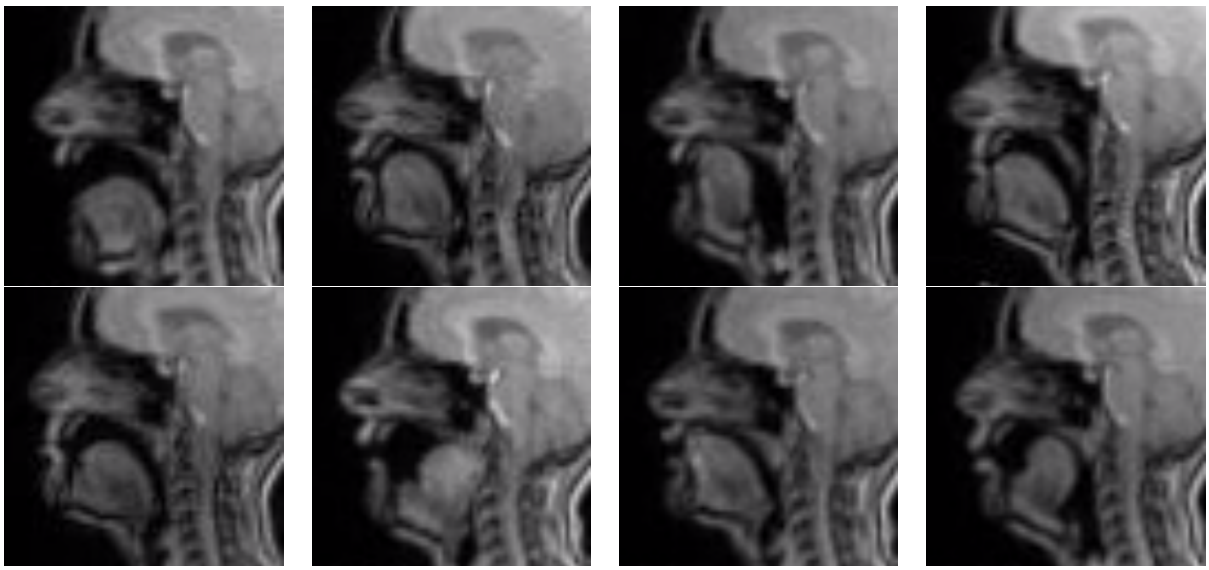


Figure 2.11: Examples of real-time MR Images from "seeing speech" database. From left to right, top to bottom there are the following phonemes: /a/, /f/, /i/, /m/, /p/, /q/, /t/, /u/. Image from [LSSS+15].

2.3 ArtSpeechMRIfr

As presented above, there are several databases that offer a large amount of resources for speech production/synthesis studies. However, most of them are for English and they do not include processed material of the data, like articulatory contours, which are required to enable investigations of speech production phenomena. Additionally, the image quality is fairly low compared to the quality that one can expect from more recent approaches. For these reasons, ArtSpeechMRIfr was chosen for this work since it provides high quality MR images and processed material for French.

2.3.1 General description of the ArtSpeechMRIfr database

ArtSpeechMRIfr is a database that includes real-time and static MR images of the vocal tract. The database contains also processed data: denoised speech, phonetically aligned

annotations, articulatory contours, and vocal tract volume information. All together this provides a rich resource for speech research. The database is built on data from two male speakers of French. It covers a number of phonetic contexts in the controlled part, as well as spontaneous speech, 3D MRI scans of sustained vocalic articulations, and of the dental casts of the subjects. The corpus for rtMRI consists of 79 synthetic sentences constructed from a phonetized dictionary that minimises the duration of acquisitions while keeping a very good coverage of the phonetic contexts which exist in French. The 3D MRI includes acquisitions for 12 French vowels and 10 consonants, each of which was pronounced in several vocalic contexts. Articulatory contours (tongue, jaw, epiglottis, larynx, velum, lips) as well as 3D volumes were manually drawn for a part of the images.

2.3.2 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 for two male subjects. 3D static data (3D MRI) for all twelve subjects and rtMRI for the rest ten subjects was recorded at Nancy Hospital, France. Midsagital images of the ten subjects can be seen in Fig. 2.12

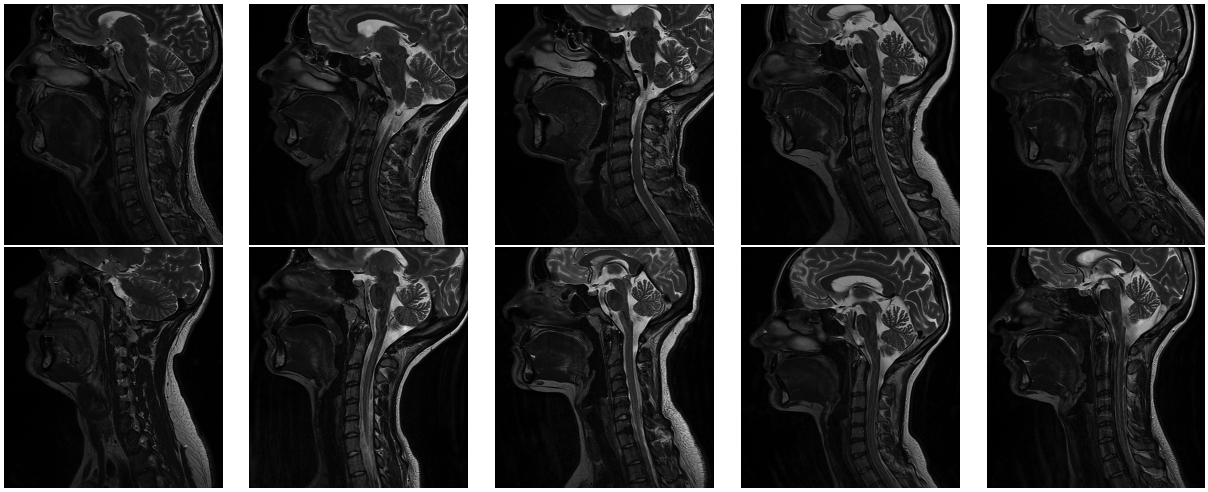


Figure 2.12: Examples of midsagittal slices from ten of ArtSpeechMRIfr subjects.

Subjects

The selected subjects are 12 adult French native speakers speaking French, 7 are male and 5 female. The MRI data was recorded at Nancy Central Regional University Hospital under the approved medical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053).

2D data

rtMRI dataset was recorded on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). A radial RF-spoiled FLASH sequence [UZV$^+$10] was used, with TR = 2.02 ms,

TE = 1.28 ms, FOV = 19.2 × 19.2 cm, flip angle = 5 degrees, and slice thickness 8 mm. Pixel bandwidth is 1600 Hz/pixel. Image resolution is 136 × 136. The acquisition time varied from 34 sec to 90 sec, mostly about 60 sec. The protocol described in [NZK+13] was followed. Images are recorded at a frame rate of 50 frames per second with the algorithm presented in [UZV+10].

3D data

For one male subject, data was recorded on a General Electric Signa HDxt 3T scanner (GE healthcare, Chicago, Illinois, United States). A 3D FGRE (TR = 3.12 ms, TE = 1.084, FOV = 26 × 26 cm, flip angle = 10 degrees) was used for the acquisition. Scan slice thickness is 2 mm, spacing between slices is 1 mm and pixel bandwidth is 488 Hz/pixel. Acceleration factor is 2. The image resolution is 256 × 256 with 76 slices. Duration of one acquisition is 12.7 seconds. Examples of mid-sagittal cuts of this data are shown on figures 2.13 and 2.14.

For the remaining eleven subjects, data was recorded on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). A 3D VIBE (TR = 3.57 ms, TE = 1.43, flip angle = 9 degrees) was used for the acquisition. Acceleration factor is iPAT = 3. Scan slice thickness is 1.2 mm, FOV = 22 × 20 cm and pixel bandwidth is 445 Hz/pixel. Data is divided in two parts.

In the first part, audio was recorded just before the MRI scan started, and the subject was asked to keep the same articulatory position without phonation for 15 seconds, i.e. during the acquisition. The image resolution is 256 × 232 with 120 slices. One example of this type of acquisition is shown in Fig. 2.13.

In the second part, audio was recorded simultaneously with the MRI acquisition. Duration of the acquisition was 7 seconds. The image resolution is 320 × 290 with 36 slices. Some data from 15 seconds acquisitions are presented on Fig. 2.14.

Sound recording

Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner with 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 denoise it, the denoising algorithm proposed in [OVB12] was used. This algorithm requires sufficiently long segments of signal without speech so that denoising will give good results.

Transcription of the continuous speech corpus

Text alignment was done with Astali [FMJ15], which can exploit an optional pronunciation dictionary if some words do not exist in the default lexicon. The transcription procedure is based on the guidelines described in [SMWC03].

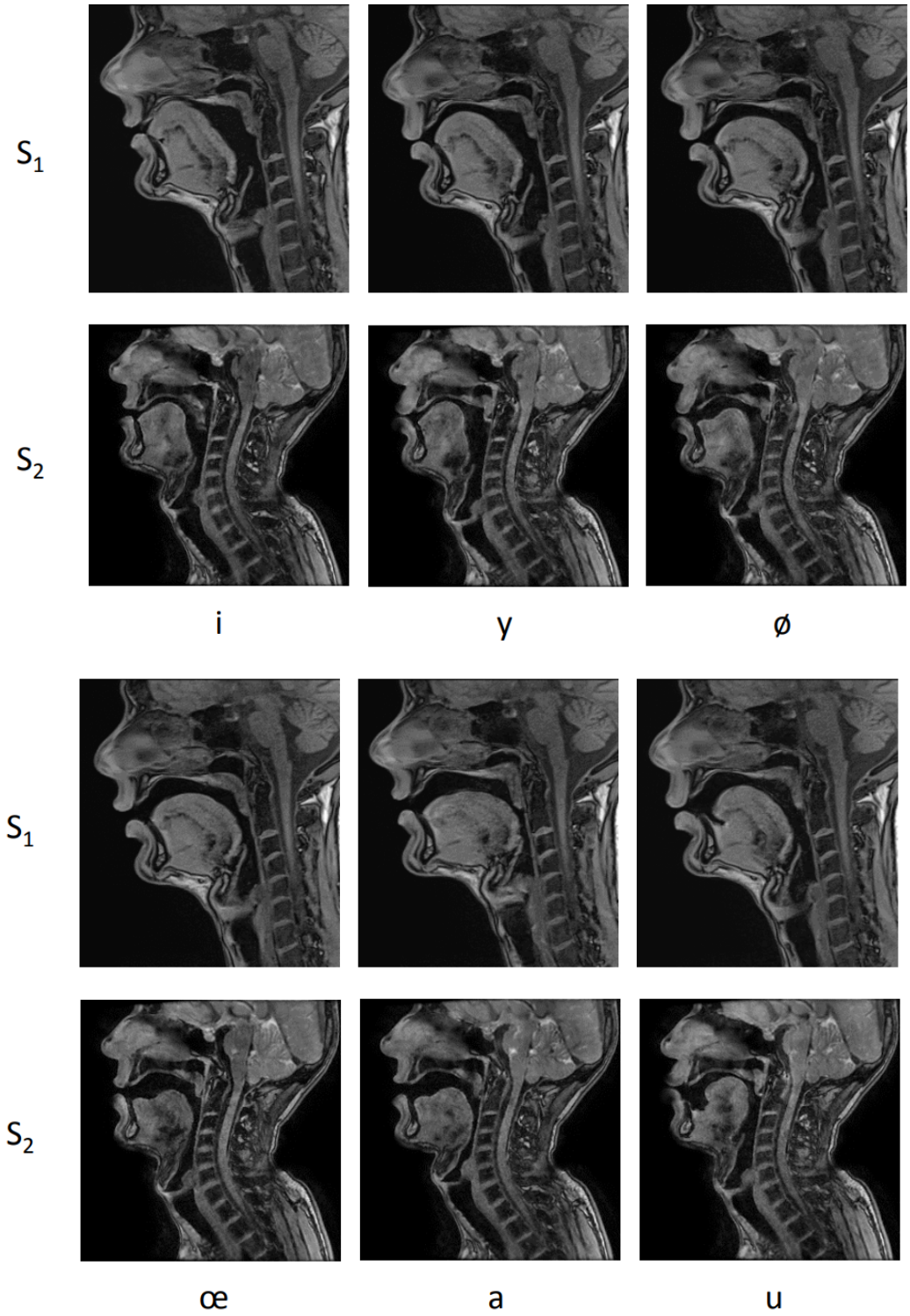


Figure 2.13: Mid-sagittal slice of the static 3D images of subjects two subjects (S_1 and S_2) for six oral French vowels.

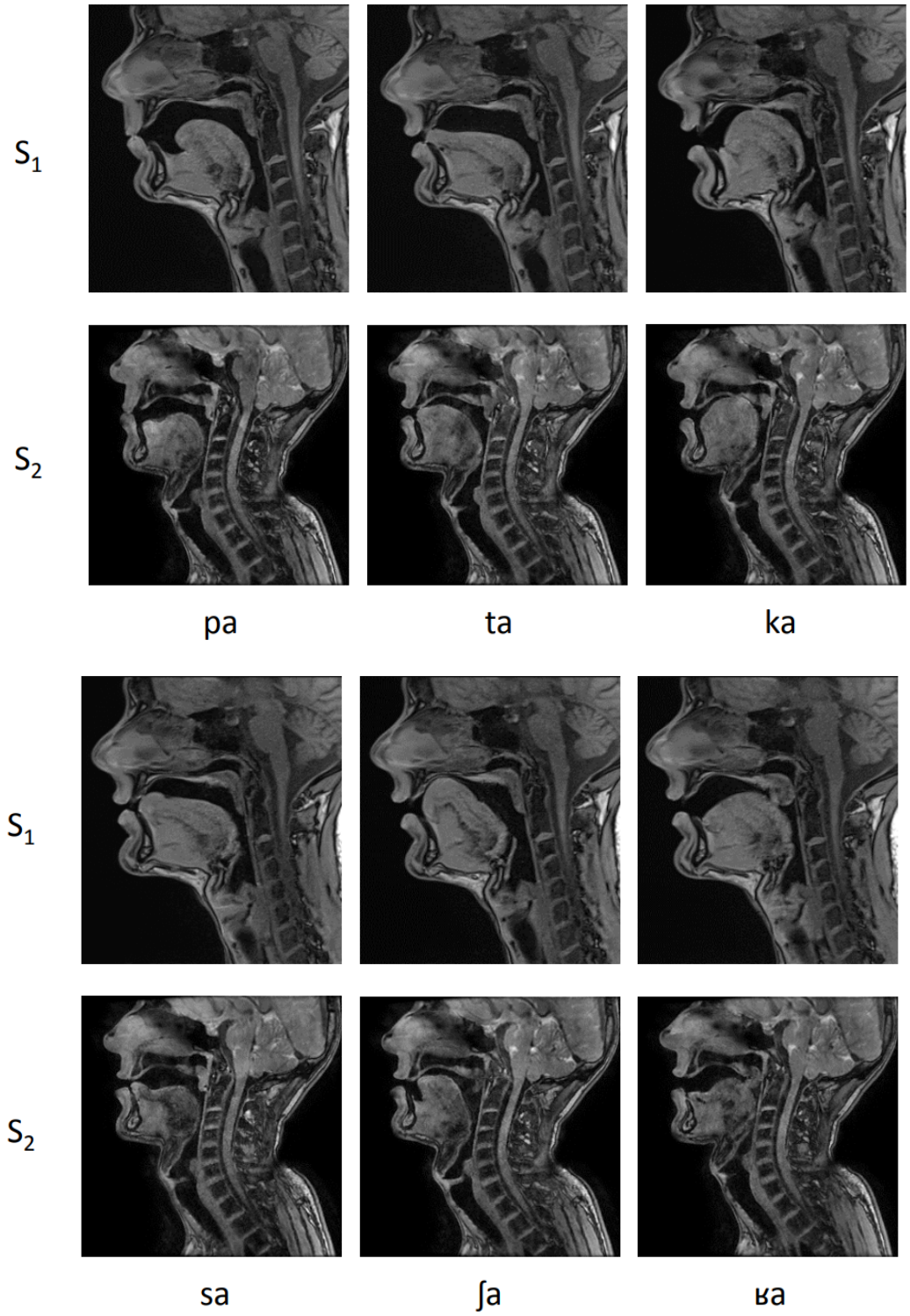


Figure 2.14: Mid-sagittal slices of the static 3D images of subjects S_1 and S_2 for some of the French consonants. Blocked articulation of the consonant within the context of a following vowel.

2.3.3 Database description

The objective of the database is to enable the exploration and modeling of coarticulation phenomena. It is thus necessary to get a good geometric description of the whole vocal tract and to get running speech which exhibits how the global geometry of the vocal tract evolves over time during speech production. So far, despite the presence of techniques for dynamic 3D MR acquisition [LZL+19], time and spatial resolution of such images is still quite low. The corpus construction strategy therefore consists of collecting a number of static configurations of the vocal tract corresponding to sustained vowels, or blocked CV articulations in 3D on the one hand and running speech in 2D (in the mid-sagittal plane) on the other hand.

The images must provide the greatest possible variability of articulatory shapes since these shapes are used to build an articulatory model. Besides, the teeth are not visible on the MRI images, and therefore it is necessary to merge these data with a numerical scan of the subject's dental cast. This requires additional MRI volumes to derive the position of teeth which are not visible. In this case, there were three: one by pressing the tongue against the upper teeth, especially in the area of the incisors, one by pressing the tongue against the lower teeth, and one with the lower and upper incisors in contact. The tongue is clearly visible on MR images, and pressing it against the teeth makes them appear in negative (Fig. 2.15).

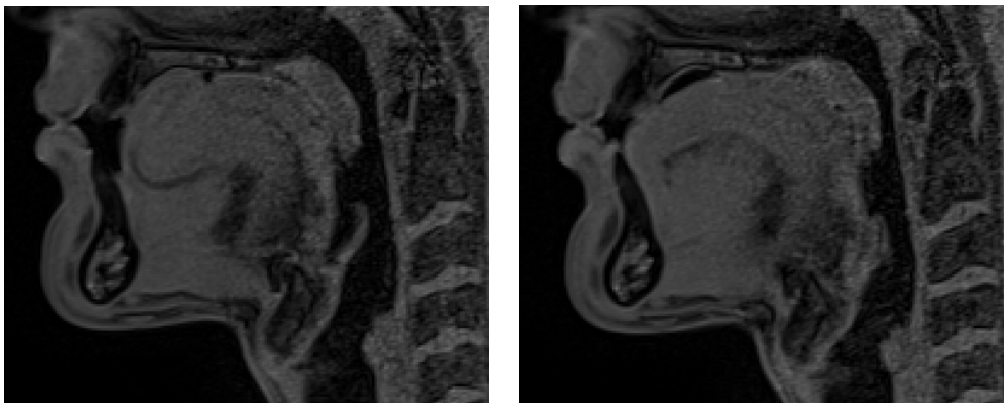


Figure 2.15: Examples of a subject pressing the tongue against the teeth. On the left the tongue is pressing against the upper teeth, on the right against the lower

Real-time 2D data

The analysis of coarticulation requires a good coverage of all phonetic contexts which can appear in French. Reading specifically prepared sentences meets this objective. For some specific issues, and to remove distant contextual effects, nonsense words (for instance a selected few CV or complex consonant clusters followed by a vowel) are better and are thus included in the database.

However, coarticulation is often more pronounced in the case of spontaneous speech which is less controlled and gives rise to stronger articulatory adjustments. Similarly, there is intra-speaker variability, and it is thus interesting to add several repetitions of

the sentences, or at least some of them. In previous work [LSVE14], X-ray films were used [Vax93] and so some of these sentences were added to compare both techniques of acquisition.

Each of these aspects gave rise to a part of the dynamic corpus of speech, which is described below.

In speech synthesis, the classical way of constructing a corpus consists of adding sentences from a vast written corpus, newspaper for instance, so as to enrich the linguistic coverage iteratively. Despite the efficiency of the construction algorithm each sentence contributes to a limited number of new phonetic contexts. To prevent a very long recording, the corpus design strategy consisted of constructing sentences by hand from a phonetized dictionary so as to add the expected phonetic contexts, i.e. those not present in the existing sentences. The dictionary is the phonetized version of the French Morphalou lexicon [lex] which provides 620.000 flexed forms [RSAF04].

Several levels of criteria were used in order to guide the manual construction of new sentences from words. After the insertion of each sentence the first level of criteria evaluated is the number of VV for all the vowels, the number of CV for C in /ptkfsʃlʁ/ and V in /i, a, u/ plus /y/, the number of VC with C as a coda and C in /l, ʁ, n, m/ and V in /i, a , u, y, e, ɛ, o, ɔ/, the consonant clusters C1C2V with C1 in /ptkbdgf/, C2 in /ʁl/ and V in /a, i, u, y/ (the other CCV following the same pattern with /sʃv are rare in French), and 15 complex consonant clusters (at least a sequence of 3 consonants, between two vowels).

This first level of criteria covers the very heart of the corpus in terms of mandatory phonetic contexts. Well constructed sentences of French were preferred, therefore words not corresponding to target's context were added. They provide new contexts, and in particular context with vowels outside the set of cardinal vowels plus /y/. VCV are considered by taking into account groupings of close vowels. There are 6 groups of vowels (/i,e/, /ɛ,a/, /u, o, ɔ/, /y, ø/, /oe,ə/ and nasal vowels /ã, õ, ɛ̃, œ̃/. This provides a second level of evaluation which guides the choice of words required to build well constructed sentences.

In total this corpus is made up of 77 sentences offering a very good coverage of all the phonetic contexts in French. Even if these sentences are sometimes a little bit curious they remain perfectly readable. There are only two non French words ("cartoons" and "squaw") but they can be easily pronounced by French speakers. One objective was to enable the comparison with an old X-ray database of 15 very short utterances (each repeated three times) [SHL+11]. In total those sentences represent 138 phonemes, and less than 30 seconds of reading, were included in the corpus. Below is presented the list of the 77 sentences.

1. Le filou et la fripouille manipulent de l'acrylique antirides dilué sous le tipi.

2. En haut du cumulus Pierre prit dix choux, du rouge et du clafoutis puis se camoufla en clown.

3. L'actionnaire des yaourts Caprice des Dieux couvrit le doigt sur le cahier aimanté.

4. Ne repoussez pas l'écrou de la galtouse de ris à la pomme.

5. Du coup l'oculiste tout fou dévissa sans scrupule le volant du véhicule.
6. La ciguë de l'homme de loi roux est dans la grue sur le parking.
7. Il n' a pas voulu, ou n'a pas pu injecter un sousmultiple de la dose en sous-cutané.
8. Je veux annuler pour éviter le raffut du saut dans le grand bassin.
9. Plus nous y croyons pire est le trou dans le lit du pauvre.
10. Trois sacs carrés. (à répéter 3 fois)
11. Vous dactylographiez sa soupe sirupeuse au lit.
12. Le chouan qui parle wolof, et a l'ouïe fine prépare une mixture bien pire.
13. Le chouchou du fou truqua le chargeur du fusil de leur nounou taciturne.
14. Sonne le glas à plat sans faire glouglou dans le foin et les plumes.
15. Le stupide toutou sous-nutri anticipa coucicouça l'africanisation des bikinis.
16. Les attabler. (à répéter 3 fois)
17. Il pouffa quand il ouvrit l'incunable qui montrait un prunus et les outils des Manouches.
18. Nous galopâmes avec peine jusqu'au bout sous le soleil.
19. Elle l'accuse de la diffamer en disant qu'elle a couru et s'est amusée avec du ciment et du sable humide.
20. J'exultais car elle joue et fume comme jamais avec les poules.
21. Crabes bagarreurs. (à répéter 3 fois)
22. Il a pourri. (à répéter 3 fois)
23. Nous analysions avec courroux l'humus du bois touffu, où tu voyais des bombes antichars et des coucous.
24. En écoutant la flûte, le chevreau mangea la robe à froufrous de Maurine.
25. Lui as-tu pris ta presse pour les piles du roi des Zoulous ?
26. Elle culbuta et accoucha huit fois dans les choux de Gilles.
27. Drapé dans son manteau mais pas du tout alourdi par le poids du chat il dut chuter sur la mosaïque.
28. Paul jugeait le Vésuve sans danger depuis le môle.
29. Il sut ça si tôt, qu'il fit tout pour diffuser les coupures à ras bord.

30. Des nuages gris et un cyclone destructeur s'approchent du groupe polaire.
31. Je vois le loubard, le wagon et des ficelles qui chutent dans la rivière.
32. Il l'a daté. (à répéter 3 fois)
33. A la cantine, un Druze cache ses frasques et ses vices, en fricotant avec un plouc.
34. Au bilan, les députés juxtaposeraient la souspoutre.
35. Pour tout casser. (à répéter 3 fois)
36. Amoindri par les tirs, le flibustier vadrouille à hue et à dia sans détour.
37. Finalement le loup du roi a vu la squaw redoutée des alouettes de Laval.
38. Elle moulut du pou chilien et du loup pour les enfants affamés du ru.
39. L'azimut chimique partira sans hachurer les sinus acquis avec humour pendant la pénurie.
40. L'aménageur qui est venu cherchait l'anthologie des appareils se réparant seuls.
41. L'ouvrage qui disposait d'une boussole était carbonisé de part en part.
42. Le premier des voyous ment très fort avant de souffler sur le nageur.
43. Il zappe pas mal. (à répéter 3 fois)
44. Comme alternative, j'ai agglutiné des tours de fil pour avoir un aimant supranaturel.
45. L'exclusivité fait peur à l'administrateur de biens du port.
46. Où irait-il en nu-pied dans cette cohue de grande taille, avec ces billes ?
47. Lustrage et pâturage riment un peu plus que fluor et météore.
48. Elle propose des activités de saut kilométrique en altitude à Soumatra.
49. Pis, p, paix, pas, port, peau pou, pu, peux, peur, pan, pont, pain.
50. Très acariâtre. (à répéter 3 fois)
51. La bise et le soleil se disputaient chacun assurant qu'il était le plus fort quand ils ont vu un voyageur.
52. Il disputait le voltigeur qui veut de chauds pantalons et des habits de mode sans plis ni goût.
53. Jouer du biniou électromagnétique ça fait bing contre le givre.
54. Blagues garanties. (à répéter 3 fois)

55. Quand la peur se répandit ils ont couru aux voitures enveloppées d'aluminium.

56. Il éblouit le veau et les pioupious qui sautaient à une encablure du Cher.

57. L'humanité uniquement hallucinée, et assoupie par la politique du sous ministre coula dans l'abîme.

58. Couds ta chemise. (à répéter 3 fois)

59. Pas de dates précises. (à répéter 3 fois)

60. Elle a tout faux. (à répéter 3 fois)

61. Chose inouïe il imita l'anti-roulis sans pâte à choux ni hachis.

62. Les scouts s'enivrent et papillonnent vers les cailloux où le wombat fait la loi sur sa mule.

63. Au milieu du lit où elle dessine des pions sur des cartoonz, le clou rouillé fait un tour.

64. Est ce un syllogisme de dire que l'homme pédant est un animal mortel.

65. Tout bouffi il dissout la moumoute à l'embouchure de la rivière moussue.

66. Le Chinois républicain Liou cacha les poissons et des agneaux dans la rue.

67. Paule prit les tamtam que la copine utilisera pour annoncer la panne.

68. Puis la structure de l'astragale va glisser doucement dans le ruisseau.

69. Le sextuple adjoint aux sports a un caillot au cerveau.

70. A l'île du saint, la crue du rio vert les submerge tous sans un cri.

71. Nous palissons. (à répéter 3 fois)

72. Infamie suprême, un fou encapuchonné fit mouche avec du gui à la proue.

73. Le truffage du choux nécessite du chiffon et du fil à rouler.

74. Frustré parce-que le cliché est flou, le paranoïaque va là où le climat est meilleur.

75. Avec du culot, la perruquiniste enrichie s'occupa du baby-foot du futur graphiste.

76. Il a pas mal. (à répéter 3 fois)

77. Des abat-jour. (à répéter 3 fois)

Static 3D data

The acquisitions were made for vowels and blocked /CV/ articulations. For vowels the subjects are instructed to phonate the vowel before the acquisition noise starts. They had to stop the phonation just before the acquisition starts and keep the same articulation during the acquisition. Asking subjects to phonate the vowels allows them to adjust the articulation. In order to get similar articulations between speakers they were asked to phonate the vowel in the context /pV/, with a very long vowel. /p/ was chosen because all the /pV/ correspond to a French word (except /pɔ/). For consonant articulations subjects were instructed to choose the articulatory position that would allow them to produce the expected /CV/. The accepted ones were /p,t,k,f,s,ʃ,l,ʁ,m,n/.

The ArtSpeechMRIfr covers:

- all French vowels /i,e,ɛ,y,ø,œ,ɔ,o,u,ɑ̃,ɔ̃,ɛ̃/ with a single acquisition for /œ̃, ɛ̃/ since the vast majority of French speakers no longer realizes the contrast between both vowels [Mad84]. Despite the precautions taken to ensure that the articulation of vowels inside an MRI machine is as close as possible to natural speech, subjects generally reduce the aperture, which is therefore underestimated in the models built from those images. For this reason the subjects were asked to record an extra vowel which is a "very open" /a/, i.e. similar to a vowel that would be articulated with a loud voice. Examples of the mid-sagittal slices for vowels are shown on Fig. 2.13.

- /p,t,k,f,s,ʃ,l,ʁ,m,n/ followed by vowels /i,a,u/ as a minimal set of CV. According to the subject, his judgment about his immobility (or the closeness with the target) which could require some acquisitions to be repeated, for vowels this minimal set can be extended. Extensions consist of adding other intermediate vowels. Examples of some consonant articulations are given on Fig. 2.14.

2.3.4 Applications

Applications of such a database can vary from the field of speech production and articulatory synthesis to speech recognition and speaker identification studies. Some of the potential uses of ArtSpeechMRIfr are presented below.

Articulatory modeling

The articulatory model is intended to generate the geometric shape of the vocal tract from a small number of parameters corresponding to the speech articulators. It is a key component of an articulatory synthesizer and the challenge is to design a model that can generate all the possible forms of the vocal tract that can appear during speech production. The proposed model takes into account the links between the articulators so as to find out the intrinsic deformation factors for each of the articulators by applying Principal Component Analysis (PCA) on articulatory contours extracted from static images [LB11]. This model has been improved several times, most recently for elongated articulators [LETV18], i.e. epiglottis and uvula. Usually, the input of PCA are the contours of the target articulator for all the static MRI images of one speaker. However, due to

delineation errors, direct application of PCA in this case leads to unrealistic PCA components, and especially irrelevant swelling deformation modes. The new model relies on the application of the PCA to the central line of these two articulators. The model improved in this way gives much better results on the epiglottis and uvula. This model built from the ArtSpeechMRIfr database's static images can be tested on the database's dynamic images. The contours of the articulators were carefully extracted semi-automatically or by hand, and corrected if necessary, for about 500 images, which allows the accuracy of the articulatory model to be assessed.

Comparison between static and dynamic data

The existence of static and dynamic data makes it possible to know to what extent 3D static data (of very good quality) can approach 2D dynamic data (of lower quality). This question arises as soon as an articulatory model is developed or when the objective is to reconstruct better quality dynamic images by exploiting static images. In [TDSL19] an automatic method is proposed to measure the differences between static and dynamic articulation. MR images of the vocal tract were used for this purpose by applying three image similarities measurements (EMD, SIFT, SSIM). It has been shown that there is a high intra-speaker variability and that there are some dynamic images, which are far from static images, that cannot be approximated with an articulatory model (Fig. 2.16). It is therefore necessary to include some dynamic images in addition to static images to improve the articulatory model.

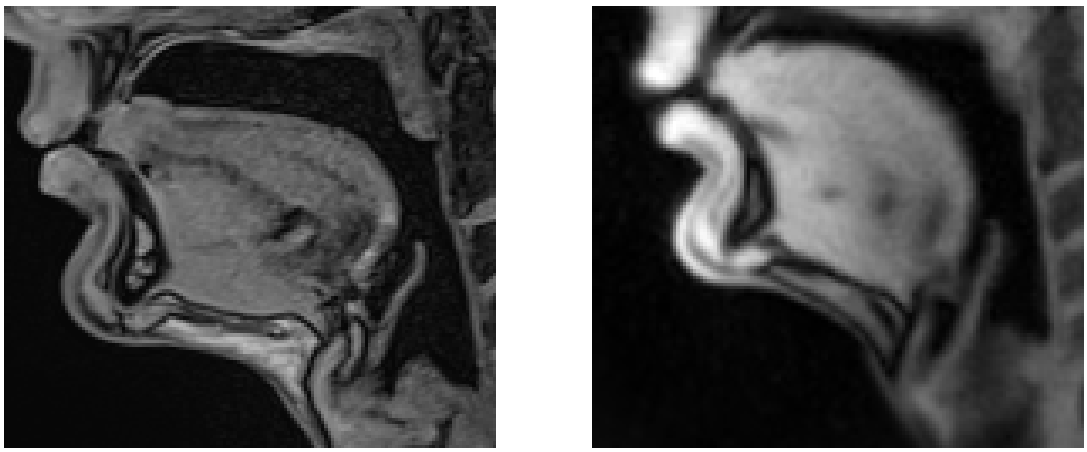


Figure 2.16: Static (left) and dynamic (right) recordings of /t(u)/ of the same speaker. Differences in articulation can be mainly seen in the region of the oral cavity.

Acoustic simulations

To carry out synthesis, several simplifications are generally made to keep a reasonable calculation time. The first consists in making the hypothesis of a plane wave propagating through the vocal tract which allows the analogy with an electrical transmission line to

be made. The second consists in moving from three-dimensional volumes to mid-sagittal slices. The existence in ArtSpeechMRIfr of 3D data together with the acoustic signal enables the impact of those simplifications on acoustics to be explored [DLVE19].

2.4 Conclusion of Databases

There are various MRI databases for speech production purposes for several languages. They include 3D static images of the vocal tract, dynamic images, different emotions, simultaneous audio recordings etc. However, there is limited work regarding the MRI data available for such studies in French language. The rich data that ArtSpeechMRIfr includes like high resolution 2D dynamic images, 3D static images, denoised sound recordings etc, as well as the rich phonetic context of both continuous speech and static articulation, makes this database a very good choice for someone to study speech production, or speech synthesis of French. Given the rich data and the great variety of potential applications of ArtSpeechMRIfr, this database will be used for Chapter 3 and 4 of the thesis. In the following Chapter, we are going to present the first part of our speech production study where we conducted acoustic simulations in order to examine the effect on phonation of several vocal tract geometry simplifications.

Chapter 3

Acoustic Simulations

Among the various types of speech synthesis like concatenative or deep learning, we believe that articulatory synthesis has the long term best potential because it allows the direct control of the speech production parameters and does not treat the whole speech production system as a black box. As in [SD01] we are convinced that *"In the long term, articulatory synthesis has more potential, not only for extending our knowledge of speech science, but for high-quality speech synthesis"*.

As stated in [EL16] articulatory synthesis is an invaluable tool to study speech production therefore the main inspiration for the work in this Chapter comes from some questions that appear mainly during the creation of statistical models for articulatory speech synthesis.

- First, is it necessary to use the 3D vocal tract shape in all cases or the midsagittal plane is sufficient [LET15]? This is important to know if the problem can be addressed in the 2D domain only.

- Second, how the head position of the speaker during data acquisition can affect phonation and how articulatory models can adapt to such changes [BBRS98]? This is important because there is variability between several MRI recording sessions in particular.

- Third, is it possible to simplify the vocal tract geometry in order to reduce the parameters of the model [LLM$^+$13b]? Or equivalently, is it necessary to take epiglottis and small cavities into account?

To try to find out the answer to these questions we used MRI data of the vocal tract properly processed and modified in order to conduct our experiments using acoustic simulations. Part of the work presented it this Chapter in Sections 3.1,3.2,3.3 have already been object of communication in several conferences (Paper 1 - Paper 4) as mentioned in the introduction of this thesis in Chapter 1 Section 1.2.

3.1 Comparison between various types of simulations

3.1.1 Introduction about acoustic simulations

Unlike other approaches, which only model the result of speech production, i.e. the acoustic speech signal, articulatory synthesis [TEL17, Bir13, LLM+13a] explicitly models the link between the vocal tract, vocal folds and aero-acoustic phenomena. This is achieved by solving the equations of aerodynamics and acoustics in the vocal tract and by using its geometry as an input.

In [TMK10] 3D MRI of the vocal tract was used to extract the volume of the vocal tract during the production of five Japanese vowels. Physical models of the vocal tract were constructed using 3D printing technology and area functions were calculated. Results were compared with those from acoustic simulations and the original audio signal. In [SHT96] the area function of the vocal tract was estimated from MRI images and then the results were compared with the sound acquired during a different session. Even though these authors tried to make the audio recording condition as close to the original MRI as possible [KWMPM00], there could still be significant difference between the recorded audio signal and the signal pronounced during MRI acquisition, mainly due to the difference in the auditory feedback for the speaker during the noisy MRI recordings and the quiet condition.

One of the challenges of statistical articulatory modeling consists of collecting and processing data to construct the model. As the delineation of the articulators in the MRI images is a task that requires a certain amount of interpretation of the geometry of the vocal tract and has a very long processing time, it is often preferable to construct a two-dimensional model in the mid-sagittal plane and then calculate the transverse area at each point of the vocal tract from the glottis to the lips [Eri07].

One question that arises is to evaluate the acoustic impact of the approximations corresponding to the transformation from the true three-dimensional shape to the two-dimensional shape completed by the calculation of the transverse area.

In this study, our purpose is to examine to what extent the 2-dimensional data can describe articulatory information, compared to 3-dimensional data, since it is much more efficient to use with a 2-dimensional model [Eri07]. We also examine how acoustics is affected in each case by comparing the sound signal against the results from 2 and 3-dimensional acoustic simulations, and 2-dimensional electrical simulations on the other hand.

3.1.2 Data acquisition

For the purposes of our experiments, we used five of the vowels from the database described in Chapter 2 Section 2.3, /a, oe, i, o, u/. We also used some additional MRI data part of a study approved by an ethics committee and the subject gave written informed consent (ClinicalTrials.gov identifier: NCT02887053). The subject retained for the data acquisition is a healthy male French native speaker at the age of 57, without any reported speaking or hearing problems.

The additional MRI data was acquired on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) with gradient of $80mT/m$ amplitude and $200mT/m/ms$ slew rate.

We used the 3-dimensional cartesian gradient echo RF spoiled vibe sequence ($TR = 3.8$ ms, $TE = 1.41$, $FOV = 235 \times 260$ mm, $flip\ angle = 7.5$ $degrees$) for the acquisition. The pixel bandwidth is $445Hz/pixel$ with an image resolution of 320×290. Scan slice thickness is 1.2 mm and the number of slices is 36. The pixel spacing is 0.8125 and the acceleration factor is 3 $iPAT$.

The subject's vocal tract is imaged while he lay supine in the MRI scanner (Fig. 1.1, left image) with his head in three positions: up, middle/normal and down. Additional foams were used to stabilize the head in each position and help the subject reach and maintain the two extreme. Between the different head positions, the foam position was re-initialized to ensure the maximum possible lengthening and shortening of the vocal tract. However, there were limitations to how far the subject could tilt his head due to the coil.

The recording time for the subject, including initializations, calibrations and pauses between head positions and phonemes, was 2 hours.

Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner using FOMRI III (Optoacoustics, Or Yehuda, Israel) fiber optic microphone. The subject pronounces each vowel to be recorded twice before the MRI acquisition starts and once as the MRI machine is on. The latter repetition takes around 7.4 s of sustained phonation.

The subject wears ear plugs for protection from the scanner noise, but is still able to communicate orally with the experimenters via an in-scanner intercom system.

Since the sound is recorded at the same session of the MRI acquisition, there is additional noise in the audio signal. Details on how we treated this issue are described in 3.1.6.

3.1.3 Data processing

For the purpose of this work we used the 3D static MRI data of the vocal tract for /a, oe, i, o , u/ of a male subject and simultaneous audio recordings from ArtSpeechMRIfr database.

For the purposes of our experiments, we used the ITK-SNAP software [YPH$^+$06] to segment the volume of the vocal tract. ITK-SNAP provides a great variety of tools for segmenting images, both automatically and manually.

As far as automatic segmentation is concerned, ITK-SNAP implements two active contour segmentation algorithms, region competition and geodesic active contours [CKS97], [ZY96].

For manual segmentation, ITK-SNAP offers two types of tools, the most interesting is the adaptive brush that adjusts itself to follow the image boundaries. The brush tool can be used for both to $2D$ and $3D$ image segmentation.

3.1.4 Acoustic simulations

For the acoustic simulations, we employ the k-wave Matlab toolbox [TC10a]. This toolbox has a wide range of applications like photoacoustic tomography ultrasound wave propagation [WT13], and acoustic propagation. [TC10b].

Several numerical methods have been developed to solve the partial differential equations of acoustics, like finite differences, finite elements, and boundary element methods [TMK10]. These methods offer significant advantages as they can calculate acoustic characteristics accurately and implement frequency dependent losses at boundaries. However, in many cases these methods are significantly slow. This is due to the fact that they require a small time step to achieve adequate accuracy and a lot of grid points per wave length. In the method used by k-wave these problems are solved by interpolating a Fourier series through all of the grid points in order to get the estimation of the signal propagation. This approach solves the problems faced by the previously referred methods as it a) requires fewer grid points (only two) per wave length since the base function of the Fourier series is a sinusoid and b) it can be fast since it employs FFT (Fast Fourier Transform) to calculate the amplitudes of the simulated signals. A problem that arises is that when a wave approaches the computational grid boundaries, it keeps propagating to the medium by entering from the opposite site of the computational grid. This happens because of the usage of the FFT algorithm for the computation. To tackle this issue, k-wave adds a specific type of layer to the boundaries of the computational grid by implementing an absorbing boundary condition, called PML (Perfect Match Layer), which prevents this phenomenon. Finally, kwave toolbox has a great number of parameters that can be customised for a simulation, most of them concerning the grid and time sparsity, the properties of the mediums, the sensors, the sources, the number of dimensions ($1D/2D/3D$), the number of PML, etc.

Instead of using the second order wave equation

$$\nabla^2 p - \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{3.1}$$

k-wave simulation code is based on the equivalent first order equations.

$$\frac{\partial u}{\partial t} = -\frac{1}{\rho_0}\nabla p \tag{3.2}$$

$$\frac{\partial \rho}{\partial t} = -\rho_0 \nabla u \tag{3.3}$$

$$p = c_0^2 \rho \tag{3.4}$$

as it provides more flexibility for modeling some quantities like acoustic intensity (u is the acoustic particle velocity, p is the acoustic pressure, ρ is the acoustic density, ρ_0 is medium density, and c_0 is the sound speed). Equations 3.2-3.4 describe the simple case of a wave propagating through a homogeneous medium. In practice however it is common to have heterogeneous medium which results in loses of some acoustic energy due to acoustic absorption. Sound absorption can be described by the frequency power law of

$$\alpha = \alpha_0 \omega^y \tag{3.5}$$

(α is the absorption coefficient, α_0 is the power law factor and y is the power law exponent). Therefore in a heterogeneous medium equations 3.2-3.4 are transformed to

$$\frac{\partial u}{\partial t} = -\frac{1}{\rho_0}\nabla p \tag{3.6}$$

$$\frac{\partial \rho}{\partial t} = -\rho_0 \nabla u - u\nabla\rho_0 \tag{3.7}$$

$$p = c_0^2(\rho + d\nabla\rho_0 - L\rho) \tag{3.8}$$

with

$$L = \tau\frac{\partial}{\partial t}(-\nabla^2)^{\frac{y}{2}-1}, \tau = -2\alpha_0 c_0^{y-1} \tag{3.9}$$

(d is the acoustic particle displacement). Since a grid with elements is used to describe the medium, equations should be discrete. In k-wave the discrete version of the equations 3.6-3.8 used are

$$\frac{\partial}{\partial\xi}p^n = \mathcal{F}^{-1}\{ik_\xi\kappa e^{ik_\xi\Delta\xi/2}\mathcal{F}\{p^n\}\} \tag{3.10}$$

$$u_\xi^{n+\frac{1}{2}} = u_\xi^{n-\frac{1}{2}} - \frac{\Delta t}{\rho_0}\frac{\partial}{\partial\xi}p^n + \Delta t S_{F_\xi}^n \tag{3.11}$$

$$\frac{\partial}{\partial\xi}u_\xi^{n+\frac{1}{2}} = \mathcal{F}^{-1}\{ik_\xi\kappa e^{ik_\xi\Delta\xi/2}\mathcal{F}\{u_\xi^{n+\frac{1}{2}}\}\} \tag{3.12}$$

$$\rho_\xi^{n+1} = \rho_\xi^n - \Delta t\rho_0\frac{\partial}{\partial\xi}u_\xi^{n+\frac{1}{2}} + \Delta t S_{M_\xi}^{n+\frac{1}{2}} \tag{3.13}$$

(ξ is the space direction (x,y,z), $\mathcal{F}$ and $\mathcal{F}^{-1}$ is the forward and inverse Fourier transform, i is the imaginary unit, k_ξ is the wavenumbers in direction ξ, $\Delta\xi$ is the grid spacing, Δt is the time step, $\kappa = sinc(c_{ref}k\Delta t/2)$, c_{ref} is the reference sound speed, S_M is the source mass and S_F is the source force). In order to include the PML into the simulations, equations 3.11 and 3.13 are transformed to

$$u_\xi^{n+\frac{1}{2}} = e^{-\alpha_\xi\Delta t/2}(e^{-\alpha_\xi\Delta t/2}u_\xi^{n-\frac{1}{2}} - \frac{\Delta t}{\rho_0}\frac{\partial}{\partial\xi}p^n) \tag{3.14}$$

$$\rho_\xi^{n+1} = e^{-\alpha_\xi\Delta t/2}(e^{-\alpha_\xi\Delta t/2}\rho_\xi^n - \Delta t\rho_0\frac{\partial}{\partial\xi}u_\xi^{n+\frac{1}{2}}) \tag{3.15}$$

$$\alpha_\xi = \alpha_{max}(\frac{\xi-\xi_0}{\xi_{max}-\xi_0})^4 \tag{3.16}$$

which is the form that PML equations are implemented in k-Wave (α_ξ is the anisotropic absorption, α_{max} its maximum value, ξ_0 and ξ_{max} are the coordinates at the beginning and the end of the PML). A simple example of simulation code can be seen in Fig. 3.1. Further details can be found in [TCJ12].

```
% define the computational grid
Nx = 2^6; % number of grid points in the x (row) direction
Ny = 2^6;% number of grid points in the y (column) direction
dx = 0.001; % grid point spacing in the x direction [m]
dy = 0.001; % grid point spacing in the y direction [m]
kgrid = kWaveGrid(Nx, dx, Ny, dy); %create grid

% create the time array
sim_time=1e-4; %duration of simulation
sim_step=3e-08; %simulation time step
kgrid.t_array = 0:sim_step:sim_time; %time points of simulation

% define the medium properties
mss=1440; %medium sound speed [m/s]
md=900; %medium density [kg/m^3]
ssr=350; %sound speed reference [m/s]
mdr=1; %medium density reference [kg/m^3]

%apply medium properties
medium.sound_speed=mss*ones(Nx,Ny);
medium.density=md*ones(Nx,Ny);
medium.sound_speed(1:Nx/2,:)=ssr;
medium.density(1:Nx/2,:)=mdr;

%source properties
disc_x_pos =Nx/2; %x-axis position
disc_y_pos = Ny/2; %y-axis position
disc_radius = 5; %source radius
disc_mag = 1; %source magnitude [Pa]
source.p0 = disc_mag*makeDisc(Nx, Ny, disc_x_pos, disc_y_pos, disc_radius); %create source

%create sensor
sensor.mask=[kgrid.x_vec(Nx-11);kgrid.y_vec(Ny/2+20)];

% run the simulation
sensor_data = kspaceFirstOrder2D(kgrid, medium, source, sensor,'DataCast','single');
```

Figure 3.1: Simple example code of acoustic simulation using k-wave toolbox

3.1.5 Electrical simulation

To perform the electrical simulation we used some of the tools provided from the Xarticul software [LSVE14], [SHL+11]. Xarticul offers multiple tools, like an easy way to delineate and process articulator contours, semi-automatic articulatory measurements and construction of articulatory models [LB11]. Xarticul can perform acoustic simulations from the area function by using the algorithm based on the TLCA (Transmission Line Circuit Analog) method [Mae82]. The main idea of the algorithm is to model every tube used to describe the vocal tract as a circuit of electrical units whose parameters (electrical) correspond to the physical (acoustic). Fig. 3.2 shows the electric-acoustic parameter analogy.

For example, the current and the voltage of the circuit in the TLCA corresponds to the volume velocity and acoustic pressure respectively. Therefore, instead of describing the vocal tract as a continuous connection of tubes, one can describe it as a continuous

Electric	Acoustic
Current	Volume velocity u
Voltage	Acoustic pressure p
R_i	Energy loss ($R_i = \frac{4\pi\mu l_i}{a_i}$)
C_i	Air compliance ($C_i = \frac{a_i l_i}{(\rho c_s)^2}$)
L_i	Air inertance ($L_i = \frac{\rho l_i}{2a_i}$)
R_{w_i}	Wall resistance ($R_{w_i} = \frac{W_R}{2l_i\sqrt{\pi a_i}}$)
C_{w_i}	Wall compliance ($C_{w_i} = \frac{2l_i\sqrt{\pi a_i}}{W_C}$)
L_{w_i}	Wall inertance ($L_{w_i} = \frac{W_L}{2l_i\sqrt{\pi a_i}}$)
U_{d_i}	Flow source ($-\frac{\partial}{\partial t} l_i a_i$)
P_{n_i}	Frication noise source ($P_{n_i} = \max\{0, \xi w \frac{U_{DC}^3}{a_{i-1}^{3/2}} (Re^2 - Re_c^2)\}$)
R_{n_i}	Internal resistance of noise source ($R_{n_i} = \kappa\rho\frac{U_{DC}}{a_{i-1}^2} + 8\pi\mu\frac{l_{i-1}}{a_{i-1}^2}$)

Figure 3.2: Parameter equivalence for electric-acoustic analogy. Image from [EL16]

connection of electric circuits. As a result, the problem of the acoustic wave propagation within a tube described by the equations,

$$\frac{\partial u}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{3.17}$$

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A}\frac{\partial u}{\partial t} \tag{3.18}$$

(u is velocity, p is pressure, A is cross-sectional area of tube, c is speed of sound in medium, ρ is medium density) is transformed to the equivalent problem of electrical wave propagation through an electric circuit described by the equations

$$\frac{\partial i}{\partial x} = -C\frac{\partial v}{\partial t} \tag{3.19}$$

$$\frac{\partial v}{\partial x} = -L\frac{\partial i}{\partial t} \tag{3.20}$$

(i is the current, v is the voltage, C is capacitance, L is inductance) The main advantage of this approach is that it allows time-varying geometries of the vocal tract [EL16] to be modeled.

3.1.6 Experiments

Our experiments can be divided into three main stages: a) image segmentation, b) acoustic simulations and c) electric simulations.

Image segmentation

For the purposes of our experiments, we used five of the vowels from the database described previously (Chapter 2, Section 2.3), /a, œ, i, o, y/. First, we processed the images with 3DSlicer [FBKC+12] (http://www.slicer.org) to apply "lanczos" interpolation in order to make corrections to the image's axis. Then, we used the tools provided by the ITK-SNAP software to automatically segment the 3D volume of the vocal tract and manually corrected the result. The area of interest begins at the glottis and extends up to the lips at the point where the lips stop being simultaneously visible in the coronal plane. We used two classes and 10000 points as nearest neighbours in order to assign each point to the appropriate class for the creation of the probabilistic map, and 10 balls on average per vowel as "seeds" with various sizes based on the region of the vocal tract where they were placed. Active contour algorithm was applied, using extreme narrow banding method [Whi98] which required between $300 - 500$ iterations to cover the whole vocal tract. The amount of iterations is greatly based on the vowel and the initial number, size and position of the "seeds". For the manual segmentation we used the adaptive brush tool with the default parameters in order to manually correct the vocal tract mesh (Fig. 3.3). Finally we used meshlab [CCC+08] to smooth every mesh by applying Laplacian smoothing filter with step 3. For each vowel, about 4 hours of processing was required, with the biggest amount of time spent on the manual segmentation step.

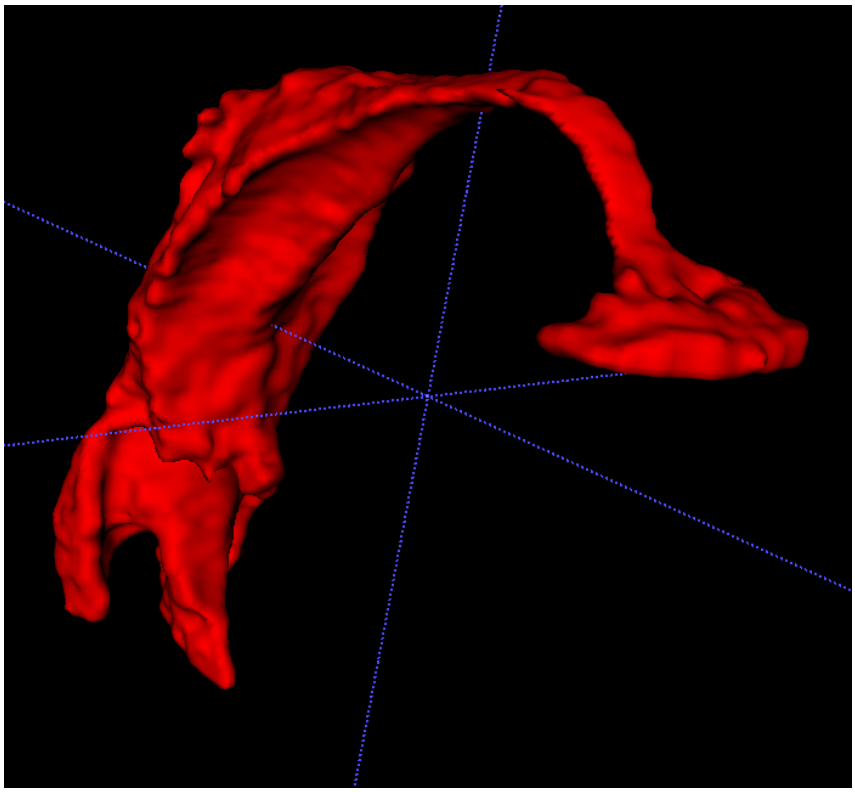


Figure 3.3: 3D volume of /i/ vowel

Acoustic simulations

For the acoustic simulations, we used k-wave toolbox for Matlab [TC10a]. For every vowel examined, simulations were carried out in both $2D$ and $3D$. First, the mesh was transformed into a volumetric representation using voxels. Then we specified the parameters for the $2D$ and $3D$ simulations. Since k-wave uses FFT, the number of grid points was set so as to have low prime factors, ideally a power of 2. For the $3D$, we used a grid size of $128 \times 128 \times 128$ grid points ($sagittal \times coronal \times axial$) with $d_x = d_y = d_z = 1mm$. We also used a PML layer of 10 grid points at the boundaries of every side of the grid, to avoid the wave penetrating the opposite side. As a source we used a ball which emits a *delta* pulse of pressure, spreading equally in all directions. The source has radius of 5 grid points, amplitude 1 Pa and was placed at the input of the vocal tract (Fig. 3.4), which was specified manually for every vowel. To record the simulated pressure we used a sensor placed at the end of the vocal tract. The medium properties inside the vocal tract were $c_{in} = 350m/s, d_{in} = 1kg/m^3$ and the properties outside, i.e. in the tissues that delimit the vocal tract, were $c_{out} = 1440m/s, d_{out} = 900kg/m^3$, where c_{in}, c_{out} are the speed, and d_{in}, d_{out} are the densities inside and outside the vocal tract respectively. These parameters were adapted to our experiments based on the values from [WL05, RWCL10]. The time step is set according to the two medium characteristics (here tissues and air) and the accepted value is $3 * 10^{-8} sec$ to guarantee a good stability. The amount of time steps computed was 1000001. The maximum allowed frequency of the grid was $175KHz$. For the $2D$ case, we run the simulations on the $y - z$ plane using a disc instead of a ball on the mid-sagittal plane of the vocal tract. All the other parameters remained the same between the two simulations. An example of the simulation code can be seen in Fig. 3.5. Finally, we calculated the transfer function of every vocal tract and computed their peak frequencies (Table 3.1), to compare them with the formants computed with the electrical simulation.

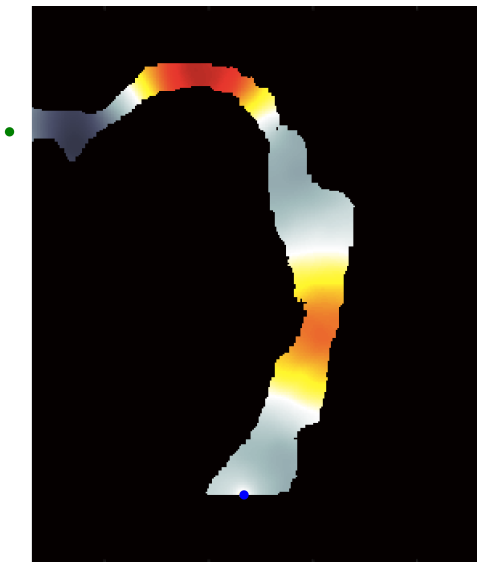


Figure 3.4: Simulation example of acoustic propagation. Blue dot at the vocal folds is the source, green dot on the left of the lips is the sensor

```
% create the computational grid
Lx=0.064; % grid length
Ly=0.128;
Lz=0.128;
Nx = 2^6; % number of grid points
Ny = 2^7;
Nz = 2^7;
dx = Lx/Nx; % grid point spacing
dy = Ly/Ny;
dz = Lz/Nz;
kgrid = kWaveGrid(Nx, dx, Ny, dy, Nz, dz);

% create the grid and the time array
sim_time=30e-3;
sim_step=3e-08;
kgrid.t_array = 0:sim_step:sim_time;

% input vocal tract geometry
Vox_x=60;
Vox_y=80;
Vox_z=80;
[i_grid]=VOXELISE(Vox_x,Vox_y,Vox_z,'VT.stl','xyz');
i_grid_d=ones(Vox_x,Vox_y,Vox_z).*i_grid;

% create vocal tract geometry mask
mask=zeros(Nx,Ny,Nz);
mask((Nx-Vox_x)/2:(Nx-Vox_x)/2-1+Vox_x,(Ny-Vox_y)/2:(Ny-Vox_y)/2-1+Vox_y,(Nz-Vox_z)/2:(Nz-Vox_z)/2-1+Vox_z)=1;
mask((Nx-Vox_x)/2:(Nx-Vox_x)/2-1+Vox_x,(Ny-Vox_y)/2:(Ny-Vox_y)/2-1+Vox_y,(Nz-Vox_z)/2:(Nz-Vox_z)/2-1+Vox_z)=...
mask((Nx-Vox_x)/2:(Nx-Vox_x)/2-1+Vox_x,(Ny-Vox_y)/2:(Ny-Vox_y)/2-1+Vox_y,(Nz-Vox_z)/2:(Nz-Vox_z)/2-1+Vox_z).*i_grid_d;
patch_tr=zeros(Nx,Ny,Nz);
for i=1:Nx
    patch_temp(:,:)=mask(i,:,:);
    patch_tr(i,:,:)=patch_temp(:,:);
end

% define grid properties
mss=1440; % sound speed
md=900; % density
ssr=350; % reference sound speed
mdr=1; % reference density
medium.sound_speed=mss*ones(Nx,Ny,Nz);
medium.density=md*ones(Nx,Ny,Nz);
patch_tr(:,103:Ny,1:Nz)=1;
patch_tr(:,Ny/2:end,1:Nz/2)=1;
medium.sound_speed=medium.sound_speed+(ssr-mss)*patch_tr;
medium.density=medium.density+(mdr-md)*patch_tr;

% define source
disc_x_pos = Nx/2; % source position
disc_y_pos =30;
disc_z_pos = 28;
disc_radius = 5;
disc_mag = 1;
source.p0 = disc_mag*makeBall(Nx, Ny, Nz, disc_x_pos, disc_y_pos, disc_z_pos, disc_radius);

% define sensor
sensor.mask=[kgrid.x_vec(Nx/2);kgrid.y_vec(103);kgrid.z_vec(85)];

% run the simulation
sensor_data = kspaceFirstOrder3D(kgrid, medium, source, sensor,'DisplayMask', Wall);
```

Figure 3.5: Example code for acoustic simulation using k-wave toolbox

	$F1$	$F2$	$F3$
/a/	613 / 796	1164 / 1226	2758 / 2329
/œ/	429 / 429	1532 / 1348	2635 / 2451
/i/	337 / 304	2394 / 2207	3136 / 3237
/o/	444 / 326	999 / 782	2219 / 2475
/u/	350 / 347	1399 / 1389	2098 / 2083

Table 3.1: 2D / 3D formants computation from acoustic simulations in Hertz

Electrical simulations

For the electrical simulations we used the mid-sagittal planes from the $3D$ MRI acquisition. We used Xarticul to manually delineate the articulator contours of each vowel (Fig. 3.6 left). Afterwards, we used 40 tubes (Fig. 3.6 right) to estimate the area function of the vocal tract in order to compute its formants (Table 3.2). Finally, we computed the formants of the original audio signal based on the method described in 3.1.6 (Table 3.3). We also made a comparison with the values given in the literature for French [Lon84] (Table 3.3).

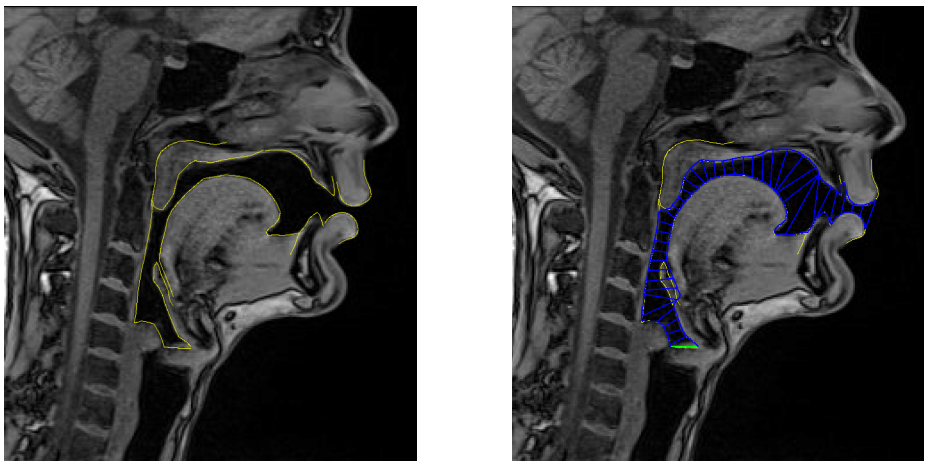


Figure 3.6: Delineation of vowel /o/ using Xarticul (left), separation of the vocal tract into acoustic tubes for /o/ (right)

	$F1$	$F2$	$F3$
/a/	510	1200	2190
/œ/	408	1276	2168
/i/	280	1684	2927
/o/	491	905	2185
/u/	393	1181	2000

Table 3.2: 2D formants computation from electrical simulations in Hertz

	$F1$	$F2$	$F3$
/a/	684 / 689	1256 / 1256	2503 / 2604
/œ/	517 / 443	1391 / 1335	2379 / 2436
/i/	308 / 380	2064 / 2306	2976 / 3193
/o/	383 / 430	793 / 732	2283 / 2619
/u/	315 / 393	764 / 815	2027 / 2185

Table 3.3: Theoretical/measured values of French vowels formants in Hertz

Formant estimation

In the case of a standard speech signal (recorded in a fairly quiet room) formants can be extracted by applying standard algorithms, e.g. LPC (Linear Prediction Coding) which is used in Praat [BW01]. For speech recorded in an MRI machine, the situation is quite different. First, the amplitude of the signal becomes much higher when the machine starts acquiring images. Since the signal must not be clipped even when the noise machine is intense, the recording level is low, and consequently the signal is poorly defined. Second, the transfer function of the optical microphone (for instance in the case of our FOMRI III (Optoacoustics, Or Yehuda, Israel)) attenuates the energy at low frequencies, and consequently the energy of the first formant is always lower than expected. This is important because LPC cannot be used anymore since F1 is often too weak to be detected. We therefore resorted to an algorithm derived from the standard linear cepstral smoothing called "true envelope" [IAse, RR05]. The advantage of cepstral smoothing is that it does not impose the implicit assumption of an all-pole model. Compared to linear cepstral smoothing, true envelope algorithm provides the additional advantage of approximating harmonics instead of smoothing the spectrum.

Fig. 3.7 shows the narrow band spectrum (curve with harmonics) and the true envelope spectrum (smooth curve), obtained with Winsnoori [Lap99].

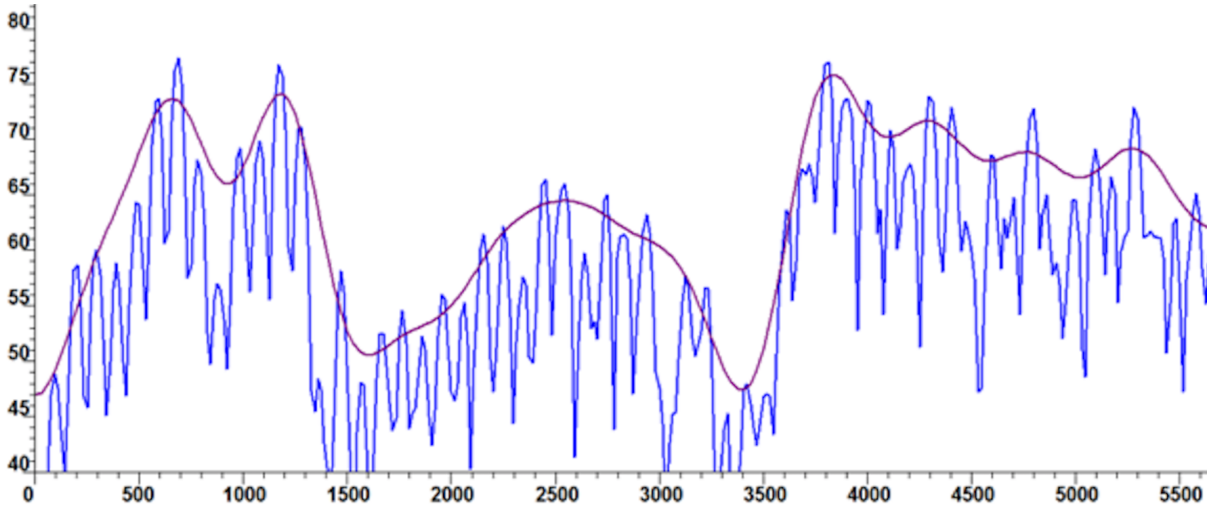


Figure 3.7: Narrow band spectrum (curve with harmonics) and the true envelope spectrum (smooth curve) of /a/ (x-axis in Hz, y-axis in dB)

Once the MRI acquisition starts the MRI noise takes over speech. Denoising this speech offers a much better speech quality perception, but at the same time degrades the intrinsic acoustic properties of speech. In particular, the detection of spectral peaks corresponding to formants becomes chaotic, and we had great difficulty in determining the formants in the denoised speech. We thus visually checked that harmonics of speech still present were compatible with the formants of the vowel just before acquisition (although they were strongly dominated by the MRI machine noise).

3.1.7 Discussion about various types of simulations

The first remark concerns the sounds produced by the speaker in the MRI machine. As shown in Table 3.3 which gives the average values for French speakers, the measured values are a little far from the expected values. This is especially true for the first formant of close vowels /i,o/ which is higher in frequency. One explanation could be that the pharyngeal cavity is smaller due to the subject posture. The visual examination of the images shows a slightly shifted articulation in some cases. However it could also be due to the difficulty of measuring formants in the denoised signal, especially when F1 is small. Second there is a good agreement between the results of 2D/3D simulations and the formants F1 and F2 determined from the speech signal recorded. However, for F3 the 3D simulation turns out to give results closer to those of natural speech than those of the 2D simulation, probably because the 3D volume gives a geometry closer to the real one.

The third remark concerns the comparison between the acoustic and electrical simulations. It turns out that the electric simulation is not as good as the acoustic simulation to reproduce the formants. Since there is a good agreement between the 2D and 3D acoustic simulation, the most probable hypothesis is that either splitting of the vocal tract into small tubes or the estimation of the area function from the mid-sagittal shape is not completely satisfactory.

Future direction for research will focus these points so as to improve the quality of articulatory synthesis.

3.2 Impact of head position on phonation

3.2.1 Introduction about the effect of head position on phonation

The geometric shape of the vocal tract can be derived from images of a film or be generated at each time point by an articulatory model [Mae90, BBB01, BJK06, LB11]. Apart from models based on geometric primitives, the models are generally constructed using factor analysis [BBB01] applied to a corpus of two or three-dimensional MRI images of the vocal tract. One of the issues raised by articulatory models derived from medical images of one subject is the validity of the model for other speakers.

Maeda [Mae90] developed a procedure that consists of separately adapting the sizes of the mouth and pharynx. This distinction between the two parts of the vocal tract is based on the observation that the size of the pharyngeal and mouth cavities depends on both gender and, predictably, age. A slightly more elaborate approach adapted to a more

complex articulatory model has been developed in [SHL+11] and tested with a model developed on one speaker which was used to fit mid-sagittal vocal tract shapes of another speaker.

In the first works dedicated to articulatory modeling carried out with X-ray images, speakers were sitting and adopting a fairly natural position to produce speech. More recent articulatory data are acquired with MRI in a supine position, and the head posture is largely dictated by the position of the MRI antenna and foam, which is used to prevent it from moving during acquisitions. Consequently, the position of the head is not natural, and above all it can vary significantly between two acquisitions, and a fortiori between two machines. The articulatory models that can be derived from those data implicitly incorporate the head posture. Experiments carried out to fit dynamic MRI data of one speaker with an articulatory model built for a reference speaker have shown that the adaptation procedure fails to approximate the whole vocal tract. More precisely, it turned out that the tongue can be fitted fairly well, which is not the case for the pharyngeal cavity whose width deviates from what is predicted by the model. In addition, this deviation is likely to change the acoustic properties of speech, and formant frequencies in particular.

For this reason we are interested in assessing the geometrical and acoustic consequences of head posture in speech production from MRI data by using direct formant estimation and acoustic simulations.

3.2.2 Experiments

The experiment can be divided into three parts: 1) image information extraction, 2) formant estimation, and 3) acoustic simulations.

Image information extraction

The data that we used for the experiment is 3D MRI data of the vocal tract of five vowels of French language /a/, /œ/, /*i*/, /*o*/, /*u*/, in three different head positions: up, natural and down. Using tools provided by ITK-SNAP, we manually segmented the vocal tract of the mid-sagital slice. We then used meshlab [CCC+08] to apply Laplacian smoothing filtering with a step of 3 to all the images.

In order to measure the head position, we used the measurement proposed in [PKSF17]. The main idea is to use an angle defined by two lines to define the head position. The first line is the one that connects the interior edge of the C2-C3 cervical vertebrae. The second line is the one that connects the posterior tip of the spinous process of C1 and the tuberculum sella. As shown in Fig. 3.8, number 1 corresponds to the first line, number 2 corresponds to the second line, while number 3 corresponds to the calculated angle. We used imageJ software [SRE12] to manually specify the lines and make the angle computations. The average angle was $144.8 \pm 0.6°$, $124.7 \pm 0.8°$, $101.3 \pm 1.1°$ for the up, normal and natural position respectively. Since there is at least 20° of difference between the three positions, we were expecting to notice some difference in phonation [JMD94].

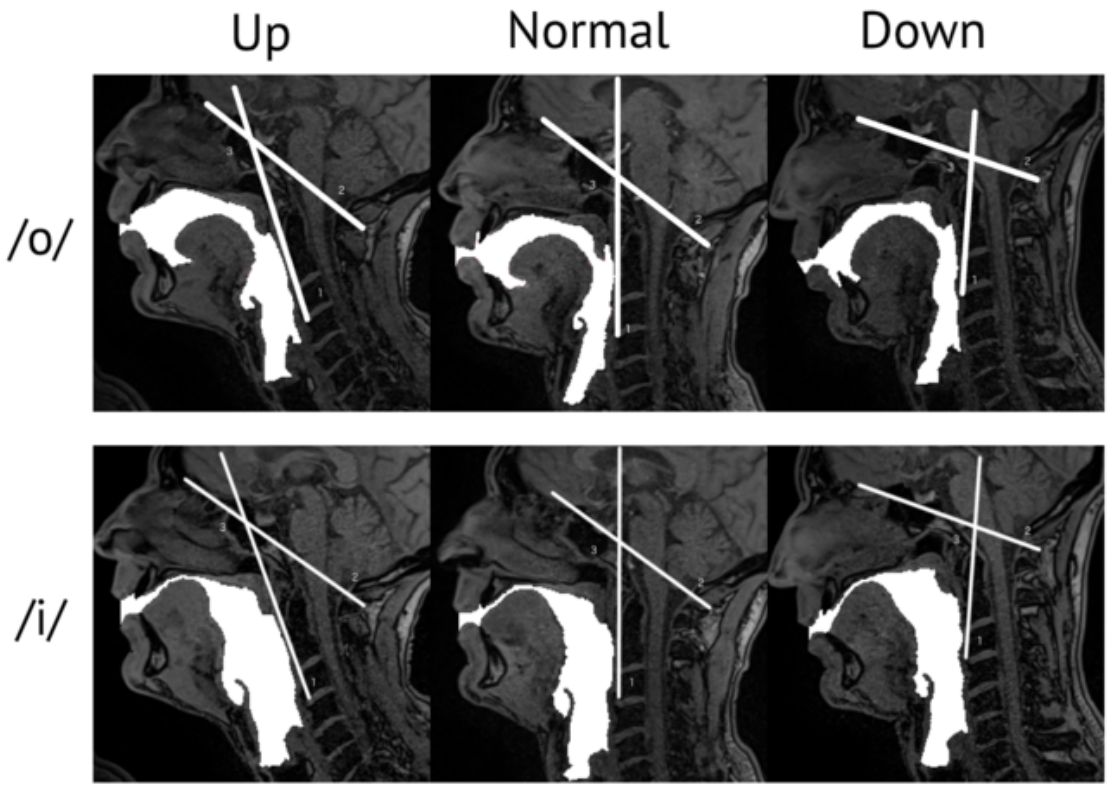


Figure 3.8: 2D segmentation of /o/ (top row) and /i/ (bottom row) vowels at up, normal and down position from left to right. Lines 1 and 2 are used to define the angle 3 of the head position

Acoustic simulations

For the purposes of our experiments, we used k-wave toolbox for MATLAB [TC10a] to simulate how the acoustic wave propagates through the vocal tract until it reaches the lips as described previously.

Finally, we computed the transfer function of every vocal tract and computed the peaks that appear in the frequency domain to compare them with the formants of the original audio signal as shown in Table 3.4.

3.2.3 Discussion about the effect of head position on phonation

The visual inspection of images shows that the up position (bigger angle between the pharyngeal and mouth cavity) increases the volume of the back cavity corresponding to the pharynx but reduces its length. Conversely, the down position essentially results in a change in the angle between the two cavities, and in a smaller size of the front cavity but does not significantly change the volume of the pharyngeal cavity. The acoustic impact of these modifications are visible in Table 3.4. It should be noted that the variations in formant frequencies between the neutral position and the other two positions are

	$F1$	$F2$	$F3$
/a/ - up$_{sp}$	648	1155	2213
/a/ - natural$_{sp}$	699	1112	2261
/a/ - down$_{sp}$	696	1065	1970
/a/ - up$_{sim}$	667	1156	2213
/a/ - natural$_{sim}$	660	1321	2294
/a/ - down$_{sim}$	637	1074	2089
$/oe/ - up_{sp}$	367	1345	2134
$/oe/ - natural_{sp}$	388	1245	2077
$/oe/ - down_{sp}$	375	1257	1901
$/oe/ - up_{sim}$	314	1421	2168
$/oe/ - natural_{sim}$	313	1301	1975
$/oe/ - down_{sim}$	311	1323	1867
/i/ - up$_{sp}$	269	1900	3040
/i/ - natural$_{sp}$	250	1830	2940
/i/ - down$_{sp}$	272	1810	3215
/i/ - up$_{sim}$	275	2049	2983
/i/ - natural$_{sim}$	243	1867	2801
/i/ - down$_{sim}$	281	1742	3147
/o /- up$_{sp}$	378	774	2159
/o /- natural$_{sp}$	360	754	2013
/o /- down$_{sp}$	360	750	1839
/o/ - up$_{sim}$	379	790	2139
/o/ - natural$_{sim}$	331	867	1954
/o/ - down$_{sim}$	325	799	1873
/u/ - up$_{sp}$	294	686	2008
/u/ - natural$_{sp}$	257	760	1949
/u/ - down$_{sp}$	299	769	1822
/u/ - up$_{sim}$	276	799	1940
/u/ - natural$_{sim}$	231	829	1915
/u/ - down$_{sim}$	281	875	1826

Table 3.4: Formants of the five vowels in three positions. The formants of the speech signal are marked as *sp*, and the formants from the simulations as *sim*

confirmed by acoustic simulations in terms of direction, even if their magnitude is sometimes different. This last point can be explained by the fact that numerical simulations are bidimensional and that formants were measured with some difficulties in the noisy speech.

In terms of formant frequencies the effect is quite negligible for the first formant of vowels that have a large pharyngeal cavity because the volume increase is proportionally quite small. On the other hand, the effect is more pronounced for $F2$, either because it corresponds to the half wavelength for the pharyngeal cavity for /i/ which results in

a lower value for the neutral and down positions (longer pharyngeal cavity), or because it corresponds to the oral cavity in the case of /u/ and /o/ for the neutral and down positions, which results in a higher value (smaller volume). Future work will focus on techniques for adapting articulatory models from these data and observations.

3.3 Impact of approximation at the level of velum and epiglottis

3.3.1 Introduction about geometric simplifications of the vocal tract

Geometric modeling of the vocal tract is used in particular to produce input data for articulatory synthesis [Bir13]. One of the challenges is to obtain a concise description and to remove geometric details that do not change the acoustic parameters significantly, from a perceptual point of view. Those simplifications could lead to a reduction of the number of parameters used to describe the vocal tract geometry, and consequently make the calculation simpler.

In general, more attention is paid to the jaw, tongue, lips and larynx, compared to the velum and epiglottis. The velum is indirectly taken into account more for representing the opening of the velopharyngeal port than for its impact on the oral cavity.

Concerning the epiglottis, its position depends on the size of the pharyngeal cavity, and thus on the tongue position. For a vowel with a large back cavity (as in /i/), the epiglottis stays apart from the back of the tongue. On the other hand, when the back cavity is more constricted (as for /œ/), the epiglottis is sometimes pressed against the back of the tongue (Fig. 3.9).

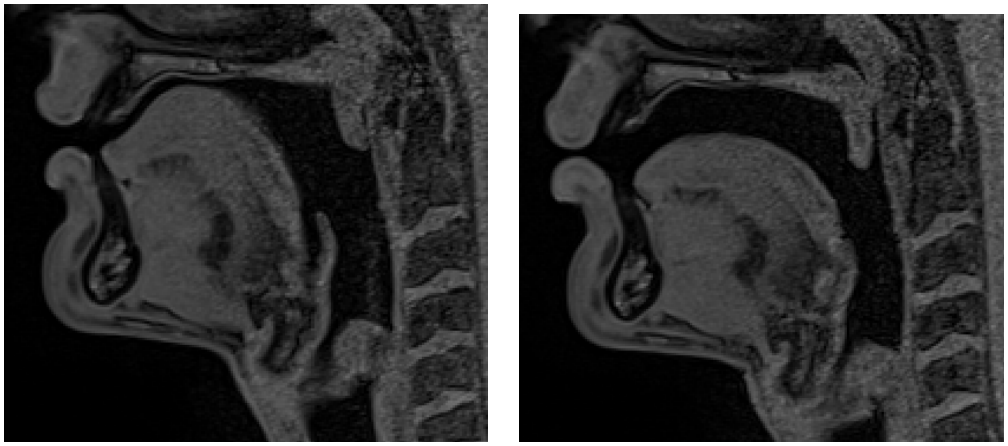


Figure 3.9: Epiglottis separate from tongue in case of /i/ (left), epiglottis pressed against the tongue in case of /œ/ (right)

Our approach to geometric modeling of the vocal tract is based on an articulatory model that independently controls each of the articulators. The first articulator is the mandible (which corresponds to the opening of the jaw) because it influences the tongue

and the lips. Two parameters are sufficient to control the opening of the jaw with good precision (jaw angle, jaw horizontal position). The tongue is the articulator that achieves the greatest number of articulation places, and its description must be fine enough to reach a precise position and shape. For this reason, (unlike the Maeda model [Mae90]), there are attempts that use between 6 and 10 deformation factors. The influence of the jaw is taken into account to determine the influence of the tongue and lips.

The epiglottis is actually a cartilage, and therefore the influence of other articulators that interact with the epiglottis, i.e. the mandible, tongue, and larynx, is decisive. Hence, their contribution through linear regression factors is more important than its intrinsic deformation factors. Once the midsagittal shape is calculated, it is necessary to find all the resonating cavities, their area functions and the global topology to run the acoustic simulation [EL16]. Geometrical simplifications allow faster simulations and avoid changes of the global topology when a small cavity appears.

The objective is to investigate the impact of geometric simplifications in order to better understand those that can be made without removing important acoustic cues. Unlike Arnela's work [ADB+16], which treats the vocal tract as a whole by transforming it into a piece-wise elliptical and then cylindrical tube, we treat the articulators separately because articulatory synthesis requires that each of them be controlled independently of each other.

We used MRI data of the vocal tract with simultaneous speech recordings of five French vowels to study the articulators' effects. The real speech signal was used as a reference. We edited the images to remove the velum and the epiglottis and then used acoustic simulations to see how the transfer function of the vocal tract was affected, and, therefore, what the role is of these two articulators in phonation.

3.3.2 Experiments

The experiment consisted of two parts: 1) image processing and 2) acoustic simulations.

Image processing

In this experiment we used five vowels of the French language, /a, œ, i, o, y/. First, we used 3Dslicer software to correct the axis orientation of the $3D$ images since there was a small angular offset of about -6 degrees in the sagittal field. We used Lanczos interpolation to resample the image with the Lanczos filter parameter chosen as $a = 4$.

When then employed the ITK-SNAP software for semi-automatic segmentation of the vocal tract $3D$ volume.

The Nearest Neighbourhood algorithm was used to create the probabilistic map to use for the automatic segmentation of the vocal tract. We used two classes and 10000 points for consideration as the nearest points for class categorization for the creation of the probabilistic map. Then an active contour algorithm was applied and the vocal tract mesh generated (Fig. 3.3).

Editing of the vocal tract geometry

For every mid-sagittal slice/vocal tract shape (for 2D/3D experiments respectively), three additional versions were created by processing the segmented images (4 images per vowel in total with the original). In the first version we edited the vocal tract geometry to withdraw the epiglottis, in the second we used a constant wall approximation at velum (by withdrawing the velum extremity) and in the third version we combined the previously described simplifications. These three versions of every vowel along with the original were the data used in the simulations (Fig. 3.10-3.13).

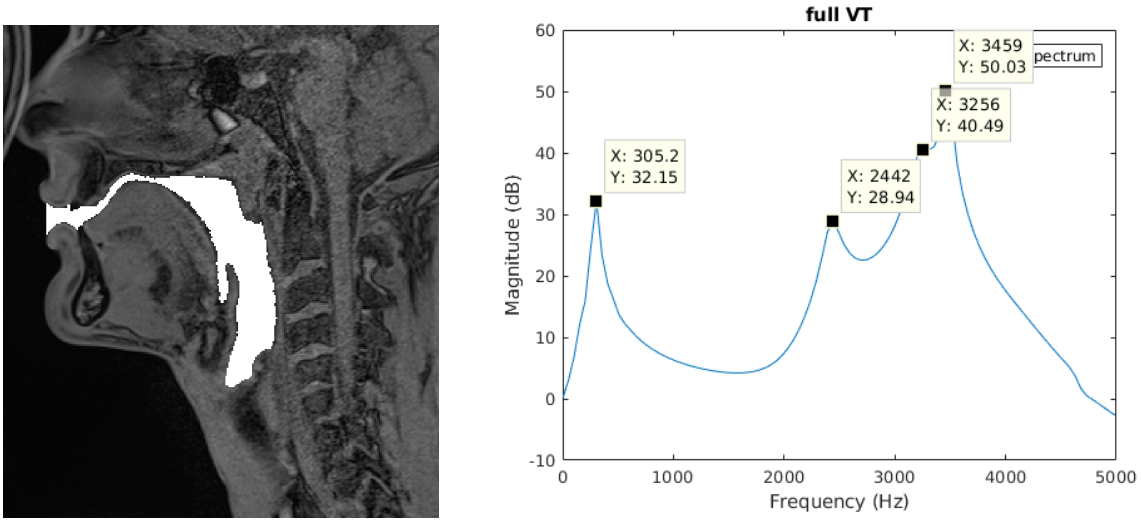


Figure 3.10: 2D segmentation of /i/ with full vocal tract (left) with its 3D spectrum (right)

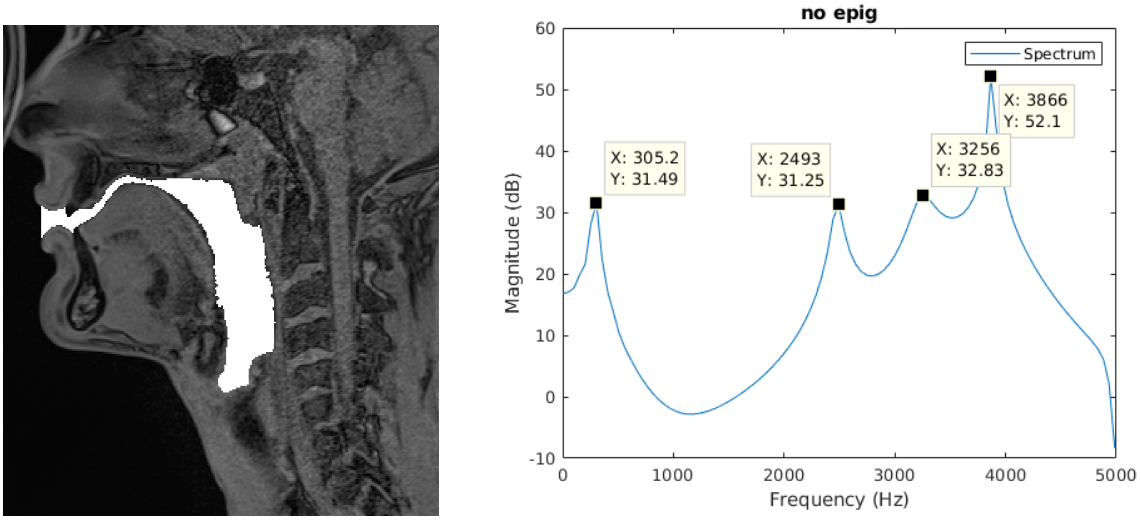


Figure 3.11: 2D segmentation of /i/ without epiglottis (left) with its 3D spectrum (right)

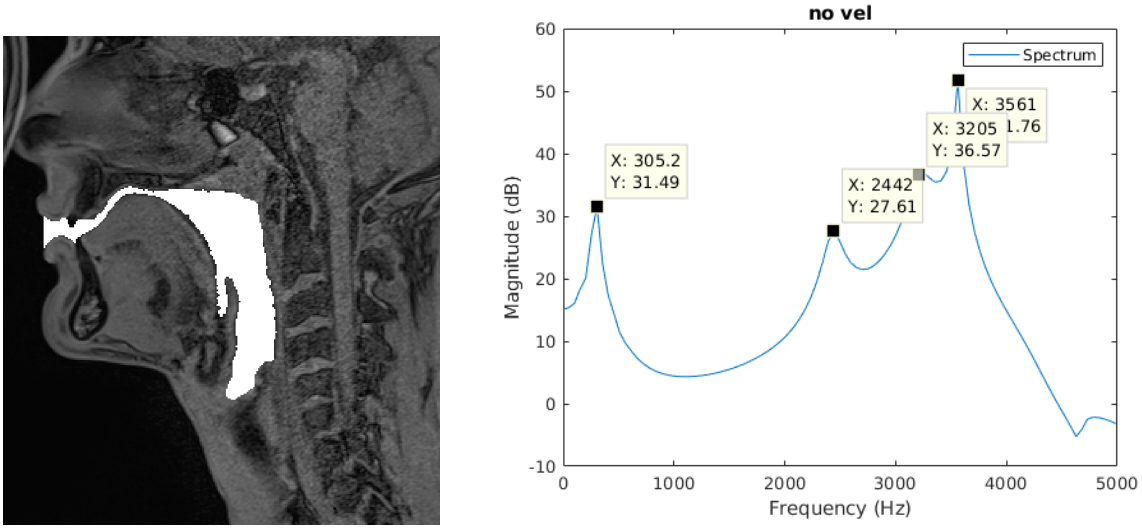


Figure 3.12: 2D segmentation of /i/ without velum (left) with its 3D spectrum (right)

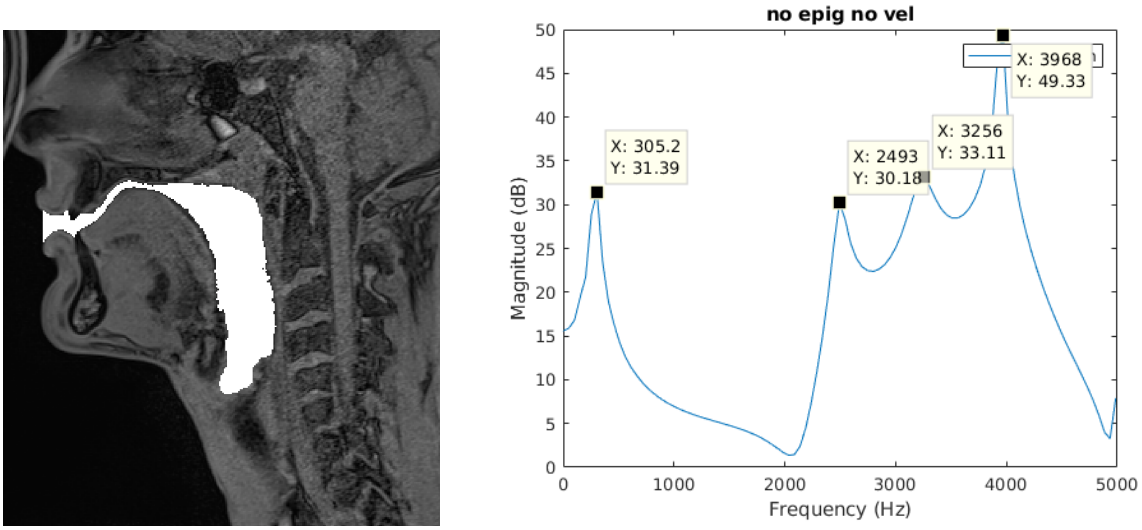


Figure 3.13: 2D segmentation of /i/ without epiglottis and velum (left) with its 3D spectrum (right)

Finally we used Meshlab [CCC+08] to smooth every mesh by applying Laplacian smoothing filter with step 3. For every vowel, about 5 hours of processing were required, with the biggest amount of time spent on the manual segmentation step.

Acoustic simulations

To run the simulations we used k-wave toolbox. We computed the transfer function of every vocal tract and computed the peaks that appear in the frequency domain to compare them with the formants of the original audio signal (Table 3.5).

	$F1$	$F2$	$F3$
/a/ - speech signal	689	1256	2604
/a/ - full vocal tract	613/796	1164/1226	2758/2329
/a/ - no_epig	552/796	1164/1226	2880/2512
/a/ - no_vel	613/798	1532/1412	2451/2334
/a/ - no_epig_vel	613/796	1532/1409	2512/2390
/œ - speech signal	443	1335	2436
/œ - full vocal tract	429/429	1532/1348	2635/2451
/œ - no_epig	429/429	1532/1348	2635/2512
/œ - no_vel	429/429	1593/1471	2451/2329
/œ - no_epig_vel	429/429	1593/1409	2512/2390
/i/ - speech signal	380	2306	3193
/i/ - full vocal tract	337/305	2394/2442	3136/3256
/i/ - no_epig	304/305	2428/2493	3304/3256
/i/ - no_vel	304/305	2428/2442	3136/3205
/i/ - no_epig_vel	304/305	2461/2493	3271/3256
/o/ - speech signal	430	732	2619
/o/ - full vocal tract	444/326	999/782	2219/2475
/o/ - no_epig	444/326	943/977	2219/2801
/o/ - no_vel	444/326	1110/717	1997/1172
/o/ - no_epig_vel	444/260	1110/716	1942/1238
/y/ - speech signal	336	1854	2228
/y/ - full vocal tract	281/320	1798/1918	2192/2283
/y/ - no_epig	253/320	1798/1918	2220/2329
/y/ - no_vel	251/320	1826/1964	2192/2238
/y/ - no_epig_vel	253/320	1826/1918	2164/2329

Table 3.5: speech signal / simulations with full vocal tract / simulations without epiglottis (no_epig) / simulations without velum (no_vel) / simulations without epiglottis and velum (no_epig_vel) formants computation in Hertz for the five vowels (2D/3D).

3.3.3 Discussion about the effect of velum and epiglottis simplification

The first remark concerns the values of the formant frequencies of the normal vocal tract without simplifications. The lying position and noise in the MRI machine largely explain the deviations from the expected values for these vowels of a male speaker.

The second remark, which is illustrated by Fig. 3.10-3.13 is that the original geometry with the small cavity between the epiglottis and tongue root gives rise to a zero in the spectrum. This is all the more pronounced since the epiglottis is well separated from the tongue. In the case of the vowel /i/, it is also noted that a zero appears in the region of F3-F4. Additionally, the impact of the focal point between F3-F4 decreases with simplifications, resulting in a larger distance between F3 and F4.

Regarding simplifications, it should be noted in Table 3.5 that they do not have a very significant impact on the first formant. For F2, the velum simplification has a more pronounced effect. The changes at the velum have an impact on the constriction between the front and back cavities of the vocal tract. This mainly affects F2 which is more sensitive to the length of the cavities.

As can be seen on the spectra of numerical simulations, the impact of the epiglottis corresponds to the appearance of a small cavity that adds zeros in the spectrum, essentially at high frequency since this cavity is small. Nevertheless, the impact of this cavity is far from being negligible above $2500Hz$ when the epiglottis is well separated from the tongue (thus not for /a/) because the zero appears in the region of F3. Further work will focus on the development of simplification control algorithms to ensure that they have as little impact as possible on the formants.

3.4 Discussion about acoustic simulations

Simulations is a good way to explore issues such as vocal tract simplifications like those presented earlier, which is hard to study in real life conditions. One can further extend these studies by involving more phonemes, more subjects and creating more generic models of the impact of several vocal tract simplifications or of the effect of various head positions on phonation. Another interesting approach proposed by people in the conference where this work was presented is exploring strategies and models to automatically decide the cases that simplifications on the vocal tract geometry can be applied without significant side effects. These models could also be applied to the field of articulatory speech synthesis with potential uses in reducing models' complexity and transferring/adapting articulatory models of one speaker to another.

The presented studies were performed using static 3D data therefore a question that arises is whether it is possible to directly use these results for continuous speech and to which extend. Even though there are studies that try to explore at which cases static articulation matches dynamic speech the results are not very clear and further research is needed[TDSL19]. Additionally, there are works like [LETV18] that try to create articulatory models using static images. However, there are cases that the articulatory model from static images cannot describe dynamic speech which requires the use of dynamic images for the model creation.

By examining this question from another point of view, instead of trying to find the cases with similarities between static and dynamic speech, an interesting approach would be to try to combine the strong points of both of these types of acquisition and create post processed data that have the advantages of static acquisitions (good spatial resolution) and dynamic ones (good temporal resolution) at the same time. The main advantages of this approach over the previous one are 1) it could give information about the 3D dynamics of the articulators during speech and therefore enlarging its application further than the scope of the previous studies and not just limit its scope to these studies 2) this approach could provide information about what is happening at the cases that dynamic speech is different from static speech (rather than a low similarity metric) which could allow to build systems and models for such cases and not just report that in these cases

the static approximation cannot be used for dynamic speech.

Chapter 4

2D to 3D extension

As stated previously in Chapter 3 Section 3.4, it would be an interesting advancement in studies if it could be possible to process a 2D real-time MRI video of the vocal tract and generate the 3D dynamic shape of it with the same quality as 3D data. Another solution could be to transform 2D data into 3D. In this Chapter we present our experiments towards this direction. More specifically in Section 4.2 we explore a way to generate 3D dynamic shapes of the vocal tract from static and real-time MRI data and in Section 4.3 we present an algorithm that estimates dynamic slices of the vocal tract at sagittal and parasagittals planes. Some of the work presented here in Sections 4.2,4.3 have already been object of communication in several conferences (Paper 5 - Paper 9) as mentioned in the introduction of this thesis Chapter 1 Section 1.2.

4.1 Introduction about 2D to 3D extension

Nowadays, MRI can capture a position of the vocal tract that is 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. There are attempts to incorporate the temporal influence (coarticulatory effects) into such static data [Bir13, TEL17], 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 [LETV18] and imposing control over nasalization.

The protocol of rtMRI, on the contrary, selects only one slice (within the context of speech production research, typically the midsagittal one) and captures the tissues within that slice in real time [LZL$^+$19, RFBM19, FBH$^+$17]. Since no constrain is imposed, 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 [Mer73] 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 [HS65, ML13] 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 speech movements [NTR+14, TN16, RTP+18]. 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 in the volume of the thick slice gets integrated 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 held 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 image segmentation especially difficult.

In [ZKP+12] a method is presented which visualises the 3D shape of the vocal tract during speech using rtMRI data. The main idea of this approach is acquiring rtMRI data on parallel sagittal slices with simultaneous (and aligned) audio recordings. MFCCs (Mel-Frequency Cepstral Coefficients) are extracted from audio recordings and DTW (Dynamic Time Warping) is applied on them to temporally align the sagittal rtMRI data in order to create the final 3D dynamic visualisation of the vocal tract. A disadvantage however of this approach is that even though it profits from the high acquisition rate of rtMRI, it does not take into account the advantages that static MRI provides in order to synthesise the 3D volume.

Hence the motivation for the work proposed in this Chapter: the need to combine the strong points of MRI and rtMRI. MRI has good image quality and volume information; rtMRI on the other hand has a good temporal resolution. To attain such a goal, we need to learn how to overcome their weaknesses, 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. Our strategy is based on the hypothesis that the first frame of the dynamic CV will have similar articulation as the midsagital frame of the 3D C(V) and likewise the final frame will have similar articulation as the midsagittal frame of the 3D V. Using these two cases as starting points, we create two versions of 3D dynamic estimation. The main idea is (for the case of consonant) 1) image transformations between the midsaggital and the other sagittal slices of the 3D static C(V) images are computed, 2) these transformations are applied to the first frame of the dynamic CV sequence. This results in an estimation of the 3D shape of the vocal tract for the first frame of the rtMRI C(V). 3) A set of transformations are computed between the first (midsagittal) frame of the dynamic CV sequence and the rest frames of this sequence. 4) These transformations are applied to the estimated (at step 2) shape of the vocal tract to create 3D dynamic version based on the consonant. The corresponding procedure is applied to vowel in order to create the second version of the 3D dynamic shape. Finally the two versions are properly fused to create the final 3D dynamic shape.

4.2 Dynamic 3D vocal tract shape generation

4.2.1 Acquiring the data

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. 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. Data used are part of the ArtSpeechMRIfr database.

4.2.2 Phonetic alignment of sound recordings

The transcription of the continuous speech corpus was phonetized by eLite HTS [RBBD14], and those phonetic labels were force aligned with HTK [YEG+02] using Merlin as frontend [WWK16]. Alignment results were manually checked and corrected in case of errors.

4.2.3 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 [Thi98] To find the transformation between the images the algorithm described in [VPPA09] 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. More details are presented in subsection 4.2.5.

4.2.4 Denoising procedure

To denoise the images, we applied thresholding, cutting off all the pixels with values less than 10% of maximum intensity. Therefore for an $N \times M$ size image I we have: if $I(i,j) \leq 0.1 \times \max\{I\}$ then $I(i,j) = 0$, with $0 \leq i \leq N-1$, $0 \leq j \leq M-1$, i, j being pixel position. 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 were smoothed out with a median filter.

The only problem that remained then was to treat the artifacts that were mostly caused by the speaker articulator movements. 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 [kov] toolbox. After this, we applied a Gaussian filter (size 15×15, $\sigma = 5$) so that the image pixel

convoluted 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.

4.2.5 Experiments on 3D shape generation

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 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 comprises the following steps: image pre-processing, double (C and V) 2D mapping, space and time extension, image combination and denoising.

Image pre-processing

We selected the images corresponding to the target syllables by using the phonetic alignment information. One of the problems that we faced 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 the pixel level. We increased the images so that the new images have 1.55 times the dimensions of the original images using cubic interpolation.

Since our algorithm is mainly designed for studying speech, our main priority was to create a realistic VT (vocal Tract) 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. 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 smallest 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 keep just a thin layer of the palate 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 sufficiently large to fit the VT of the other speaker with a different anatomy. In the end, a window with dimensions 146×131 was selected (Fig. 4.1).

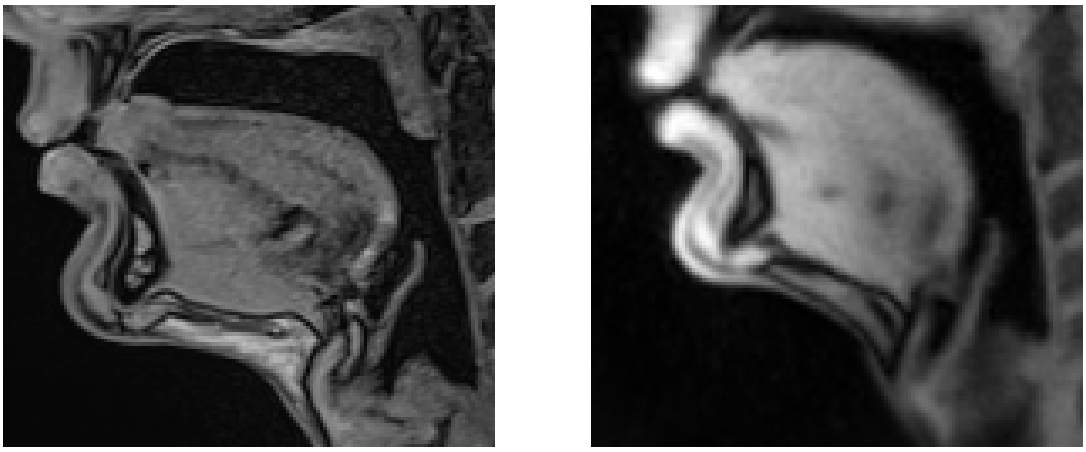


Figure 4.1: Static (left) and dynamic (right) recordings of /t(u)/ by S_1

Double 2D mapping

We 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 Chapter (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 image transformation to have a similar contrast between 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.e. I_c, and one for the V, i.e. 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 (n stands for normal) and T_i (i stands for inverse) respectively.

Space and time extension

This stage is divided into three steps: deformation in space, deformation in time 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 one 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 $T_s(i)$ (s stands for space).

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 $T_{d-n}(i)$ ($d - n$ stands for dynamic-normal). We do the same calculations for the inverse sequence (from frame $i + 1$ to i, the reason will be explained later in 4.2.5) and we call it $T_{d-i}(i)$ ($d - i$ stands for dynamic-inverse).

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, i.e. T_n, using the transformation for sagittal transformation, i.e. T_s. The resulting transformation

$$T_{s-r}(i) = T_s(T_n(i)) \tag{4.1}$$

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. This transformation is important as it allows us to apply the information extracted from space of the static image to the space of the generated images. One could think of directly using T_n however this will not work properly due to differences in articulation between static and dynamic domain (Fig. 4.3 top row).

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, i.e. S_f, and one starting from V and propagating backwards to C, S_b.

Image combination

Since we have two versions of 3D dynamic images (i.e. S_f, S_b), we need to combine them. The reason that we chose to have two versions is due to the core idea of our approach. If we had one version created for example directly from the consonant, as we start propagating the 3D shape along time towards the vowel, estimation errors stack resulting in a poorer estimation the further we move away from the starting point. However, the further we move away from the starting point, the closer we are to the end point for which 3D data is available. Therefore by creating a second version of 3D dynamic images with the inverse direction (starting from the vowel and moving backwards towards the consonant) we have better estimations at the time points closer to the end where the quality of the first 3D estimation has degraded and vice versa. An issue that appears at this point is how to combine the two versions in order to acquire the final one. 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} ($f-t$ stands for forward-transformed). 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 (by visually checking) since the articulation of vowels are more similar between them for dynamic and static speech compared to consonants. Finally, we cropped the left part of the transformed images S_{f-t} for the duration of the consonant that covers the place of articulation [FKJ06] (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. This is an empiric choice but in practice it works quite well since the gluing line is not visible at the final images (Fig. 4.3 bottom row). 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. A visual representation of the algorithm can be seen in Fig. 4.2

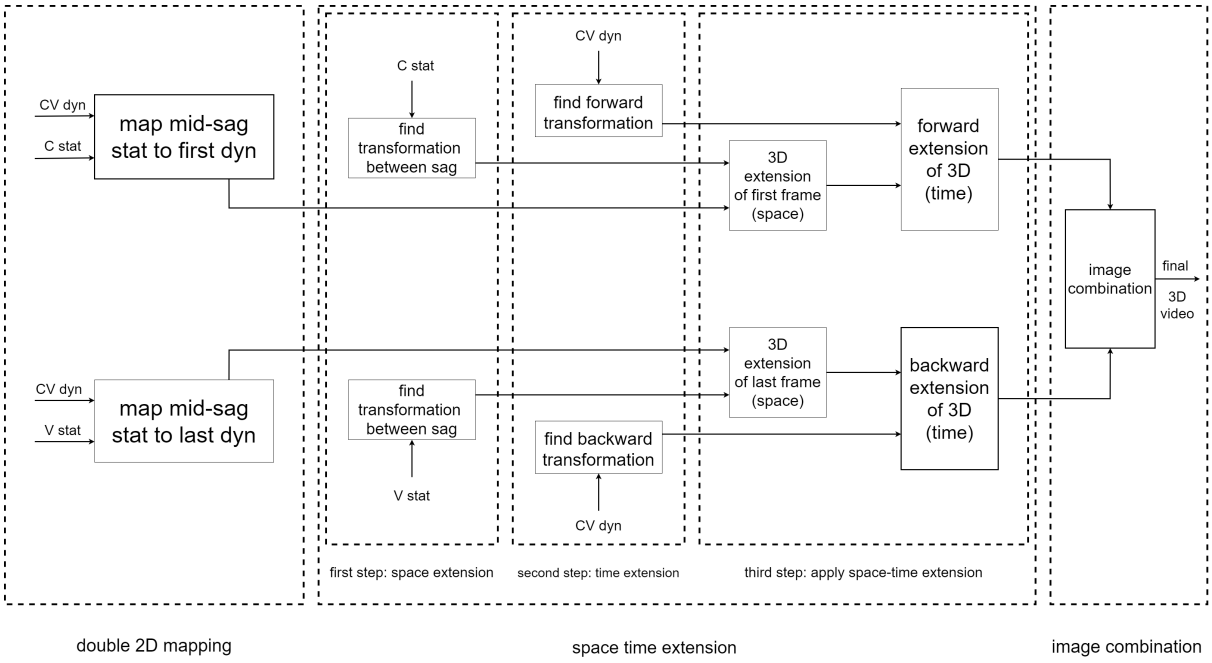


Figure 4.2: Block diagram of the dynamic 3D vocal tract shape generation algorithm

Denoising

The denoising procedure that we used was same as described in Section 4.2. 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 Fig. 4.3.

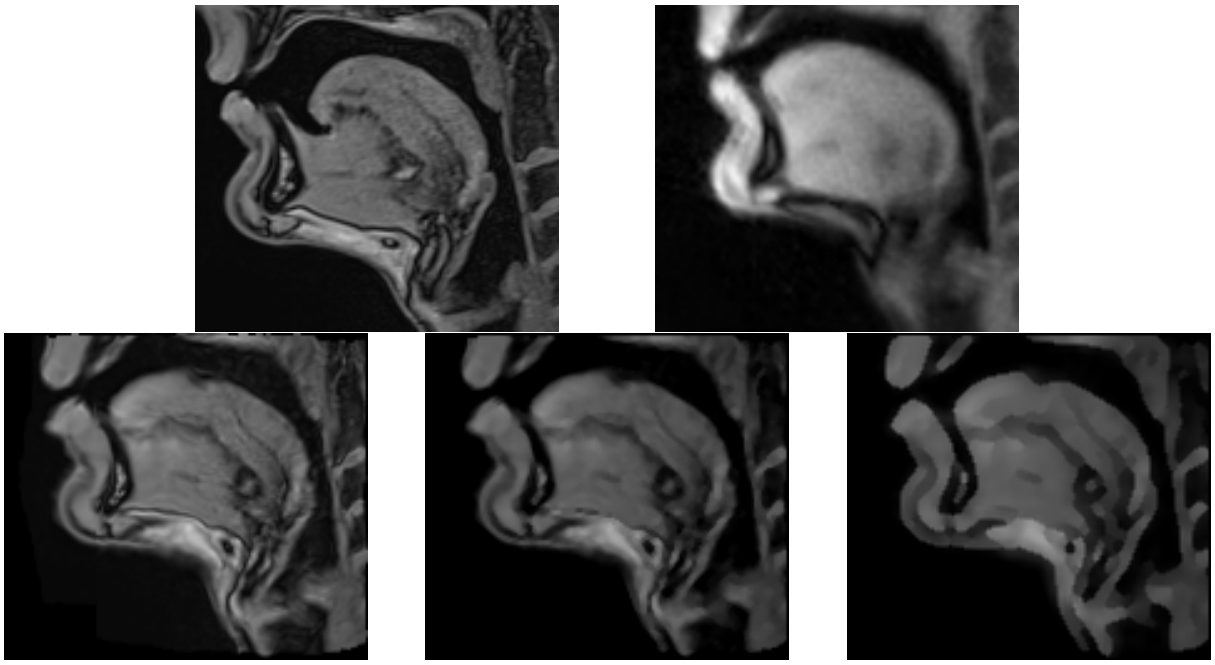


Figure 4.3: From left to right, top to bottom: /f(a)/ by S_1 in 3D; 2D; the raw generated midsagittal sequence when transforming to S_1; its denoising; its denoising and smoothing

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 and consequently we cannot perform numerical evaluations. Therefore, we asked two people from our team 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 more or less ok, both in the same and cross-speaker cases (Fig. 4.4).

For the non-midsagittal slices, we visually inspect them to check if we have smooth transitions based on what we would expect from the static 3D MRI. Although there were some problematic cases, either related to insufficiently good image combination or due to filtering, in most of the cases the non-midsagittal slices looked for some regions as we would expect while for other there were some problems (Fig. 4.5).

Of course since there several image transformations and estimations, errors that stack reduce the quality of the generated vocal tract shapes. For example motion-looking artifacts like at the back of the tongue in Fig. 4.5 bottom row or thicker lips Fig. 4.3. Even though images were filtered, resulting in better image quality, sometimes they cause small distortion of the images like in 4.3 where a small hole appears on the surface of the tongue.

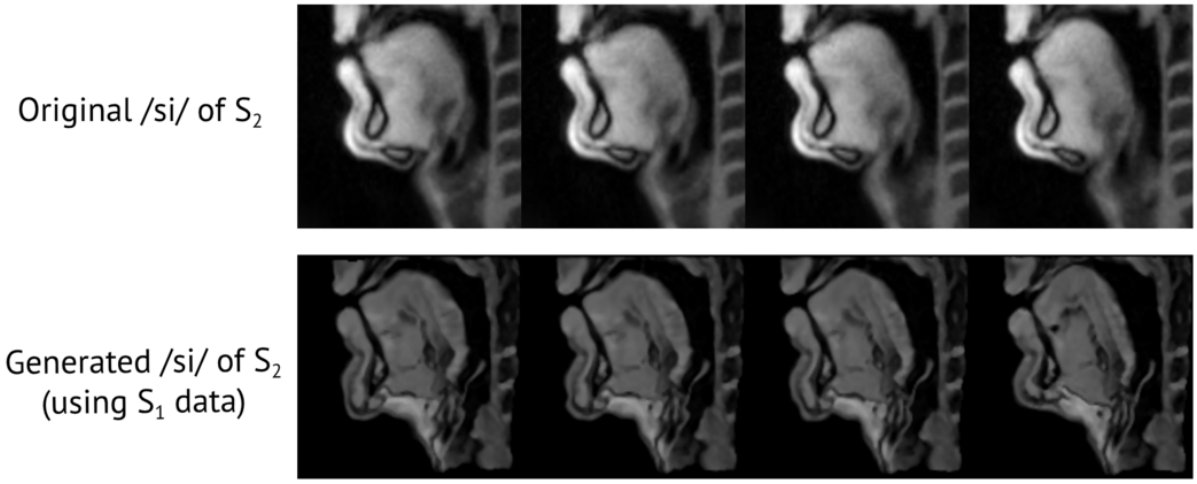


Figure 4.4: Every 3rd image in the original dynamic 2D articulation of /si/ by S_2 (above, left to right) and the generated midsagittal slice sequence when transforming to the 3D data of S_1 (below, left to right)

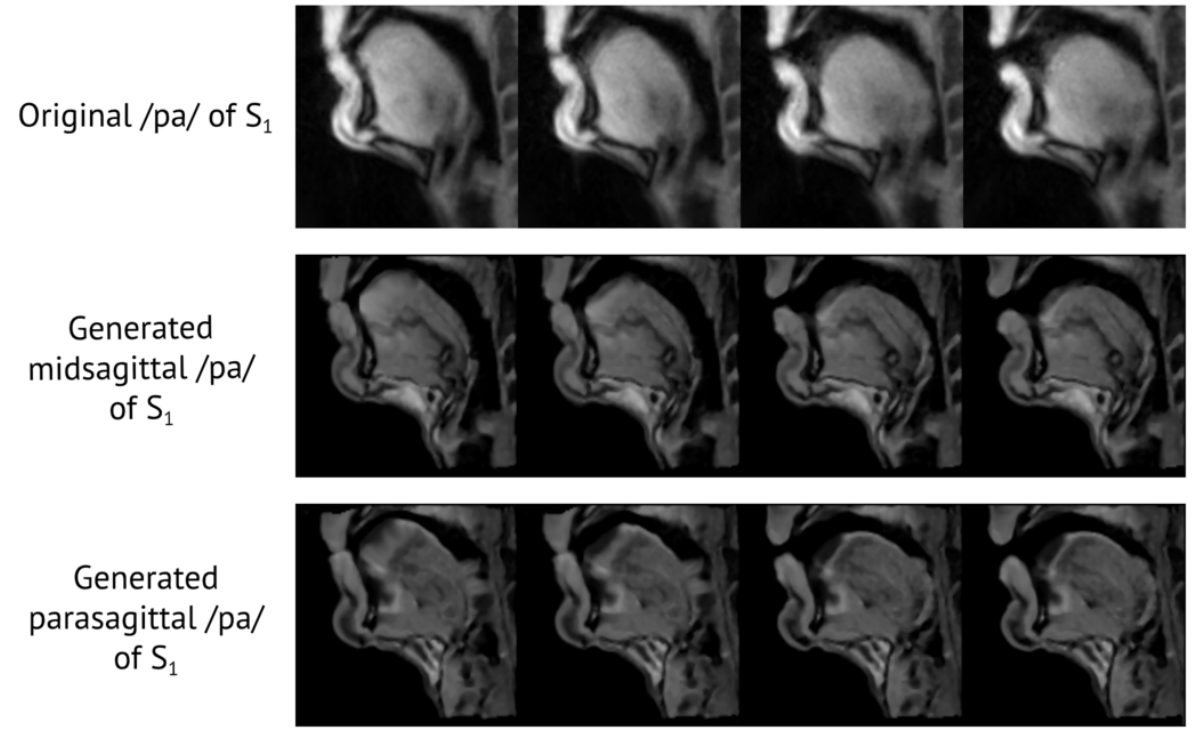


Figure 4.5: Every 2nd image in the original dynamic 2D articulation of /pa/ by S_1 (above, left to right), the generated midsagittal slice sequence when transforming to the 3D data of S_1 (middle) and sagittal sequence (slice 67 out of 120; below)

This is because in the raw generated image the surface of the tongue at this point has relatively lower pixel intensity from the rest of the tongue and during edge detection step of the filter the surface of the tongue was not detected properly and this small hole filtered out as noise. Even though the resulting images are not perfectly synthesised (especially in the parasagittal frames) there is still some value mainly in the midsagittal ones. For example, by comparing the top left image with the bottom middle one in Fig. 4.3 we can visually notice that the generated vocal tract shape is very close to the original one, the region near the vocal folds is less noisy and the velum and epiglottis are more clear. One can notice however that the epiglottis in the generated images is out of position compared to the original images, which is mainly due to differences in articulation styles between the two subjects.

4.2.6 Conclusions about dynamic 3D vocal tract shape generation

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 4.2.1). Especially in the plosives /t/ and /p/, just maintaining a stable articulation is a challenge for the speaker. However, in the case of vowels, the situation is better since vowels can be phonated alone. 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 the capture of the articulators' contact (lip closure, tongue touching the palate) due to the calculations of the deformation field, even though the constriction narrows in the C part. This is mainly due to the smoothing constrain which prevents deformation field from having discontinuities. That is why we extract this type of information from S_f which captures 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 acquisitions (Fig. 4.1). 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 because /s/ and /f/ are fricatives that can be sustained unlike stops.

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 for the midsagittal plane. 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.

4.3 Further extensions

As stated above, the two main limitations of the algorithm are the fact that it is not fully automated and that there is no numerical evaluation. In the following subsection we explored another way of acquiring 3D dynamic information of the vocal tract by using midsagittal slices to generate additional sagittal slices. The main constraint that we put is the possibility of the approach to be numerically evaluated and secondly to be an automated process. In order to be able to do numeric evaluation we need to have reference data for the sagittal slices that we are going to synthesise. Additionally, these data could be used to create more realistic estimations of the sagittal slices compared to the approach described previously in Section 4.2. Therefore, we used 2D rtMRI data acquired on several parallel sagittal planes. The advantage of this approach is that 1) it does not require multiple types of estimations at the model construction (training), only one from the midsagittal slice to the neighbouring sagittal one therefore reducing potential stacking errors and 2) the generated results (test) can be directly compared with the reference.

4.3.1 Vocal tract sagittal slices estimation from MRI midsagittal slices

Firstly, we conducted a two speaker experiment therefore we chose a non-favourable case where the speakers were of different gender. One male and one female subjects were asked to repeat /fi/,/fa/,/fu/,/pi/,/pa/,/pu/,/si/,/sa/,/su/,/ti/,/ta/,/tu/ three times. Before the pronunciation of every CV, subjects were instructed to breath from the nose with closed mouth. This is important because we want the articulation to be as similar as possible for every repetition. Three sagittal planes (8mm slice thickness) were chosen for the acquisition, the midsagittal one and its left and right ones (with no space in between them). One plane was acquired per CV repetition. 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 [UZV$^+$10]. In total 2000 images (including silence) were acquired. As a data preprocessing step, images should be labeled with their corresponding phoneme. Even though we could have used standard forced alignment phonetic labelling tools we did it manually to achieve a better temporal precision. The first step is to apply piece-wise linear alignment between the midsagittal frames of the train speaker and all the sagittal frames of the train speaker, using the test speaker as reference. A non-rigid image transformation T_L, T_R of the train speaker that transforms the corresponding midsagittal frame to its left and right sagittal frame was calculated for every time frame. To compute the non-rigid image transformation a MATLAB function was used based on

the algorithm described in [VPPA09] as presented in 4.2.5. Additionally, a transformation A was computed every time frame that maps the midsagittal frame of the train speaker to those of the test speaker. Finally, using the group of transformations A we adapt T_L, T_R to the test speaker, getting T_{La}, T_{Ra}. By applying them to the midsagittal frames of the train speaker we acquire the synthesised left and right slices across time.

Secondly, we extended these result by using seven speakers for training to create seven separate single speaker models and then fuse them by averaging T_{La}, T_{Ra} transforms (across train speakers, not between them) in order to get T_{LF}, T_{RF}. T_{LF}, T_{RF} are now applied to the midsagittal frames of the test speaker to get the multi speaker estimated left and right slices across time. A visual representation of the algorithm can be seen in Fig. 4.6.

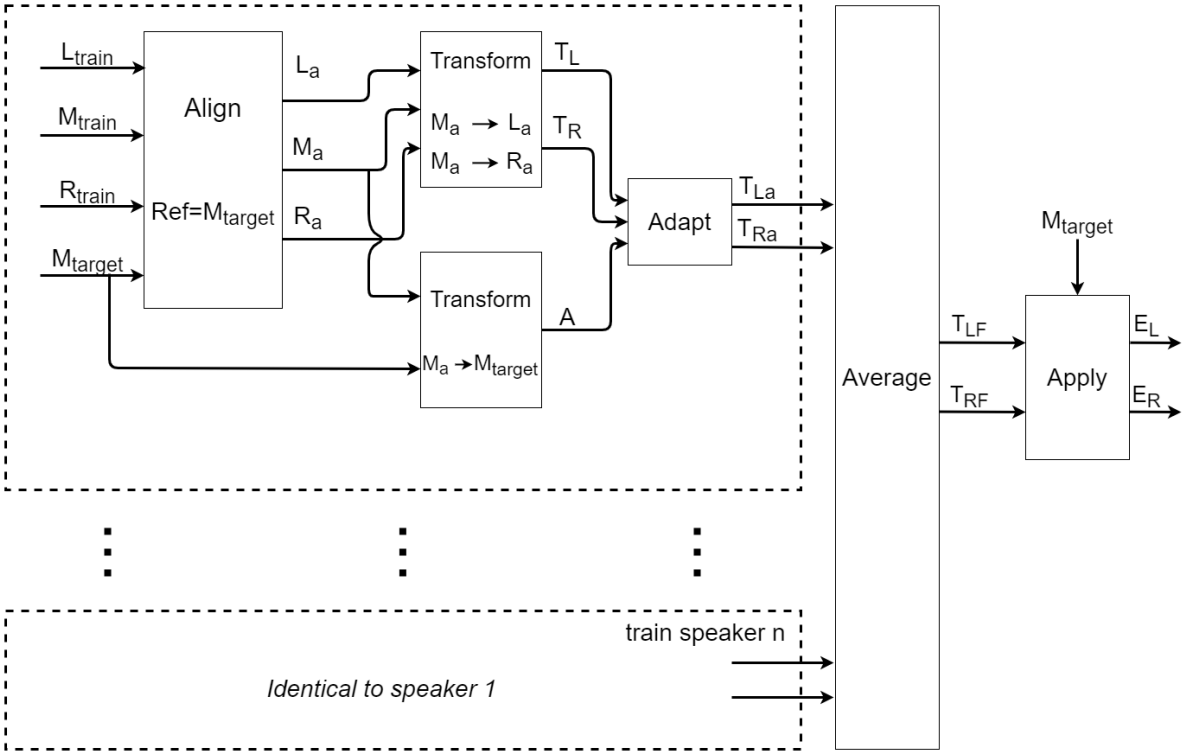


Figure 4.6: Visual representation of the proposed algorithm for sagittal slices estimation. First, several single speaker transformations are computed that are fused together at the next step to create the final estimations

To validate the results from both single and multi speaker version, cross-correlation between the synthesized and the original images [SAB+12], normalized by the autocorrelation of the original images, was used. For the multi speaker case, 8-fold cross validation was also used. Further details about data acquisition protocol and validation process are presented in Chapter 5 Section 5.1.

Results

In Fig. 4.7 we can see chosen images of the right side during /pu/ in both synthesized and the corresponding original form for the single speaker version. Synthesized images have average correlation of 0.9354 (±0.0106) 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 of 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.

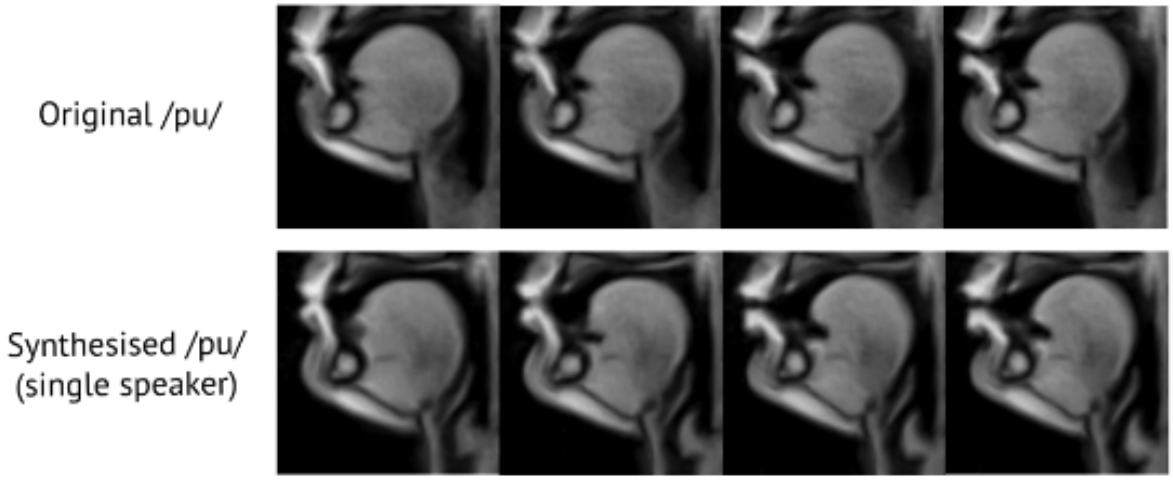


Figure 4.7: Selected right frames of /pu/ (single speaker). Top: original images and bottom: corresponding synthesised ones.

In Fig. 4.8 we can see chosen images of the right side during /ti/ in both synthesized and the corresponding original form for the multi speaker version. Synthesized images have average (8-fold cross validated) correlation of 0.9552 (±0.0073) with the original ones, using normalized image cross-correlation. The image quality is better and the similarity was increased compared to the single speaker model. By visually inspecting the images, we can notice that synthesised and original images are similar in terms of vocal tract shape, with few exceptions mainly in the palate, epiglottis and velum where the synthesised images look actually better. This point is further discussed in 4.3.1.

Conclusion about sagittal slices estimation

Our numerical results show that synthesised images are quite similar to the original ones with an average similarity of 0.9552 across all CVs and all planes (maximum value could be 1 which would show identical images). This fact can also be noticed by visually inspecting the original and the synthesised images in Fig. 4.8 as the biggest part of the

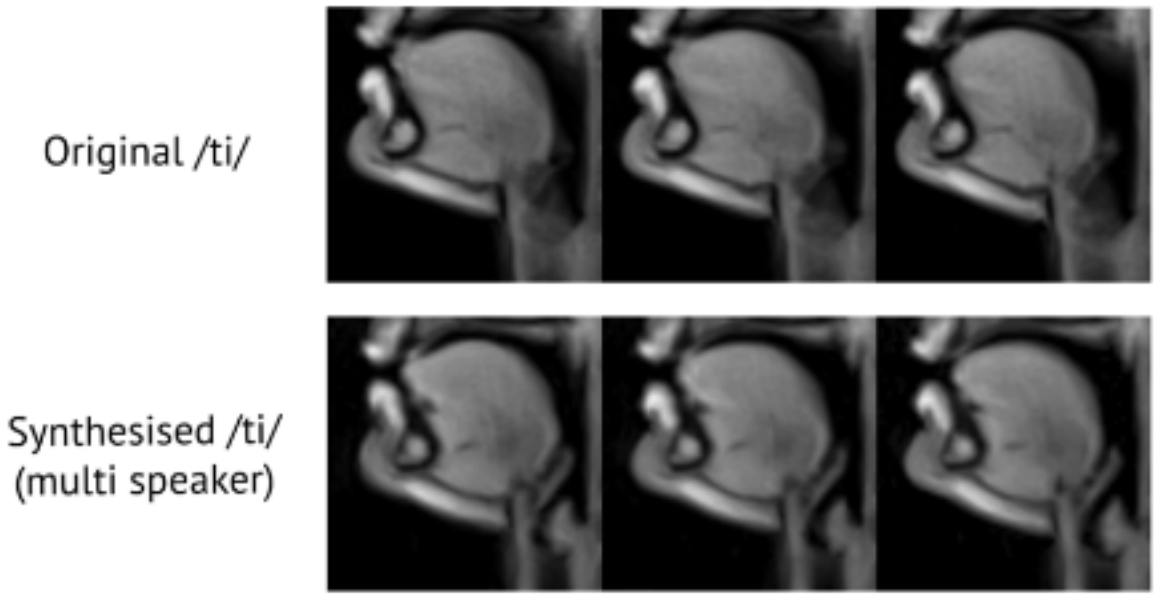


Figure 4.8: Selected frames for /ti/ of speaker 8 of right plane. Top: original images Bottom: synthesised images

vocal tract is almost identical between them. The main differences can be noticed at the palate and at the epiglottis/larynx where in the original images these regions are a little more blurry. This improvement in quality in the synthesised images is because they were based on images from the midsagittal plane where this region is more clear than on the Left or Right plane. We should note however that the fact that synthesised images look better in these regions, does not necessarily mean that they depict the true better. It should be at the judge of the user at which cases synthesised images represent the truth better. Additionally, the algorithm is quite robust since the standard deviation of the total average is 0.0073. This can also be inferred by examining the differences in anatomy, head position and articulation styles between subjects (Fig. 4.9) and checking how close the synthesised images with the original ones are.

This algorithm can be easily transformed to a fully automated algorithm by using standard techniques to replace the manual labeling at the pre processing step. The presented algorithm can be used to enrich information of midsagittal vocal tract slices. However, more work needs to be done towards the direction of generalizing these results by exploring cases like VCV, CVC, whole words or phrases. Finally, one could think of estimating further sagittal planes of the vocal tract.

4.3.2 Synthesize MRI vocal tract data using "silence" MR Images

The previous work of estimating parasagittal frames in subsection 4.3.1, inspired the idea of trying to synthesize midsagittal frames using limited data. In this work, we propose a method that captures the dynamics of CVs by using some image transformations and

Figure 4.9: Silence frames from two speakers. One can notice the differences in anatomy and articulation

adapts these transformations to a target speaker by using its "silence" frame in order to synthesize the data of the target speaker pronouncing the training CVs. By the term silence in this work we refer to a position of the vocal tract where the mouth is closed and the subject is breathing from the nose. Synthesized images were compared to the original images of the target speaker pronouncing the same CVs using image cross-correlation.

Algorithm description

The data used for this study were the same with those used for the previous experiment in subsection 4.3.1. Again, two experiments were conducted, the single speaker and the multispeaker one. As for the experimental part, 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 frame. To compute the non-rigid image transformation a MATLAB function which is based on the algorithm described in [VPPA09] was used (see 4.2.5 for more details). 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 $T_{train-test}$ transform. The newly created transformation T_{single} is then applied to the first frame of the test speaker and then propagated to every newly synthesized frame, creating the synthesised sequence CV_{syn}. The last step is to map the corresponding training time frames CV_{train} to the synthesized ones CV_{syn} to suppress some artifacts that are created due to the transformations. The resulting denoised sequence FE_{single} is the output of the single speaker model. To create the multispeaker model, several single speaker estimations are created, one per training speaker that are afterwards aligned and averaged to give the final multi speaker estimation FE_{multi}. A visual representation of the algorithm can be seen in Fig. 4.10. In order to validate the results, cross-correlation between the synthesized and the original images [SAB+12], normalized by the autocorrelation of the

original images, was used. For the multi speaker case, 8-fold cross validation was also used. Further details about data acquisition protocol and validation process are presented in Chapter 5 Section 5.1.

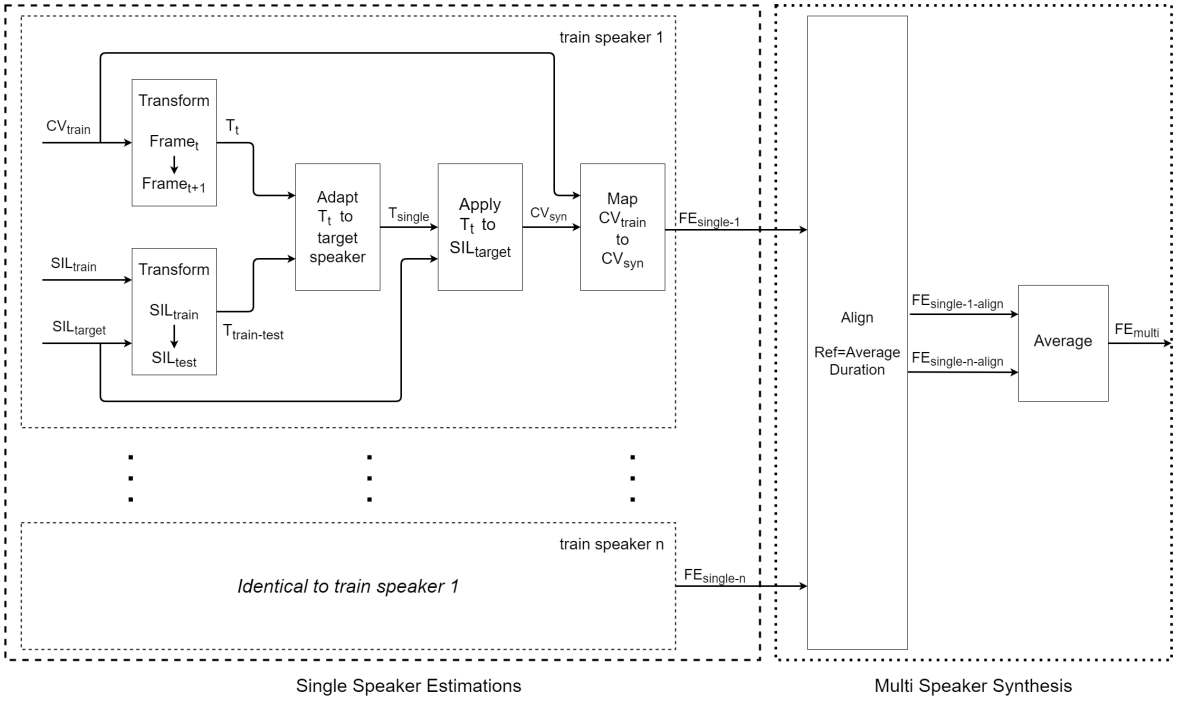


Figure 4.10: Visual representation of the proposed algorithm to synthesise midsagittal frames. First, several single speaker sequence estimations are computed that are aligned and averaged together at the next step to create the final multi speaker estimation

Results

In Fig. 4.11 we can see chosen images during /pi/ in both synthesized and the corresponding original form for the single speaker models. We used normalized image cross-correlation as a similarity metric and synthesized images have average similarity of 0.9437 ($\pm$0.0096) with the original ones over the set of syllables studied. By visually inspecting them, we can see that some differences appear 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. Furthermore the position of the glottis appears to be quite sifted which could affect formant frequencies. Apart from this, images look quite similar in terms of vocal tract shape with an exception of 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.

In Fig. 4.12 we can see chosen images during /pu/ in both synthesized and the corresponding original form for the single speaker models. Synthesized images have average

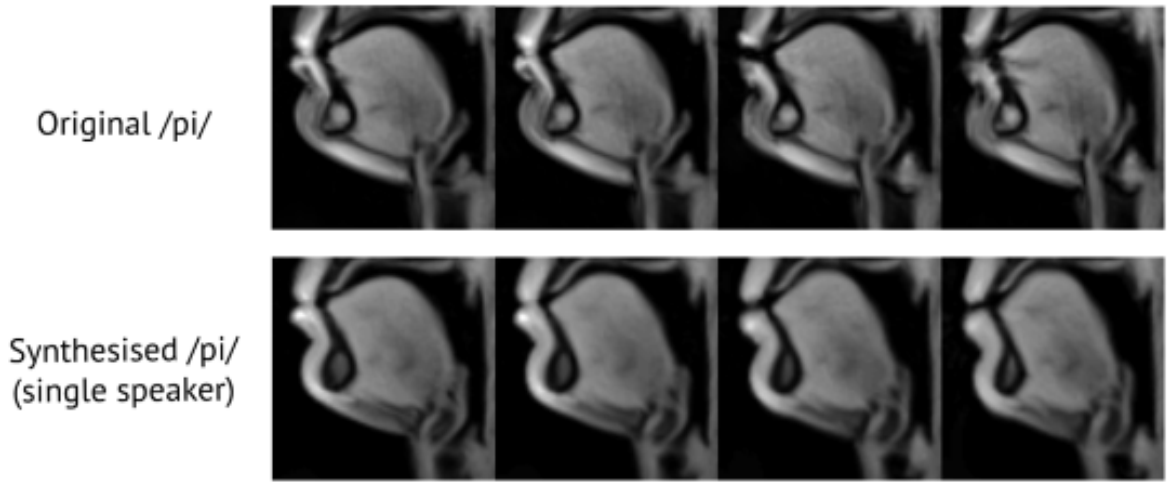


Figure 4.11: Selected frames of /pi/. Top: original images and bottom: corresponding synthesised ones.

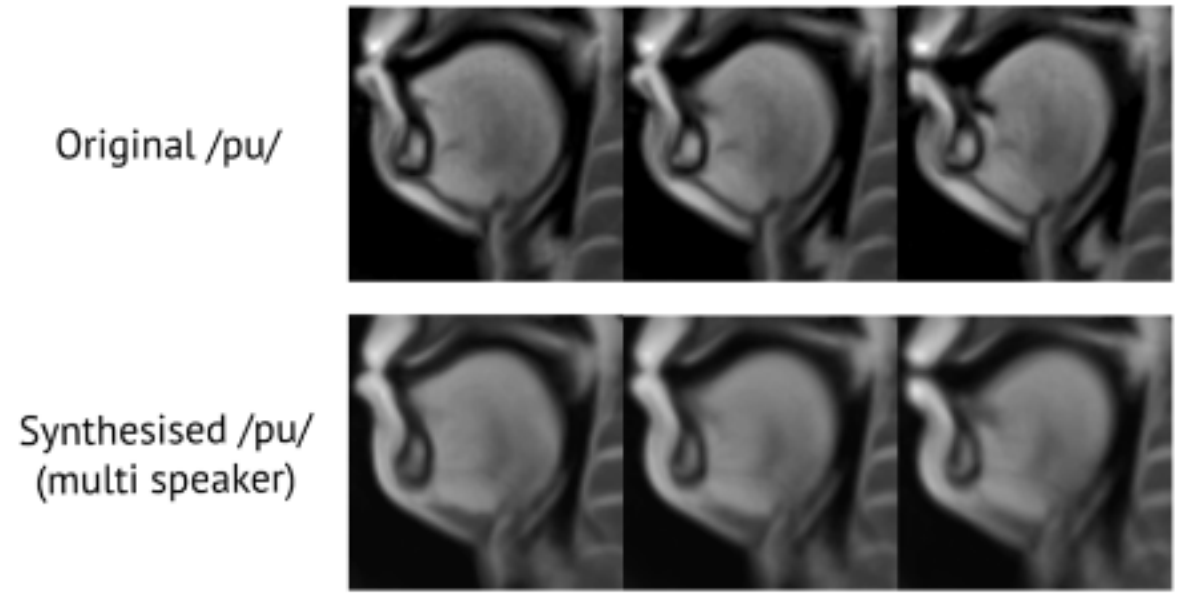


Figure 4.12: Selected frames for /pu/ of speaker 6. Top: original images; Bottom: synthesised images

match of 0.9361 ($\pm$0.0046) with the original ones over the set of syllables studied, using normalized image cross-correlation. The average similarity is slightly lower that the single speaker model but multi speaker model is a little more robust as it has half the standard deviation of the single speaker model. By visually checking the images we can notice that the epiglottis is more blurry and that the tongue is a bit blurry with some artifacts appearing at the front part. However, compared to the single speaker model, the shape of the back of the tongue is closer to the original one for the multispeaker model.

Conclusion about dynamic vocal tract shapes synthesis

Some differences can be observed in the images mainly due to different articulation and vocal tract shapes (Fig. 4.9). Additionally, the algorithm uses as phoneme duration the average phoneme duration of the train speakers since it does not take into account any known information about the speaking style of the test speaker.

By visually examining the results in Fig. 4.12, one can notice that synthesised images look quite similar with the original ones. However, 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 error 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. 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. Further research could include learning this registration more globally, or by using DNN learning techniques or other ways to implement the duration model.

4.4 Discussion about 2D to 3D extension

Generating dynamic shapes of vocal tract offers a great potential for speech research since it can provide information which otherwise would have been hard to acquire or even inaccessible. One can use such approaches in order to study 3D dynamic phenomena that appear during speech production. In Section 4.2 we proposed an algorithm that combines the strong points of static and real-time MRI to generate dynamic shapes of the vocal tract and in Section 4.3 we proposed ways to estimate sagittal and parasagittal slices. All of the presented algorithms can be applied to a single speaker therefore if someone would like to study speech dynamics and create models by taking into account inter and intra speaker variability it has to generate data for each training speaker and somehow combine them to create the model for the study. Even though such approaches look very promising, a potential weak point could be the fact that data generation is speaker dependent and this could induce some bias, as discussed with experts on the conferences that this work was presented.

An interesting approach that one can tackle this issue is to try to use instead of a specific speaker, a global model as a standard speaker which would be speaker independent and therefore inducing no bias. Such a global model could be quite challenging to create but it would further advance speech studies. A potential way to construct such a standard speaker could be to 1) use the algorithms presented in this Chapter to generate speaker dependant data, 2) combine them to create a biased model, 3) instead of the specific speaker, use the newly created model to repeat the process some times. The advantage of such approach is that it is straightforward, quite easy to implement and is expected to remove at least a part of the bias. On the other hand, one can argue that the convergence point of such an algorithm may not be guaranteed and could be at a partially biased point instead of an unbiased one. These problems appear mainly because speaker variability is not tackled as an issue inside the core part of the method but at a later stage where its (speaker variability) effect is tried to be reduced from an already biased created model. To handle this issue, one could think of managing the issue of variability one step earlier, at the core part of the method during the initial creation of the model in order to be "truly" unbiased. Of course, a disadvantage of such an approach is that the previous algorithms cannot be used directly therefore the problem should be handled from the beginning and completely new algorithms should be developed. However, the strong advantage of this approach is that the resulted models could be used as a fully unbiased and speaker independent model which would open new ways to tackle some open problems in speech production.

Chapter 5

Generic speaker model

In this Chapter we propose a way that one can use in order to create an artificial speaker which can be used as a standard speaker as explained in the previous Chapter. This will enable someone not only to find potentially better solutions to partially solved/explored problems but also to study issues from a new perspective. This Chapter is taken form an article submitted to the Journal of the Acoustical Society of America entitled "A 3D dynamic spatio-temporal atlas of the vocal tract during consonant-vowel production from 2D real time MRI", Paper 10 as mentioned in the introduction of this thesis Chapter 1 Section 1.2

Anatomical and gesture variability lead to a very large variability of MRI images of the vocal tract, which prevents the creation of a 3D model that can represent any speaker. The creation of a generic approach and model that incorporates this variability starting from its construction is thus crucial.

In the medical field, a popular approach to represent inter-subject image variability is the use of one or several atlases. Even though there are several definitions of what an atlas is, in this thesis when we refer to atlas we mean a general model that represents the structure of the studied organ/region and can be used as a reference to describe the anatomy independently of a specific subject.

In particular, this approach is very often used in brain studies for tasks like automatic region segmentation, region labeling, etc. For instance, several atlases built from data of adults have been used to automatically label and segment the brain regions of young prematurely born children [GRH$^+$08]. Each of the adult atlases was registered to the target child image and the final labeling and segmentation were based on a combination of the registration results. Such approaches facilitate the creation of automatically labeled atlases for young children by taking advantage of the availability of specific adult atlases and adapting them to the case of children.

There are several techniques to create an atlas or tackle the various issues that can appear during the creation process.

One method to construct a brain atlas is to use affine registration to generate the anatomy-free reference space and then use non rigid registration to create the "average brain" template [SDM$^+$04]. Apart from creating a population specific brain atlas, one can create a subject specific brain atlas [EAR08]. The main idea is that the similarity (in terms of image, gender, age etc) between the target subject and each subject of the rest

of the population is computed and this information is used as a weighting factor when creating the atlas of the target subject.

Another type of issue could appear during the use of the atlas, and more specifically during the registration process of a new image to the atlas in order to extract atlas information for the specific subject. In order to map brain slices with severe histological artifacts to brain atlases, one can use an automatic method to identify the regions of artifacts and keep only the edge of the "correct" brain perimeter [AXG16]. The estimated edge is then sampled and these points are used as landmarks for point to point image registration with the atlas. The other possibility consists of mapping histological slices of the brain without brain reconstruction from the slices prior to registration since it can create artifacts [XRLH18]. The main problem that needs to be solved is how to find out the orientation used to acquire brain slices. In this approach every histological slice is mapped to the atlas independently. The overall similarity is checked and the atlas is rotated until the angle providing the maximal mapping similarity is found. This method is claimed to have similar or even better accuracy than previous algorithms for this task.

Even though these works are mainly focused on the static brain anatomy, there is also interest regarding the dynamics of the brain and how it evolves across time. For example, an anatomical dynamic brain atlas of the mouse was built by using brain scans of six mice at seven time points. The resulting dynamic atlas has the ability to provide a static atlas at those predefined time points [CMY+11]. The idea of predefined time points was further extended in [CBWJ13] where a multidimensional atlas is presented that includes various contrast levels for every time point in addition to the baseline dynamic information at the predefined time points.

However, using predefined time points during atlas construction can be a limiting factor not only in the data acquisition process but also when studying the brain evolution. To bypass this issue, a method is proposed in [DFBJ10] which uses kernel regression to synthesise samples at any arbitrary time points by using all samples that are close the target time point. Other methods have been proposed like the one in [LJW+12] where first a dynamic model is built for each subject before combining all these models to create the final dynamic atlas space.

Apart from creating anatomical atlases, these methods can be used to create probabilistic atlases to estimate prior probabilities for automatic brain segmentation like in [KMAS+11] where a 4 dimensional atlas is created based on affine transformations and gaussian kernels. Using kernels solves the problem of the dependency between data and atlas time points with the drawback that the resulting atlas time points could have been synthesized from a variable number of data. This may result in differences in consistency and smoothness across the atlas time points. One solution is to improve the normal kernel method and use adaptive kernels instead, as proposed in [SAB+12] which allows the same amount of data samples per synthesised atlas time point to be used.

Given the advancements and the flexibility in the atlas construction techniques, atlas could be a powerful tool for investigating speech production. Earlier studies of speech articulators and especially the tongue, used to be based on histological analyses [Tak01] or tagged cine-MRI of multiple subjects [SDD+01, PPS+07]. Later however, some works exploited the atlas idea to create a motion field atlas of the tongue [XPS+17, WXS+19] for the analysis of the correlation between the activity of tongue muscles [XSG+19].

Dynamic atlases could provide valuable assistance in the study of speech production when the vocal tract geometry changes rapidly, and consequently changes of the area function have a strong acoustic impact [STTN17, THM+06]. In the same conditions they could also improve speech imaging techniques [FWLS16]. Indeed, spatio-temporal atlases are usually based on cine MRI to capture the 3D geometry of the vocal tract and its temporal evolution [WLM+15, WXL+15, WXL+18]. Such approaches rely on the repetition of a specific sentence to create the atlas. The underlying hypothesis is that the subject repeats the same sentence several times in exactly the same way, which requires prior training to speak by following a metronome. Additionally, the resulting atlas frame rate is fully dependent on the cine MRI acquisition frame rate.

In the present work, we propose a method for constructing 3D dynamic atlases of the vocal tract using rtMRI of parallel sagittal planes at a high frame rate, without requiring prior training. The main question addressed is whether it is possible to reduce speakers' inter- and intra-variability by using the atlas space as a standard generic speaker. One of the contributions of our work is to employ the histological atlas creation approach [XRLH18] to collect the 3D information, using rtMRI to acquire data, which offers a high frame rate and reduces the amount of repetitions required by other techniques like cineMRI. Such an approach is new for vocal tract atlases.

Another contribution is the use of the adaptive Gaussian kernel technique to create the atlas samples [SAB+12] with the advantage of making the atlas frame rate independent from the rtMRI frame rate. The proposed method thus gives more flexibility to control the resulting atlas parameters. Therefore the same data can be used to create various atlases with different parameters without the need for new data acquisition every time.

Finally, and this is a determining advantage in studying speech production, the atlas built with this method can be used as a reference speaker to reduce the variability between and within subjects.

Indeed, many works devoted to the production of speech from a general point of view are based on the implicit assumption that an articulatory model built from a single speaker, which is the case of the famous Maeda articulatory model [Mae90], is valid for all speakers. This is a simplification that reduces the scope and validity of much work. In our approach, on the contrary, we have introduced the variability into the construction of the atlas itself, which therefore effectively covers a large speaker variability, provided that the speakers used are sufficiently diverse.

Throughout the paper the atlas thus refers to a specific model for a population of 3D (2D on parallel planes) vocal track dynamic images.

In this work an dynamic vocal tract atlas is generated from rtMRI using the new proposed algorithm and a 4 fold cross validation with histogram matching enable to evaluate whether it is possible to use the atlas space as a generic speaker model in order to reduce variability between speakers.

5.1 Method

Our method for constructing dynamic atlas consists of the following steps:

1) Acquire 2D dynamic rtMRI parallel sagittal planes of the vocal tract during the production of several CVs.
2) Create a subject independent space for silence.
3) Use this space to remove subject's specific anatomical information from the dynamic images.
4) Combine the previously created "anatomical neutral" dynamic images to create the dynamic atlas.

5.1.1 Subjects

Subjects used in this study were four male and four female native speakers of French without any speaking or hearing problems. The average age was 27.25 years with a standard deviation of 4.23 years.

5.1.2 Data acquisition

The data were acquired on Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) located in Nancy Central Regional University Hospital under the approved ethical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053).

For the vocal tract measurements, 3D data was recorded using a multi-slice 2D T2 turbo spin echo (TR = 4610 ms, TE = 100 ms, flip angle = 150 degrees). The thickness of scan slices is 2 mm, and pixel bandwidth is 445 Hz/pixel. Subjects were imaged while having the mouth closed and breathing through the nose.

For acquiring dynamic data, we used a 2D rtMRI sequence. Even though there are 3D dynamic sequences [LZL$^+$19], 2D still offers better spatial and temporal resolutions. In our approach, we used radial RF-spoiled FLASH sequence [UZV$^+$10] with TR = 2.22 ms, TE = 1.47 ms, FOV = 19.2×19.2 cm^2, flip angle = 5 degrees, and slice thickness is 8 mm. Pixel bandwidth is 1670 Hz/pixel. The number of radial spokes is 9, and the resulting image resolution is 136×136. The acquisition time was 44 sec. Images were recorded at a frame rate of 50 frames per second with the algorithm presented in [UZV$^+$10], using a 64 channel head-neck antenna.

To capture 3D information with the 2D rtMRI sequence, we relied on the approach employed to construct brain histological atlases. Since the maximum width of the studied vocal tracts was 40 mm, we used 5 sagittal planes in total, the midsagittal one, two on the left and two on the right, with 0 frame spacing between them.

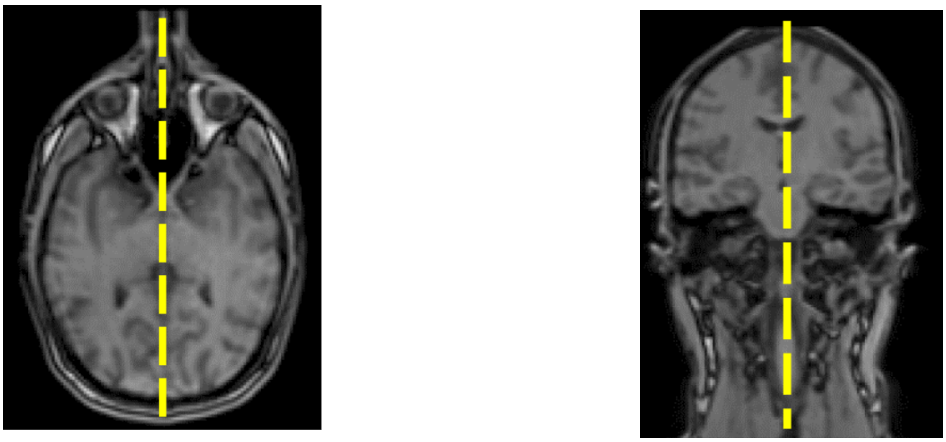
For each subject 5 contiguous sagittal planes (R2, R1, Mid, L1, L2) were acquired covering the whole vocal tract. For each slice the subject repeated the 12 CV syllables at a natural speed as instructed. To help the subject to reproduce the CVs in an identical way through the 5 repetitions, the text of the syllables was projected in the MRI for the duration of the acquisition.

As described in [XRLH18] a major issue when dealing with slices is their orientation, which should be the same for all the speakers.

Care was taken, to ensure the exact sagittal alignment of the midsagittal slice for each subject to avoid misalignement problems previously reported [XRLH18].

A way to solve this issue could have consisted of mapping the slices to an atlas and correct them afterwards. However, to the best of our knowledge, there does not exist such an atlas. Therefore, instead of correcting slices, we tackled this issue one step before, during the real time acquisition step, by using an MRI acquisition protocol designed to be as strict as we could make it to ensure that every time the target sagittal plane (i.e. L2, L1, Mid, R1, R2) was exactly the one being acquired.

The acquisition protocol was chosen to be as short as possible, keeping in mind that it should include a periodic check of the subject's initial orientation and correct midsagittal positioning. The midsagittal plane was defined as the plane which passes in the middle of C2-C3 (in the coronal view) and separates the 2 brain hemispheres (in the axial plane). An optical overview can be seen in Fig. 5.1.



(a) Axial view

Coronal view

Figure 5.1: Definition of the midsagittal plane using axial and coronal view

Alg. 1 gives the overview of the acquisition.

This study focused on 12 CV syllables with C={f, p, s, t} and V={i, a, u}, i.e. /fi/, /fa/, /fu/, /pi/, /pa/, /pu/, /si/, /sa/, /su/, /ti/, /ta/, /tu/. The choice of these syllables was made so that we have two types of consonants, i.e. stops (/p/, /t/) and fricatives (/f/, /s/), two places of articulation, i.e. labials (/f/, /p/) and alveolars (/s/, /t/), in the context of the cardinal vowels (/i/, /a/, /u/). At this point it is important to note that initially we planned to include also the plosive /k/. However, probably due to the supine position in the MRI machine and the force of gravity, some subjects were confusing it with /q/ while speaking and they randomly pronounced either /k/ or /q/ during the acquisition even after proper instructions about the place of articulation. Given the difficulty of some subjects to accurately produce /k/ through all the repetitions, we decided to exclude it.

To prevent co-articulation effects with previous random vocal tract positions, subjects were instructed to close the mouth and breath from the nose before articulating every CV so as to impose the same initial silence position every time. Additionally, the subject was instructed to finish every CV with /p/ so as to impose a minimal anticipatory coarticulation effect onto the vowel.

Algorithm 1 Acquisition scheme

Run a 3D localizer sequence after having comfortably installed the subject in the machine.
targetPlane $\Leftarrow Mid$
setMidPlane :
Acquire 3 groups of 3 slices of the vocal tract. Groups are chosen on perpendicular planes. The midsagittal plane is then defined and a short rtMRI sequence on several perpendicular planes is carried out to verify that the plane is correct.
Acquire multislice 2D images used for measuring the vocal tract.
loop:
Acquire rtMRI data in the targetPlane
Acquire a 3D localizer.
if movement is detected between the localizers goto setMidPlane
targetPlane $\Leftarrow$ next(targetPlane) ▷ The order of planes is Mid, L1, L2, R1, R2.
if targetPlane $\leq R2$ goto loop

We chose /p/ because lips this is the closest articulator to the head coil. The signal is thus stronger and the image quality is very good for this articulator. Consequently the contact between lips which is used as a temporal landmark can be detected with a very good accuracy. Therefore in practice, subjects uttered /sil//C//V//p/.

5.1.3 Vocal tract measurements

A practical way to increase the probability that subjects have different vocal tract sizes, without measuring it directly, is to measure their height before including them in our experimental protocol [RMS09]. The shortest subject was 160 cm while the tallest was 187 cm (average 174cm).

In order to assess ability of the atlas to be used as a standard generic speaker model we measured vocal tract dimensions of included subjects to ensure that there is enough variability in the dataset. Even though methods have been developed previously for speaker adaptation using the length of mouth and pharyngeal cavities [Mae91, BV17], vocal tract/head position [PKSF17] or automatic articulatory landmark extraction [ENRS20] there is no standard method for measuring the vocal tract in terms of height, length, depth since there is no strict definition of those measures due to the complexity of the vocal tract shape, which depends on the position, the articulated phoneme, etc. Therefore we proposed the following method to measure the length and height of the vocal tract. It uses the midsagittal plane and the first step is to draw a line from the outer touching point of the lips towards the anterior lower border of the body of the axis vertebra (see Fig. 5.2).

This line is stopped at pharyngeal wall and we define this length as the length of the vocal tract. The second step is to draw a line, parallel to the previous one and tangent to the palate. The intersection point between this line and the pharyngeal wall is defined as the upper boundary of the vocal tract. The third step is to draw a line from the platform of the vocal folds until the oesophagus. This point at the oesophagus is defined as the

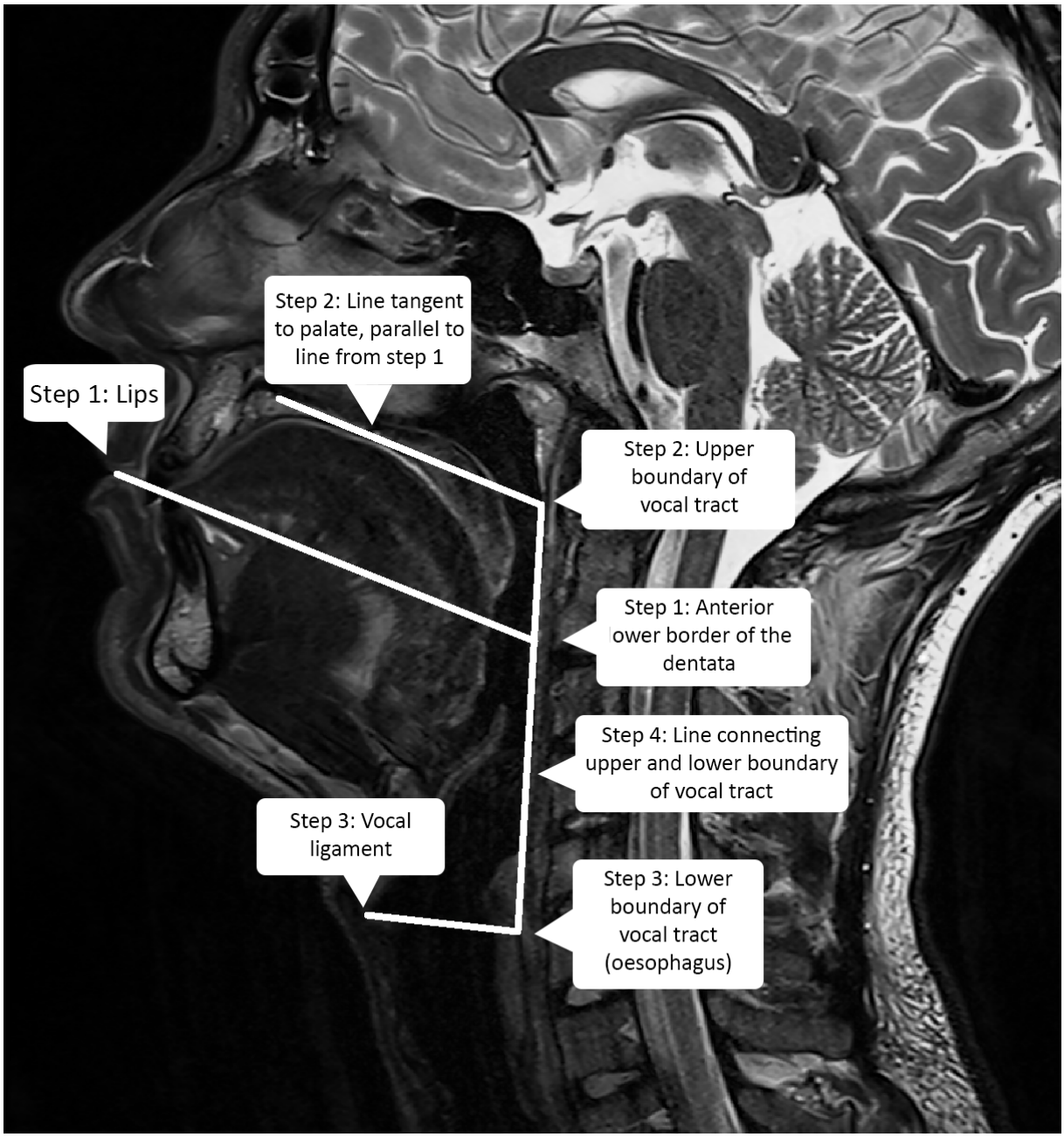


Figure 5.2: Vocal tract measurements algorithm

lower boundary of the vocal tract. The height of the vocal tract is defined as the distance between its lower and upper boundaries (Fig. 5.2). To estimate the width of the vocal tract all the sagittal planes are scanned and the number of planes where the vocal tract is visible at the bottom of the pharyngeal cavity gives the width of the vocal tract.

Table 5.1 shows the measurements for our group of subjects. The difference between the shortest and longest measure is 22 mm ($\sigma = 6.5$ mm) for the vocal tract length and 25 mm ($\sigma = 8.6$ mm) the height, i.e. more than 25 % of these dimensions approximately. For the purpose of our task we thus consider that these sizes exhibit sufficient variability [RMS09]. Fig. 5.3 shows silence frames from all the speakers in the dataset.

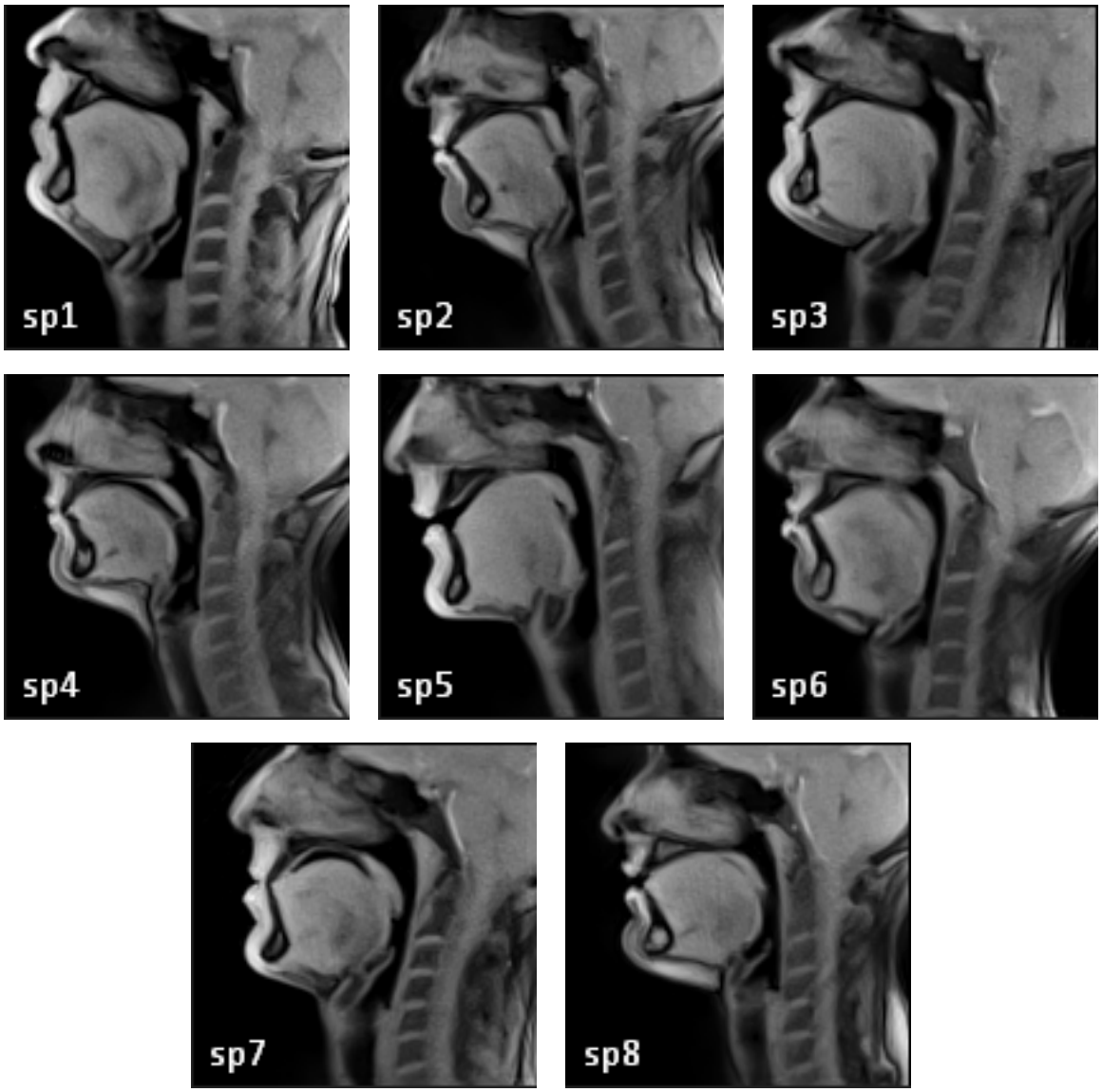


Figure 5.3: Midsagittal (M) frames for silence for all speakers (sp1-sp8 left to right, top down). sp{odd} are male and sp{even} are female speakers

Speaker	Length (mm)	Height (mm)	Width (mm)
SP1	97	92	40
SP2	77	76	32
SP3	99	81	40
SP4	89	69	34
SP5	94	86	36
SP6	87	81	32
SP7	88	90	38
SP8	87	67	34
Mean	89.8	80.3	35.8
STD	6.5	8.6	3.1

Table 5.1: Table of VT measurements

5.1.4 Atlas construction

The acquired dynamic films were manually labeled in order to achieve a better temporal segmentation. Image labelling was done by a person with around 5 years of experience working with this type of image and were then checked by an expert with more than 15 years of experience in the field. For every /sil//C//V//p/ we only kept the /C/ and the /V/ part.

The stop onset is the first image where there is a contact between the tongue tip and teeth for /t/, and pressure between lips for /p/, the vowel onset is the first image where the constriction is released, i.e. there is no more contact between the tongue tip ad teeth for /t/, and no more contact between lips for /p/. The vowel offset corresponds to the first image where lips are in contact because the subjects were instructed to articulate a /p/ after the second vowel. The average duration (number of frames at 50 Hz) per phoneme across all planes and speakers is given in Table 5.2.

syllable	C	V	CV
fi	9	5.65	14.65
fa	8.175	6.475	14.65
fu	7.525	6.9	14.425
pi	6.55	7.275	13.825
pa	7.475	8.55	16.025
pu	6.6	7.625	14.225
si	8.775	5.875	14.65
sa	8.9	6.05	14.95
su	9.025	5.2	14.225
ti	7.6	6.825	14.425
ta	6.85	6.7	13.55
tu	7.025	4.85	11.875

Table 5.2: Table of average phoneme duration (in number of frames at 50 fps)

The proposed algorithm relies on three hypotheses. First, all the slices are in the expected plane. For instance, all the central slices are in the mid-sagittal plane and all the other sagittal slices are shifted from the mid-sagittal plane accordingly. This is a direct consequence of the very strict acquisition protocol we designed, and the anatomical position we chose. As a consequence images of one given plane and speaker can be compared and mapped with the corresponding images of all the other speakers. Anatomical differences between speakers could potentially affect this hypothesis all the more since a potential error can stack as one moves further from the midsagittal plane.

However, we expect this error not to be significant because we moved just two slices away at most from the midsagittal plane and the slice thickness was big enough so that the outer parts of the vocal tract (in the sagittal direction) will lie within the R2 and L2 planes for all subjects.

The second hypothesis in that the order of events is the same for all the speakers, which is expected and reasonable at the scale of an isolated CV.

Third, due to the frame rate of 50 Hz, small piece-wise linear extensions or compressions of the images in time are not affecting significant the dynamics of articulation.

For describing the construction of the atlas silence space, we will refer to the midsagittal plane for simplicity unless it is specified differently. The process presented below for the midsagittal plane is repeated for all the other planes. Before every image transformation or averaging in this work, histogram matching is performed to transform the histogram of the moving image to the one of the reference image. This is intended to compensate for intensity differences between images [SDM+04].

The atlas construction process can be divided into four major steps:

1) Create the anatomically-free reference space.
2) Make dynamic data anatomically free.
3) Align data temporarily.
4) Synthesise the atlas samples.

The objective of **step 1** is to make the data anatomically neutral. By anatomically neutral we mean that data are independent of anatomical variability and correspond to a virtual neutral speaker. For this purpose we used a silence frame during breathing, at a resting position before speakers start recording the CV (as described in the protocol, i.e. breathing from the nose with closed mouth and without any visible articulatory movement) from all N speakers in order to create the reference anatomically free space. The average histogram was computed and all the images' intensities were transformed so that their histogram will match with it. For image registration, the transform used ($T(x, y)$ with x, y being the image coordinates) is composed of two parts, the global and the local one.

$$T(x, y) = T_{global}(x, y) + T_{local}(x, y) \tag{5.1}$$

In our case an affine transformation was used for $T_{global}(x, y)$ and a B-spline transformation for $T_{local}(x, y)$ [RSH+99]. Therefore

$$T_{local}(x, y) = \sum_{l=0}^{2} \sum_{m=0}^{2} B_l(u) B_m(v) \phi_{i+l,j+m} \tag{5.2}$$

where $\phi_{i,j}$ are the uniformly distributed control points of a $n_x \times n_y$ mesh

$$i = *x/n_x - 1 \tag{5.3}$$

$$j = *y/n_y - 1 \tag{5.4}$$

$$u = x/n_x - *x/n_x \tag{5.5}$$

$$v = y/n_y - *y/n_y \tag{5.6}$$

and B_l, B_m is the lth and mth B-spline base function. Each image was registered to all other $N-1$ images using the described non-rigid B-spline based transformation using the image_registration function of the MATLAB toolbox "B-spline Grid, Image and Point based Registration"[mat].

This toolbox was used for all the transformations performed in this work. For every image we get $N-1$ transforms. The average transformation (without any further weighting) is computed for every image and this average transformation is applied to the corresponding image to produced the anatomical free version of every image which is image dependent. Finally all the N image dependent anatomical free spaces are truly averaged to create the final reference space (image independent, anatomically neutral).

More precisely, for the ith silence image from the set of silent images $\{I_{1\ldots n}\}$ the transformations $T_{i,j}, i \neq j$ are computed and averaged to give the average transformation $\overline{T}_i = \frac{1}{N-1} \sum_{j=1..n, i \neq j} T_{i,j}$. Finally, the final reference space is created by applying the $\overline{T}_i$ transforms to the corresponding images and averaging them $\overline{I} = \frac{1}{n} \sum_{i=1\ldots n} \overline{T}_i(I_i)$ with $\overline{T}_i(I_i) \simeq \overline{I}_i$. A visual representation can be seen in Fig. 5.4.

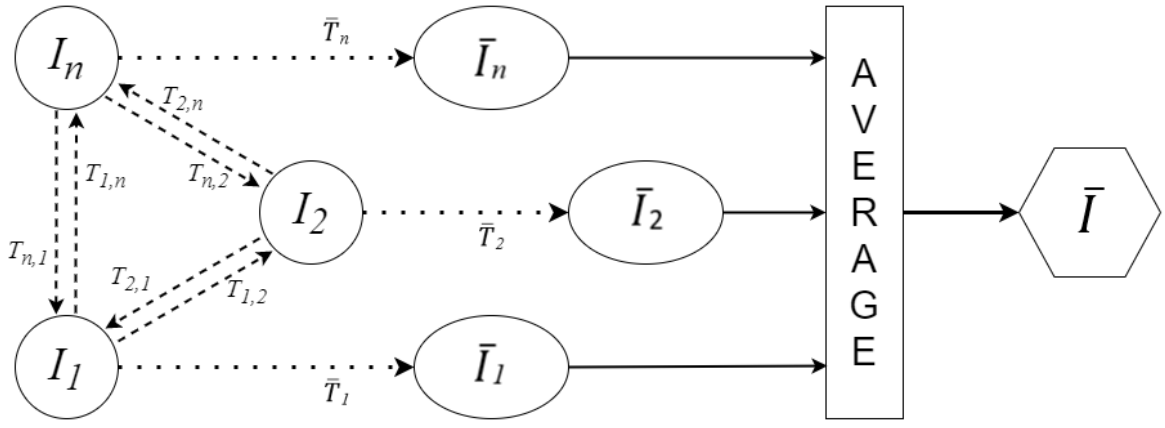


Figure 5.4: Creating the reference space. Every ith silence image is registered to all others, the computed transformations are averaged to give $\overline{T}_i$ and applied to the ith image to get $\overline{I}_i$. The resulting images are averaged to get the final reference space image $\overline{I}$

Step 2 is intended to make the data anatomically free. First, the images' histogram of all the CVs is matched with the histogram of the reference and the image then transformed to the reference space using only an affine transformation (computed with the same MATLAB function as in Step 1)

because it transforms the anatomy of the data to the reference anatomy but keep the vocal tract position variability, i.e. the position of the articulators [KMAS$^+$11].

Step 3 is intended to process the anatomical free data for applying the adaptive kernel technique. For each CV, all the planes of all the speakers were used to specify the corresponding average C, V and CV duration. These values are set as the time duration of the atlas. Data are then piece-wise linearly aligned to the previous average values using rtMRI frame rate to pass from the frame space to the time domain in order to compute the global time.

For example, in order to align a CV (to be modified) to a reference CV, the C and V parts of the modified CV are independently and linearly extended or compressed until the duration of both C and V of the modified CV match with those from the reference CV. This alignment technique (see Fig. 5.5) is intended to achieve time alignment so as the duration of the modified (Mod) CV is that of the reference (Ref) CV, but not to map frames of the reference CV to those of the current CV.

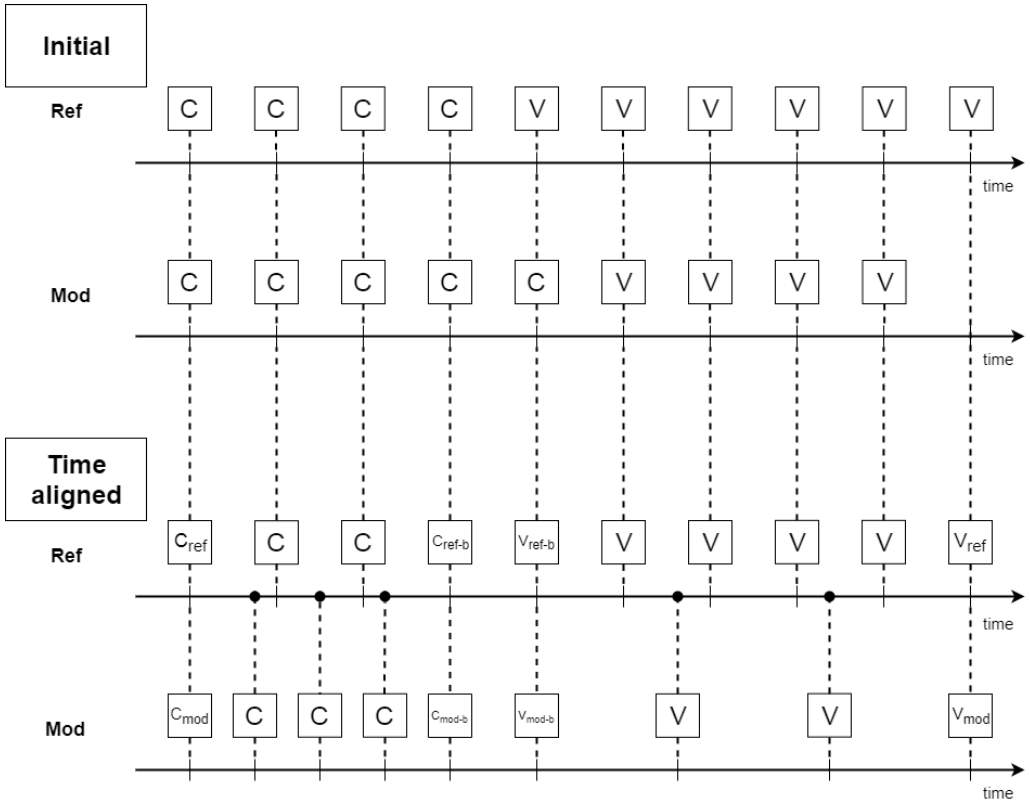


Figure 5.5: Piece-wise time alignment. Mod is the CV which duration is to be modified in order to match the duration of the reference (Ref) CV. On the top are both CVs before time alignment (Initial) and on the bottom the time aligned version of the Mod CV with the Ref CV

In practice, this procedure creates one anatomical free image series from the image series of all speakers, by putting all frames in a global time scale based on the time stretching or compressing defined by the piece-wise linear alignment. It should be noted that the resulting series may have multiple frames at one time point and that samples are not homogeneously distributed across time.

Step 4 consists of synthesising the atlas images from the global series of images, i.e. the 12 CVs involved in this work, by using the adaptive gaussian kernel method [SAB+12].

The core idea is to generate the atlas image at a given target time point from k images in the global series located in the vicinity of the target time point. k is a pre-specified number of samples to choose the closest relevant samples and the resulting image is the Gaussian weighted average of the k samples.

The advantages are that the atlas frame rate is independent of the data acquisition frame rate and that the atlas sampling may not be regular since the time points can be chosen freely. Theoretically, the initial sampling rate has some influence, but the initial frame rate is high enough to study all common speech tasks [LSMN16]. However, the number of samples used to synthesise the images and the parameters of the Gaussian weights should be tuned. In [SAB+12] the number of samples was chosen as a function of the number of subjects available in the vicinity of a target time point and could vary substantially, i.e. from 3 to 25, because the number of subjects recorded depended on time and the phenomenon monitored was much slower. Thus, when many subjects were available the gaussian was sharp, and conversely wider when fewer subjects were available. In our case the number of subjects is constant, i.e. 6, and consequently the number of samples available is almost constant if we consider that the dynamic variability is limited. We tested several choices and set k to 7 atlas samples within a window of 20 ms, which is the recording period and is expected to be sufficient for our study [LSMN16]. The Gaussian weighting was designed so that its mean value is the selected time point τ to be synthesised and the standard deviation was tuned so that the weight of the farthest k sample τ_f from the center is 0.35 of the maximum value of the Gaussian distribution. Therefore the parameters of the Gaussian distribution is $\mu = \tau$ and $\sigma = \sqrt{-(\tau - \tau_f)^2/(2 * ln(0.35))}$ [SAB+12]. A visual representation can be seen in Fig 5.6.

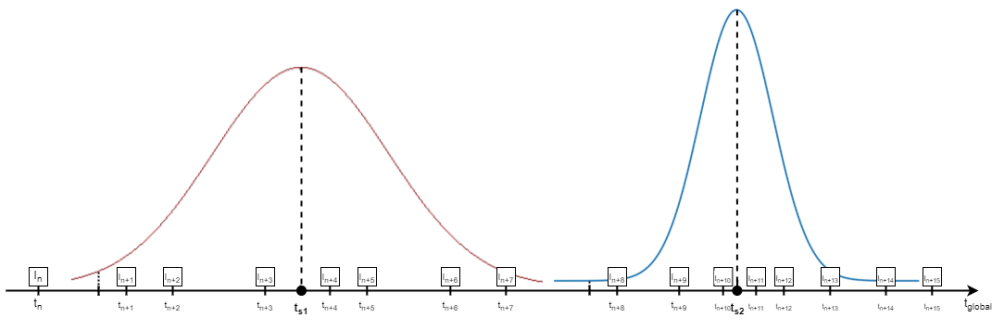


Figure 5.6: Adaptive Gaussian kernel technique. The width of the Gaussian is adapted based on the distance between the desired synthesis time points (t_{s1}, t_{s2}) with the available samples I_i. The number of the samples contributing to frame generation is stable

To evaluate the results, 4 fold cross validations were carried out using 6 subjects for training and 2 subjects for test for every fold. In every fold the two test subjects were chosen to be of different gender to get results for both genders. Both of the test CVs are piece-wise linearly temporally aligned with the corresponding atlas CV. For each frame of each atlas CV the temporally closest frame of the corresponding test CV is selected. It is thus possible for a test frame to be used more than once while some others may not be used at all. At this point, for every atlas frame there is one linked frame for every test CV (two links per frame per CV, one for each test subject).

All the frames linked with the same atlas frame form a stack of images as seen in Fig. 5.7. Each stack includes an atlas image and the corresponding images of: speaker 1 image without registration, speaker 2 image without registration, speaker 1 image after registration, speaker 2 image after registration. Examples of every stack image at the midsagittal plane can be seen in Fig. 5.9.

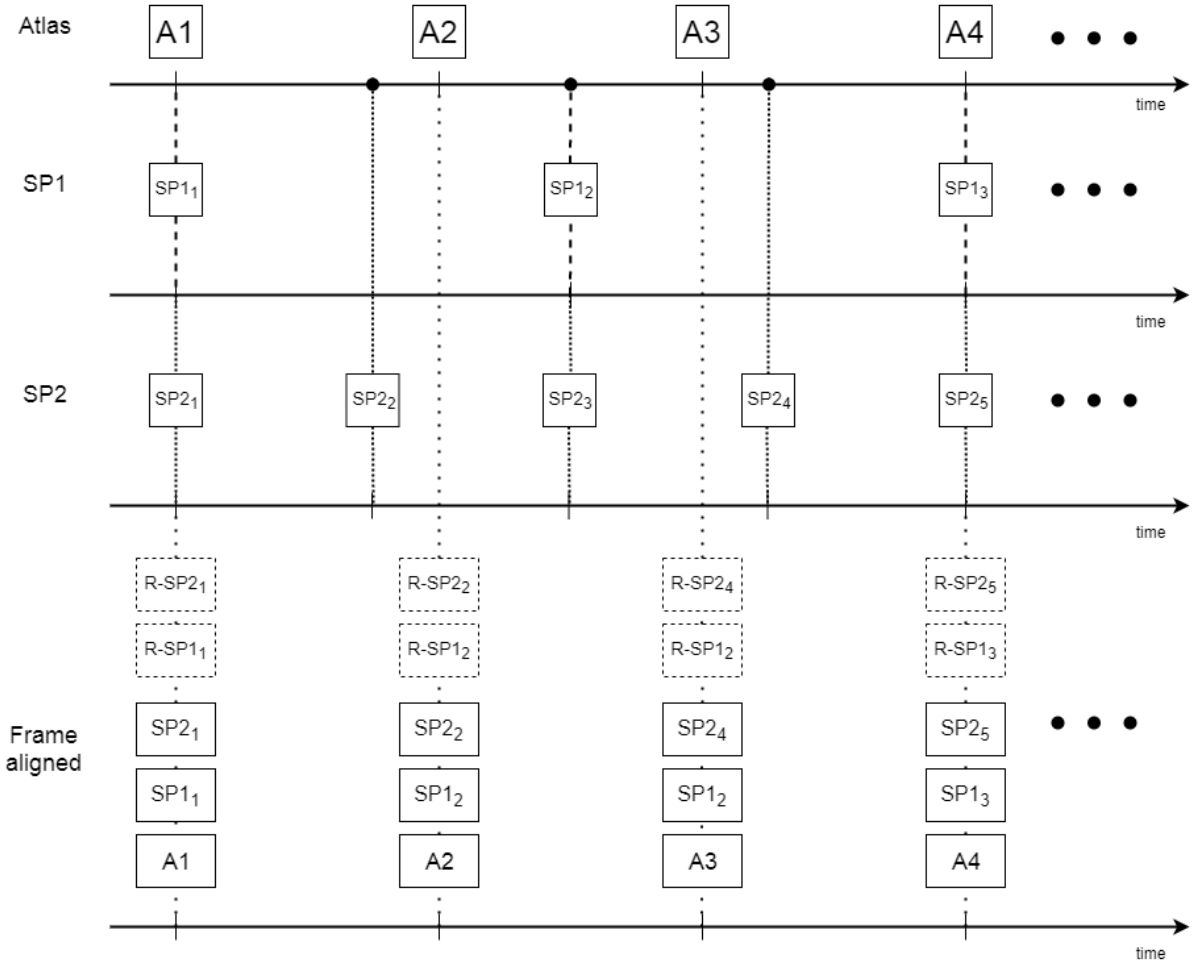


Figure 5.7: Frame alignment used for tests. A represents the atlas frames and SPi_j original frames j for speaker i and $R-SPi_j$ the registered framed within the atlas space.

Histogram matching is applied so that the histograms of the linked images fit that

of the corresponding atlas frame. Test images are mapped to the atlas image using the B-spline non-rigid transformation (the same technique as that used for construction).

The idea of this procedure is to transform any given image of a target speaker CV as close as possible to the corresponding atlas image. We used cross correlation as a similarity measurement between images mapped from the atlas and original images[SAB+12]. The cross correlation value is normalized by the auto-correlation of the atlas frame.

More precisely, for each stack of images A be an atlas image, O_1, O_2 the original images of speaker 1 and speaker 2, and R_1, R_2 the corresponding registered images to the atlas. All images represent $M \times N$ matrices of pixel density values. For the similarity measurements before the use of atlas (BA) for these frames we have:

$$BA = \frac{\max \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} O_1(m,n) O_2(m-k, n-l)}{\max \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A(m,n) A(m-f, n-g)}, \tag{5.7}$$

with
-(M-1)$\leq k, f \leq M-1$
-(N-1)$\leq l, g \leq N-1$
The similarity measurement after the use of atlas (AA) can be computed from:

$$AA = \frac{\max \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} R_1(m,n) R_2(m-t, n-c)}{\max \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A(m,n) A(m-f, n-g)}, \tag{5.8}$$

with
-(M-1)$\leq t, f \leq M-1$
-(N-1)$\leq c, g \leq N-1$
These measurements are averaged across space and time in order to produce Table 5.3. Columns 2 and 4 are the computed averages (of BA and AA respectively) and column 3 and 5 are the corresponding standard deviations.

5.2 Results

The presented methods above were applied to the acquired data on all 5 planes. Small time variations on the required atlas construction time appeared during the various registration processes due to the fact that by nature some speakers are anatomically more similar/different between them. Fig. 5.8 present examples of frames from all sagittal planes in the atlas space for /tu/. The visual assessment of the synthesised images show that they represent the natural vocal tract position with the expected dynamics. This visual conclusion is further supported by the numerical results of Table 5.3.

Images from the top of Fig. 5.9 are the reference images of atlas. Images from the second and third line of Fig. 5.9 are mapped to those of the first line and the resulting images are shown in Fig. 5.9 on fourth and fifth line. Similarity between the original images (Fig. 5.9 second and third line) for all frames of all planes are computed. The same computations applied to the transformed images after atlas registration (Fig. 5.9 fourth and fifth line) to evaluate if the similarity has increased.

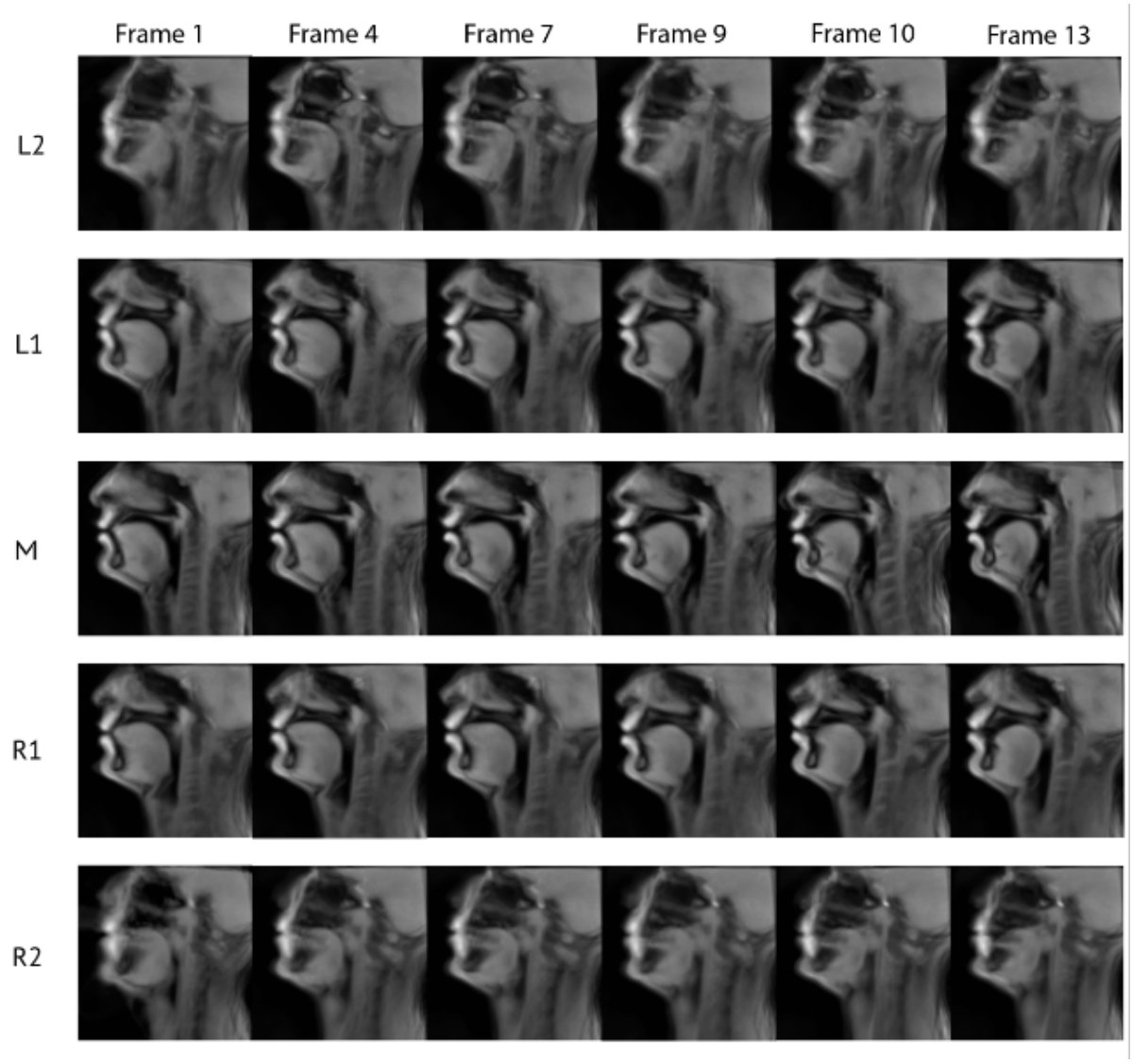


Figure 5.8: Frames 1, 4, 7, 9, 10, 13 of the all atlas planes without sp5, sp6 for /tu/

As can be seen from Table 5.3, the average similarity between the images after the use of atlas is increased with smaller standard deviation (column 4, 5) compared to the similarity and the standard deviation without the atlas use (column 2, 3).

Fig. 5.9 show the midsagittal frames of the atlas with the corresponding test subject frames before and after transformation with the atlas. The places of articulation are clear for both /t/ and /u/. We can see the dynamics of the tongue starting from the very beginning of /t/ where the tongue presses the hard palate up until the end where the tongue is lowered for the production of /u/. Fig. 5.8 show the temporal evolution of the articulator positions in the five planes. For example, by visually comparing the tongue position between midsagittal and adjacent planes (e.g. frame 9), one can notice that the tongue is lower in the midsagittal plane near the teeth region. Additionally, for most of the images of R1 and L1 planes lips are almost closed, in contrast to the midsagittal plane

Figure 5.9: The midsagittal frames of the atlas with the corresponding test subjects frames before and after transformation with the atlas

where they are clearly open. This kind of information is important for studying speech production and cannot be derived from the midsagittal frames alone. The results of the normalized image similarity before and after the use of atlas are presented in Table 5.3.

5.3 Discussion about generic speaker model

Images of the R2 and L2 planes are more blurry compared to the other planes due to the fact that the original images of the speakers at that plane suffer from a "partial volume effect". Indeed the slice thickness is 8 mm and when moving away from the midsagittal plane, the volume of one pixel may correspond to a mixture between more than one type of tissue (muscles, fat, teeth) and air, which give rise to some blurring (see Fig. 5.10).

phoneme	Mean (before)	STD (before)	Mean (after)	STD (after)
fi	0.872	0.044	0.975	0.014
fa	0.876	0.047	0.976	0.014
fu	0.869	0.043	0.974	0.015
pi	0.874	0.044	0.976	0.015
pa	0.874	0.046	0.975	0.014
pu	0.873	0.040	0.974	0.015
si	0.872	0.044	0.975	0.014
sa	0.870	0.044	0.974	0.019
su	0.873	0.045	0.976	0.016
ti	0.873	0.046	0.974	0.016
ta	0.877	0.048	0.976	0.016
tu	0.874	0.044	0.975	0.021

Table 5.3: Table of cross validated results. From left to right: CV, average similarity score before the use of atlas, standard deviation of the average similarity before the use of atlas, average similarity after the use of atlas, standard deviation of the average similarity after the use of atlas

However, one can still extract useful information about the movement of some articulators like the tongue body.

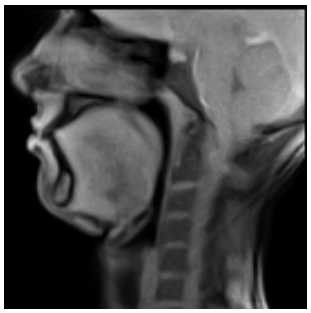
Figure 5.10: Original L2 frames during /u/ for speakers 6-8 (left to right). One can notice that images in this plane are a bit more blurry compared to the midsagittal plane (Fig. 5.9 second and third line)

By comparing the atlas images against the individual subject's images, one can notice that atlas images are less sharp. This could due to histogram matching that took place before every image transformation, or to the initial histogram matching of all the silence frames with their average histogram. It could also be due to the interpolation kernel during the spatial transform or because of the image averaging procedure both during silence creation and during the atlas sample generation. However, the main reason is that when silence is created the resulting reference silence image presents some loss of sharpness due to different anatomies and head positions. Even though silence does not

look strongly connected with the final atlas synthesised images, any loss of sharpness could further propagate until the end of the atlas creation process since on the one hand it was used as a reference to match the histograms and on the other hand it was used to transform all the dynamic data of all subjects in order to remove subjects'anatomical information and create "anatomically neutral" dynamic data.

The second noticeable point is that the spine is not very sharp in some cases for two reasons. This region is also affected by the general loss of sharpness but the main reason is that posture and anatomical differences between subjects, especially between males and females result in that more vertebra are visible for some subjects and less for others (see Fig. 5.11). This probably affects the transformation algorithm since these extra vertebra have no place to be directly mapped. They are therefore compressed, or extended in the opposite case, within the spine. However, we can see that the main articulators like the tongue are not strongly affected. Even if there is no objective criterion that specifically focuses on the articulators since every image was treated as a whole this behaviour was expected because all the images contained the whole vocal tract and thus its impact is indirectly stronger on the transformations computed compared to some vertebra (C6) that sometimes appears and sometimes not.

(a) Silence of speaker 5



(b) Silence of speaker 6

Figure 5.11: Silence frames for two speakers. One can see that more vertebra are visible for speaker 5 (a) compared than for speaker 6 (b)

The potential uses of the atlas concern the study of the dynamic 3D area function [THM+06] since it allows the use of one representative subject, i.e. the atlas, instead of either one random subject or repeating the process for multiple subjects and then processing results to draw general conclusions. Another use of the atlas concerns the transformation of 2D rtMRI videos into 3D dynamic videos [DTI+19] since the atlas incorporates the real 3D dynamic information that occurs during the production of continuous speech, and not just estimates it from from static 3D and midsagittal rtMRI.

Automatic tracking of the vocal tract contours[LBV+18, TGH+19] could also take advantage of the atlas to map a specific subject data whose data have to be delineated. The main advantage is that once the atlas is created, it could be used to process new rtMRI data without requiring every time data pre-processing, retraining models etc. Finally, the main contribution of this work is that the atlas is a true golden speaker which embodies speaker independent articularoty gestures.

The future objectives of this work are the construction of a more generic model including more speakers and languages so as to create a more universal atlas and exploring the minimal set of transformations that would enable any speaker to be copied from the atlas. Creation of more CVs or even extension to CVC, VCV or small words could also be interesting in order to study further co-articulation effects. Finally, one can envisage the use of the atlas in order to create models of the vocal tract area function and how it can be estimated from the midsagittal plane, a topic with various applications mainly in the field of articulatory speech synthesis.

Chapter 6

Discussion

In this Chapter we are going to 1) summarise the main contribution of this thesis 2) present some interesting questions that appeared during the "journey" of this thesis but we chose to follow other directions and focus on other problems therefore they left unexplored. Potential solutions on how we would have planned to tackled these problem in case that we moved towards their direction are also proposed 3) propose future directions that one can follow in order to continue the research journey established in this thesis and go further.

6.1 Contributions of thesis

This thesis initially presents some of the common ways to acquire articulatory data. Among them, MRI looks the most promising one therefore the major MRI articulatory databases are presented and their potential uses are discussed. The relevant new articulatory database for French language is then extensively described since it is the one that is used for a big part of the studies in this thesis. Various algorithms were developed and several experiments were conducted which can be mainly divided into three parts. The first part is about the study of the effect that the velum, the epiglottis and the position of the head have on speech production of /a, oe, i, o, u/ vowels. Their effect was studied by manually segmenting and processing static MRI data and using acoustic simulations. Simulated and original data were examined to asses the effect of the previously mentioned parameters based on the differences that we noticed in the formants. The second part is dedicated on developing algorithms which will enable to overcome some of the technical constraints that still exist on the current MRI acquisition technology. The proposed algorithms in this part include a way to post process midsagittal rtMRI videos of the vocal tract in order to improve image quality, a couple of approaches to generate the corresponding 3D dynamic vocal tract shapes as well as a way to synthesise pseudo rtMRI data from silent frames during CV production. All these algorithms were based on image transformations that treat each image in total, therefore no delineation of contours or other (hard to acquire) information from the image is needed prior to the application of the algorithms.

In the last part of this thesis an algorithm that generates the 3D dynamic data of

a standard speaker during CV production is presented. For the development of this algorithm, a special MRI data acquisition protocol was developed, new MRI data were acquired and a procedure of measuring vocal tract dimensions based on anatomical landmarks was described. These data generated by the proposed algorithm can be used as a standard reference space (for example to reduce subjects' variability) since they are speaker independent. The proposed generic speaker model could also be used as a base for several other studies like articulatory modelling, articulatory speech synthesis etc opening a new approach of how variability is handled between subjects.

To sum up, the contribution of the thesis is to provide an insight of the effect of simplifications of the vocal tract geometry, approaches to overcome limitations of the current MR imaging technology and an algorithm that enables to incorporate information about speaker variability within the model construction process.

6.2 Selection of unexplored research questions

Before presenting a set of unexplored questions we are going to recap our main findings. In Chapter 3 our results showed that head position has a significant effect on speech phonation which is in agreement with findings in the literature about changes in the upper airway dimensions based on neck position [JMD94]. Regarding the effect of velum and epiglottis on vocal tract geometry our results show that there is some impact on the formants. which is further supported from additional studies [CHK$^+$19]. In Chapter 4, the results of our work in generating images of various vocal tract shapes across time and space show that it is a quite competitive task that many issues appeared that had to be solved. Even though the related literature is limited, there are clues at the same direction as our results, that show that 3D dynamic vocal tract shape generation could be possible [ZKP$^+$12]. Finally, in Chapter 5 the proposed algorithm gives more flexibility compared to other approaches [WXS$^+$19] when generating the vocal tract of a global speaker model, as it makes the temporal resolution of the generated data independent from the data acquisition frame rate.

Having summarised the main results of this thesis, we are going to present some of the future directions that worth considering. An interesting question that appeared during the work of simulations is how these results can be applied to the field of speech synthesis. There are various approaches that one can use the results of vocal tract simplifications or head positions to speech synthesis. One way that the results from this work can be directly used could be to modify accordingly the formants of the synthesised speech sounds based on the impact that simplifications or head positions have to the original sound. This can be seen as a post processing step from a synthesis point of view. Another way could be to use those results indirectly, from an automatic control systems point of view, by trying to implement a controller to add at the electric part of the electro-acoustic analogy that will make the electric circuit to account for changes in the head position for example. Even though the first way looks easier, the second looks more robust since it manages the head position differences inside the method and not outside as post processing.

Another worth-mentioning question that was inspired from the several types and ways of vocal tract shapes generation is the application of these algorithms directly to whole

words or phrases. Although the extension of these algorithm to work on phrases may look straightforward, some problems are expected to appear whose solution may be a little tricky. For example, if the presented algorithm for image improvement is directly applied to a word, discontinuity of the images will appear between some of the diphthongs. A way that one can try to solve this could be to further post process the generated sequence by calculating the deformation fields from each image to the next one and apply high smoothing factor on them and then regenerate the images with the globally smoothed recalculated deformation fields. Small blurring of the images as a pre-processing step could also be useful for both image continuity and capturing the better the contact of the articulators. It should be probably also followed by a sharpening filter at the very end of the post processing step.

Finally, after the atlas construction a question that appeared is whether it would be possible to create an atlas in a similar way but with isotropic resolution [LSM+17]. One approach could be to acquire data also in coronal axial parallel planes and by properly combining them recreate the 3D isotropic dynamic vocal tract shape per speaker. The previously presented algorithm can be directly used with slight modification. Great care should be given to the field of view of the acquisition in order to acquire in all planes a broader region of the vocal tract so that the final 3D generated isotropic image will sufficiently larger than the vocal tract and the articulators. This will keep the region of interest of the vocal tract and articulators intact from compressions or extensions that would appear to the atlas images because of image transformation between speakers with different anatomy. In contrast to the creation of brain atlases [SDM+04] where the brain is segmented from the skull, in this approach segmenting the vocal tract from the rest of the head is not necessary, therefore reducing the required data preprocessing time. One should also keep in mind that the cost of image transformations in this case is significantly higher compared to the 2D approach.

6.3 Directions to expand this thesis

One possible direction that one can expand the work of this thesis is to use the atlas in order to implement all the algorithms presented in Chapter 4. This approach will make 2D to 3D extension algorithms and pseudoMRI data of the vocal tract synthesis speaker independent since the core part of the algorithm will be based on the atlas which is a generic model and therefore it can be used as a reference speaker.

Another possible way could be to use the atlas to create automatic segmentation or delineation algorithms of the vocal tract by adding this information to the atlas and then use transformation algorithms in order to map for example the atlas segmented regions of the vocal tract to the vocal tract of a specific subject [GRH+08].

Some more optimistic directions to continue the work of this thesis is to build articulatory models in the atlas speaker and the try to adapt them to a specific subject or synthesise the voice of the atlas space and then apply audio to articulatory inversion [OL05]. These ideas could be combined and by using articulatory speech synthesis techniques, one can think of an indirect way to apply voice conversion between two speakers [KKTH19] by adapting an articulatory model of one speaker to another one through the

atlas in a two-step procedure. This approach could be further extended to the even more challenging but very interesting problem of language translation. One can envisage the creation of two atlases, one for the source language and one for the target language. By creating articulatory models in both atlases and being able to adapt from one to another, it could be possible to adapt the model of the target language to the source language and then to the speaker in order to acquire a model of the speaker speaking the target language, which will be used to synthesise the voice of the speaker speaker the target language by using articulatory speech synthesis techniques.

Chapter 7

Résumé détaillé en français

7.1 Introduction

Le premier intérêt de la production de la parole est probablement les applications médicales. Les personnes souffrant d'une pathologie de la parole qui a besoin une sorte d'opération visant à retirer une partie du conduit vocal ou à ajouter un implant artificiel peut bénéficient grandement de la modélisation articulatoire. Elle peut guider le chirurgien sur la façon de pratiquer l'opération et où et comment "couper" la partie à supprimer afin que l'orateur garde la plus grande partie possible de sa capacité à parler. Dans le cas des implants, ils peuvent être réglés et ajustés hors ligne pour la de manière optimale.

La synthèse et la conversion de la voix est un autre domaine qui pourrait potentiellement bénéficier des progrès de la production de la parole la recherche. Les personnes ayant subi une laryngectomie, par exemple, peuvent avoir une voix artificielle réaliste grâce aux techniques de synthèse de la parole. Les connaissances en matière de production de la parole peuvent améliorer encore les résultats des la voix synthétique à différents stades comme l'utilisation des informations articulatoires de l'orateur cible pour le rendre sa voix originale de manière artificielle, ou en guidant la modélisation des articulateurs en fournissant des informations sur leur importance en fonction d'un son prononcé par exemple. La synthèse vocale joue également un rôle important dans le monde futur avec des maisons intelligentes et des appareils qui communiquent avec l'utilisateur par la parole.

La production de la parole peut améliorer les systèmes de reconnaissance vocale en fournissant des informations sur la la nature du discours pendant la phase de formation. Ceci est particulièrement utile dans les situations de bruit extrême conditions comme dans la plupart des applications de la vie réelle, comme donner des commandes au téléphone pendant la conduite ou à des appareils intelligents à la maison. Un autre exemple est le pilotage d'un avion de combat où le bruit est très important, notamment dans les hélicoptères, et où la précision de la reconnaissance vocale est nécessaires à une communication efficace et à la détection de la fatigue.

La connaissance de la production de la parole peut également aider les personnes qui tentent d'apprendre une nouvelle langue en fournissant des informations précises sur la manière dont les phonèmes sont censés s'articuler et comment placer les articulateurs afin

d'obtenir une bonne production orale. Discours Les thérapeutes peuvent également tirer profit de ces connaissances en étant capables de visualiser et de mieux guider les personnes qui bégaient afin d'améliorer plus rapidement leur communication.

L'objectif principal de ce sujet de doctorat est d'étudier la production de la parole à l'aide d'images IRM. Plus précisément, les objectifs de cette thèse peuvent être résumés comme suit :

- Utiliser les données statiques de l'IRM du conduit vocal et les simulations acoustiques pour explorer l'impact des articulateurs de la parole, les simplifications géométriques et des positions de tête sur la phonation de la parole.
- Travailler sur des algorithmes pour l'extension spatiale et temporelle de la forme du conduit vocal, c'est-à-dire transformer les images dynamiques 2D de l'IRM misagittale du conduit vocal en images dynamiques 3D de l'IRM du conduit vocal ou estimer la évolution temporelle de la forme du conduit vocal pendant la CV (voyelle consonne) production en utilisant des cadres de silence.
- Travailler sur un algorithme permettant de créer un atlas dynamique 3D du conduit vocal des CV qui servira de pour décrire la dynamique de la parole dans les domaines spatial et temporel.

7.2 Bases de données

Dans ce chapitre, nous présentons les bases de données IRM actuellement disponibles, décrivons leurs limites et expliquons la nécessité d'en créer une nouvelle adaptée à notre étude. Il existe plusieurs bases de données qui offrent une grande quantité de ressources pour les études de production/synthèse de la parole. Cependant, la plupart d'entre elles sont destinées à l'anglais et ne comprennent le matériel traité des données, comme les contours articulatoires, qui sont nécessaires pour permettre l'étude des phénomènes de production de la parole. En outre, la qualité de l'image est assez faible par rapport à la qualité que l'on peut attendre des approches plus récentes. Pour ces raisons, ArtSpeechMRIfr a été choisi pour ce travail car il fournit des images RM de haute qualité et du matériel traité pour le français.

ArtSpeechMRIfr est une base de données qui comprend des images RM en temps réel et statiques des voies vocales. La base de données contient également des données traitées : la parole débruitée, des annotations phonétiquement alignées, des contours articulatoires et des informations sur le volume des voies vocales. Tous ces éléments constituent une riche ressource pour la recherche sur la parole. La base de données est construit à partir de données provenant de deux hommes parlant le français. Il couvre un certain nombre de contextes phonétiques dans la partie contrôlée, ainsi que la parole spontanée, les balayages IRM 3D des articulations vocales soutenues, et des moulages dentaires des sujets. Le corpus pour l'IRMr est constitué de 79 phrases synthétiques construites à partir d'un dictionnaire qui minimise la durée des acquisitions tout en gardant une très bonne couverture des contextes phonétiques qui existent en français. L'IRM 3D comprend des acquisitions pour 12 voyelles et 10 consonnes françaises, chacune d'entre elles

étant prononcée dans plusieurs contextes vocaux. Les contours articulatoires (langue, mâchoire, épiglotte, larynx, vélum, lèvres) ainsi que des volumes en 3D ont été dessinés manuellement pour une partie des images.

7.3 Simulations acoustiques

L'inspiration principale du travail de ce chapitre provient de certaines questions qui apparaissent principalement lors de la création de modèles statistiques pour la synthèse vocale articulée.

- Premièrement, est-il nécessaire de utiliser la forme 3D des voies vocales dans tous les cas ou le plan mi-sagittal est suffisant ? Il est important de savoir si le problème peut être traité dans le domaine 2D uniquement.
- Deuxièmement, comment la position de la tête du locuteur pendant l'acquisition des données peut affecter la phonation et comment les modèles articulatoires peuvent s'adapter à de tels changements ? C'est important car il y a une variabilité entre plusieurs séances d'enregistrement IRM en particulier.
- Troisièmement, est-il possible de simplifier la géométrie des voies vocales afin de réduire les paramètres de l modèle ? Ou, de manière équivalente, faut-il tenir compte des épiglottites et des petites cavités ?

Pour tenter de trouver la réponse à ces questions, nous avons utilisé les données de l'IRM des voies vocales correctement traitées et modifiées afin de mener nos expériences à l'aide de simulations acoustiques.

En ce qui concerne la première question, la première remarque concerne les sons produits par le locuteur dans l'appareil d'IRM. Les valeurs mesurées sont un peu éloignées des valeurs attendues. Cela est particulièrement vrai pour le premier formant des voyelles proches /i,o/ qui est plus élevé en fréquence. Une explication pourrait être que la cavité pharyngienne est plus petite en raison de la posture du sujet. L'examen visuel des images montre une articulation légèrement décalée dans certains cas. Cependant, cela pourrait également être dû à la difficulté de mesurer les formants dans le signal débruité, surtout lorsque F1 est petit. Ensuite, il existe un bon accord entre les résultats des simulations 2D/3D et les formants F1 et F2 déterminés à partir du signal vocal enregistré. Cependant, pour F3, la simulation 3D s'avère donner des résultats plus proches de ceux de la parole naturelle que ceux de la simulation 2D, probablement parce que le volume 3D donne une géométrie plus proche de la réalité.

Concernant la deuxième question, l'examen visuel des images montre que la position haute augmente le volume de la cavité dorsale correspondant au pharynx mais réduit sa longueur. A l'inverse, la position basse entraîne essentiellement une modification de l'angle entre les deux cavités, et une diminution de la taille de la cavité antérieure mais ne modifie pas de manière significative le volume de la cavité pharyngienne. Il est à noter que les variations des fréquences des formants entre la position neutre et les deux autres positions sont confirmées par des simulations acoustiques en termes de direction, même

si leur ampleur est parfois différente. Ce dernier point peut s'expliquer par le fait que les simulations numériques sont bidimensionnelles et que les formants ont été mesurés avec quelques difficultés dans le discours bruyant.

En ce qui concerne la troisième question, la première remarque concerne les valeurs des fréquences des formants du conduit vocal normal sans simplification. La position allongée et le bruit dans l'appareil d'IRM expliquent en grande partie les écarts par rapport aux valeurs attendues pour ces voyelles d'un locuteur masculin. En ce qui concerne les simplifications, il faut noter qu'elles n'ont pas un impact très significatif sur le premier formant. Pour F2, la simplification du vélum a un effet plus prononcé. Les changements au niveau du vélum ont un impact sur la constriction entre les cavités avant et arrière du conduit vocal. Cela affecte principalement le F2 qui est plus sensible à la longueur des cavités.

7.4 Transformation 2D à 3D

La motivation des travaux proposés dans ce chapitre est la nécessité de combiner les points forts de l'IRM et de l'IRMtr. L'IRM a une bonne qualité d'image et des informations sur le volume ; l'IRMtr, en revanche, a une bonne résolution temporelle. Pour atteindre un tel objectif, nous devons apprendre à surmonter leurs faiblesses, c'est-à-dire comment améliorer les images de l'IRMtr avec les connaissances que nous avons des acquisitions statiques et bien contrôlées de l'IRM, afin de fixer leurs artefacts, et comment augmenter l'image qui est limitée à deux dimensions afin qu'elle devienne volumétrique.

L'objectif est d'aborder les deux questions pour les syllabes de consonnes-voyelles (CV), en prenant les séquences 2D de l'IRMtr CV ainsi que les captures 3D correspondantes de l'IRM *C(V), V*. Notre stratégie est basée sur l'hypothèse que la première trame du CV dynamique aura une articulation similaire à celle de l'IRM du C(V) 3D et, de même, le cadre final aura une articulation similaire à celle du cadre mi-sagittal du V 3D. En utilisant ces deux cas comme points de départ, nous créons deux versions de l'estimation dynamique 3D. L'idée principale est (pour le cas de la consonne) 1) de calculer les transformations d'image entre la tranche midsaggitale et les autres tranches sagittales des images C(V) statiques 3D, 2) d'appliquer ces transformations à la première trame de la séquence CV dynamique. Il en résulte une estimation de la forme 3D du tractus vocal pour la première trame de l'IRMtr C(V). 3) Un ensemble de transformations sont calculées entre la première trame (mi-sagittale) de la séquence CV dynamique et les autres trames de cette séquence. 4) Ces transformations sont appliquées à la forme estimée (à l'étape 2) du tractus vocal pour créer une version dynamique 3D basée sur la consonne. La procédure correspondante est appliquée à la voyelle afin de créer la deuxième version de la forme dynamique 3D. Enfin, les deux versions sont correctement fusionnées pour créer la forme dynamique 3D finale.

L'un des défis que nous avons dû relever était le fait que l'articulation statique est différente de la parole naturelle. Cela était particulièrement évident pour les consonnes, puisque la position des articulateurs dépend fortement de la voyelle anticipée. Ce problème a persisté malgré le protocole d'acquisition de l'IRM 3D qui tient compte de la coarticulation. En particulier dans les plosives /t/ et /p/, le simple fait de maintenir

une articulation stable est un défi pour l'orateur. Cependant, dans les Dans le cas des voyelles, la situation est meilleure puisque les voyelles peuvent être prononcées seules.

Une remarque importante est que nous avons utilisé des données provenant de différentes machines, acquises avec différents qui se traduisent par une qualité d'image très différente. De plus, nous avons traité deux types de (statique avec dynamique) et a même utilisé différents haut-parleurs avec des différences anatomiques, les images résultantes étaient suffisamment robustes pour le plan mi-sagittal. Leur comportement dans la transformation dynamique 3D pour le même locuteur et pour plusieurs locuteurs était cohérent.

Enfin, nous pouvons voir que la qualité des images synthétisées s'améliore pour les tranches mi-sagittales car elles ont un contraste et une résolution accrus tout en préservant la forme des voies vocales des images en 2D et dans le dans la majorité des cas, les informations anatomiques également. Pour les cadres non mi-sagittales, on peut encore synthétiser des images raisonnables, mais avec plus de bruit/artefacts, et dans certains cas des problèmes aux points de collage.

7.5 Modèle générique de locuteur

Dans ce chapitre, nous proposons une méthode que l'on peut utiliser pour créer un locuteur artificiel qui peut être utilisé comme locuteur standard. Cela permettra à quelqu'un non seulement de trouver des solutions potentiellement meilleures à des problèmes partiellement résolus/explorés mais aussi pour étudier les questions sous un angle nouveau.

Dans le présent travail, nous proposons une méthode pour construire des atlas dynamiques en 3D de la les voies vocales en utilisant l'IRMtr de plans sagittals parallèles à une fréquence d'images élevée, sans exiger de formation préalable. La principale question abordée est de savoir s'il est possible réduire l'inter et l'intra-variabilité des locuteurs en utilisant l'espace de l'atlas comme locuteur générique standard. Une des contributions de notre travail est d'utiliser l'approche de création d'atlas histologique pour collecter les informations 3D, en utilisant l'IRMtr pour acquérir les données, ce qui offre une fréquence d'images élevée et réduit la quantité de répétitions requises par d'autres techniques comme le cinémomètre. Une telle approche est nouvelle pour les atlas des voies vocales.

Une autre contribution est l'utilisation de la technique du kernel gaussien adaptatif pour créer les échantillons de l'atlas avec l'avantage de rendre la fréquence de trame de l'atlas indépendante de la fréquence de trame de l'IRMtr. La méthode proposée donne donc plus de souplesse pour contrôler l'atlas résultant paramètres. Les mêmes données peuvent donc être utilisées pour créer divers atlas avec des paramètres différents sans qu'il soit nécessaire de procéder à de nouvelles acquisitions de données à chaque fois.

Enfin, et c'est un avantage déterminant dans l'étude de la production de la parole, l'atlas construit avec cette méthode peut être utilisé comme locuteur de référence pour réduire la variabilité entre et au sein des sujets.

En effet, de nombreux travaux consacrés à la production de la parole d'un point de vue général sont basés sur l'hypothèse implicite d'un modèle articulatoire construit à partir d'un seul locuteur, ce qui est le cas du célèbre modèle articulatoire de Maeda, est valable pour tous les locuteurs. Il s'agit d'une simplification qui réduit la portée et la validité

d'une grande partie du travail. Dans notre approche, au contraire, nous avons introduit la variabilité dans la construction de l'atlas lui-même, qui couvre donc effectivement une grande variabilité des locuteurs, à condition que les locuteurs utilisés soient suffisamment divers.

Dans ce travail, un atlas dynamique des voies vocales est généré à partir de l'IRMtr en utilisant le nouvel algorithme proposé et une validation croisée quadruple avec la correspondance des histogrammes permettent d'évaluer s'il est possible d'utiliser l'espace de l'atlas comme modèle générique de locuteurs afin de réduire la variabilité entre les locuteurs. Les résultats montrent que l'utilisation de l'atlas proposé permet de saisir le comportement dynamique des articulateurs et est capable de généraliser le processus de production de la parole en créant un espace de référence universel des locuteurs.

7.6 Discussion

Cette thèse présente dans un premier temps quelques unes des méthodes courantes d'acquisition de données articulatoires. Parmi elles, l'IRM semble la plus prometteuse ; c'est pourquoi les principales bases de données articulatoires de l'IRM sont présentés et leurs utilisations potentielles sont discutées. La nouvelle base de données articulatoire pertinente pour la langue française est alors largement décrite puisque c'est celle qui est utilisée pour un une grande partie des études de cette thèse. Divers algorithmes ont été développés et plusieurs expériences ont été menées, qui peuvent être principalement divisé en trois parties. La première partie concerne l'étude de l'effet que le vélum, l'épiglotte et la position de la tête ont sur la parole production. Leur effet a été étudié par des études manuelles la segmentation et le traitement des données statiques de l'IRM et l'utilisation de simulations acoustiques. Les données simulées et originales ont été examinées pour évaluer l'effet des paramètres mentionnés précédemment en fonction des différences que nous avons remarqué dans les formants. La deuxième partie est consacrée au développement d'algorithmes qui permettra de surmonter certaines des contraintes techniques qui pèsent encore sur l'acquisition actuelle de l'IRM technologie. Les algorithmes proposés dans cette partie comprennent un moyen d'afficher le processus mi-sagittal vidéos IRMtr des voies vocales afin d'améliorer la qualité de l'image, quelques approches pour générer les formes 3D dynamiques correspondantes des voies vocales ainsi qu'un moyen de synthétiser des données pseudo IRMtr provenant de trames silencieuses pendant la production du CV. Tous ces algorithmes étaient basé sur des transformations d'images qui traitent chaque image dans sa totalité, donc pas de délimitation des contours ou d'autres informations (difficiles à obtenir) de l'image est nécessaire avant l'application des algorithmes.

Dans la dernière partie de cette thèse, un algorithme qui génère la 3D les données dynamiques d'un locuteur standard pendant la production du CV sont présentées. Pour le développement de ce un protocole spécial d'acquisition de données IRM a été développé, de nouvelles données IRM ont été acquises et une procédure de mesure des dimensions des voies vocales basée sur des repères anatomiques a été décrite. Ces données générées par l'algorithme proposé peuvent être utilisées comme un espace de référence standard (par exemple pour réduire la variabilité des sujets) puisqu'ils sont locuteurs indépendant. Le modèle générique de locuteur proposé pourrait également servir de base à plusieurs

autres des études telles que la modélisation articulatoire, la synthèse vocale articulatoire, etc. La variabilité est traitée entre les sujets.

En résumé, la contribution de la thèse est de donner un aperçu de l'effet des simplifications de la géométrie des voies vocales, les approches visant à surmonter les limites de la technologie actuelle d'imagerie par RM et un algorithme qui permet d'intégrer dans le modèle des informations sur la variabilité du locuteur processus de construction.

Bibliography

[ADB+16] Marc Arnela, Saeed Dabbaghchian, Rémi Blandin, Oriol Guasch, Olov Engwall, Annemie Van Hirtum, and Xavier Pelorson. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. *The Journal of the Acoustical Society of America*, 140(3):1707–1718, 2016.

[ADMC09] Mayssa Ahmad, Jacques Dargaud, André Morin, and François Cotton. Dynamic mri of larynx and vocal fold vibrations in normal phonation. *Journal of Voice*, 23(2):235–239, 2009.

[ANH97] Abeer Alwan, Shrikanth Narayanan, and Katherine Haker. Toward articulatory-acoustic models for liquid approximants based on mri and epg data. part ii. the rhotics. *The Journal of the Acoustical Society of America*, 101(2):1078–1089, 1997.

[AXG16] Nitin Agarwal, Xiangmin Xu, and M Gopi. Robust registration of mouse brain slices with severe histological artifacts. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, page 10. ACM, 2016.

[BBB01] D. Beautemps, P. Badin, and G. Bailly. Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America*, 109(5):2165–2180, 2001.

[BBR+02] Pierre Badin, Gerard Bailly, Lionel Reveret, Monica Baciu, Christoph Segebarth, and Christophe Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3):533–553, 2002.

[BBRS98] Pierre Badin, Gérard Bailly, Monica Raybaudi, and Christoph Segebarth. A three-dimensional linear articulatory model based on mri data. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[BGGN91] Thomas Baer, John C Gore, L Carol Gracco, and Patrick W Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *The Journal of the Acoustical Society of America*, 90(2):799–828, 1991.

[BHG+14] Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert. Robust articulatory speech synthesis using deep neural networks for bci applications. 2014.

[Bir07] Peter Birkholz. Articulatory synthesis of singing. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[Bir13] P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS one*, 8(4), 2013.

[BJK06] Peter Birkholz, Dietmar Jackèl, and Bernd Kröger. Construction and control of a three-dimensional vocal tract model. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[BKGN10] Erik Bresch, Athanasios Katsamanis, Louis Goldstein, and Shrikanth S Narayanan. Statistical multi-stream modeling of real-time mri articulatory speech data. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[BNNN06] Erik Bresch, Jon Nielsen, Krishna Nayak, and Shrikanth Narayanan. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, 120(4):1791–1794, 2006.

[BTL19] Pierre Badin, Marija Tabain, and Laurent Lamalle. Comparative study of coarticulation in a multilingual speaker: Preliminary results from mri data. 2019.

[BV17] Peter Birkholz and Elisabeth Venus. Considering lip geometry in one-dimensional tube models of the vocal tract. In *International Seminar on Speech Production*, pages 78–86. Springer, 2017.

[BW01] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[CBWJ13] Evan Calabrese, Alexandra Badea, Charles Watson, and G Allan Johnson. A quantitative magnetic resonance histology atlas of postnatal rat brain development with regional estimates of growth and variability. *Neuroimage*, 71:196–206, 2013.

[CCC+08] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. Meshlab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conference*, volume 2008, pages 129–136, 2008.

[CCW13] Nicolas Chauffert, Philippe Ciuciu, and Pierre Weiss. Variable density compressed sensing in mri. theoretical vs heuristic sampling strategies. In

2013 IEEE 10th International Symposium on Biomedical Imaging, pages 298–301. IEEE, 2013.

[CHK+19] C Carignan, P Hoole, E Kunay, A Joseph, D Voit, J Frahm, and J Harrington. The phonetic basis of phonological vowel nasality: Evidence from real-time mri velum movement in german. In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 413–417. Australasian Speech Science and Technology Association Inc., 2019.

[CKS97] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997.

[CMY+11] Nelson Chuang, Susumu Mori, Akira Yamamoto, Hangyi Jiang, Xin Ye, Xin Xu, Linda J Richards, Jeremy Nathans, Michael I Miller, Arthur W Toga, et al. An mri-based atlas and database of the developing mouse brain. *Neuroimage*, 54(1):80–89, 2011.

[CS86] Ff Charpentier and M Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2015–2018. IEEE, 1986.

[CSF+15] Christopher Carignan, Ryan K Shosted, Maojing Fu, Zhi-Pei Liang, and Bradley 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*, 50:34–51, 2015.

[DB10] Bart De Boer. Investigating the acoustic effect of the descended larynx with articulatory models. *Journal of Phonetics*, 38(4):679–686, 2010.

[DFBJ10] Brad C Davis, P Thomas Fletcher, Elizabeth Bullitt, and Sarang Joshi. Population shape regression from random design data. *International journal of computer vision*, 90(2):255–266, 2010.

[DHMS02] Didier Demolin, Sergio Hassid, Thierry Metens, and Alain Soquet. Real-time mri and articulatory coordination in speech. *Comptes rendus biologies*, 325(4):547–556, 2002.

[DKM18] Ioannis K Douros, Athanasios Katsamanis, and Petros Maragos. Multi-view audio-articulatory features for phonetic recognition on rtmri-timit database. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5514–5518. IEEE, 2018.

[DLVE19] Ioannis Douros, Yves Laprie, Pierre-André Vuissoz, and Benjamin Elie. Acoustic evaluation of simplifying hypotheses used in articulatory synthesis. 2019.

[DTI+19] Ioannis Douros, Anastasiia Tsukanova, Karyna Isaieva, Pierre-André Vuissoz, and Yves Laprie. Towards a method of dynamic vocal tract shapes generation by combining static 3d and dynamic 2d mri speech data. 2019.

[EAR08] Anders Ericsson, Paul Aljabar, and Daniel Rueckert. Construction of a patient-specific atlas of the brain: Application to normal aging. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 480–483. IEEE, 2008.

[EL] Subin Erattakulangara and Sajan Goud Lingala. Airway segmentation in speech mri using the u-net architecture.

[EL16] Benjamin Elie and Yves 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*, 82:85–96, 2016.

[ENRS20] Mohammad Eslami, Christiane Neuschaefer-Rube, and Antoine Serrurier. Automatic vocal tract landmark localization from midsagittal mri data. *Scientific Reports*, 10(1):1–13, 2020.

[Eri07] C. Ericsdotter. Detail in vowel area functions. In *Proc of the 16th ICPhS*, pages 513–516, Saarbrücken, Germany, 2007.

[FBH+17] Maojing Fu, Marissa S Barlaz, Joseph L Holtrop, Jamie L Perry, David P Kuehn, Ryan K Shosted, Zhi-Pei Liang, and Bradley P Sutton. High-frame-rate full-vocal-tract 3d dynamic speech imaging. *Magnetic resonance in medicine*, 77(4):1619–1629, 2017.

[FBKC+12] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[FEG+08] Michael J Fagan, Stephen R Ell, James M Gilbert, E Sarrazin, and Peter M Chapman. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics*, 30(4):419–425, 2008.

[FK01] Joe Frankel and Simon King. Asr-articulatory speech recognition. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[FKJ06] Zsuzsanna Fagyal, Douglas Kibbee, and Frederic Jenkins. *French: A linguistic introduction*. Cambridge University Press, 2006.

[FMJ15] Dominique Fohr, Odile Mella, and Denis Jouvet. De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée. In *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, 2015.

[FQXS14] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[FVVDD+06] Sidney Fels, Florian Vogt, Kees Van Den Doel, John Lloyd, Ian Stavness, and Eric Vatikiotis-Bateson. Artisynth: A biomechanical simulation platform for the vocal tract and upper airway. In *International Seminar on Speech Production, Ubatuba, Brazil*, volume 138, 2006.

[FWLS16] Maojing Fu, Jonghye Woo, Zhi-Pei Liang, and Bradley P Sutton. Spatiotemporal-atlas-based dynamic speech imaging. In *Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 9788, page 978804. International Society for Optics and Photonics, 2016.

[GFW+06] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, and J. Berg. Detecting fatigue from voice using speech recognition. In *2006 IEEE International Symposium on Signal Processing and Information Technology*, pages 567–571, 2006.

[GGM92] AR Greenwood, CC Goodyear, and PA Martin. Measurements of vocal tract shapes using magnetic resonance imaging. *IEE Proceedings I (Communications, Speech and Vision)*, 139(6):553–560, 1992.

[GN10] Prasanta Kumar Ghosh and Shrikanth Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–2172, 2010.

[GRH+08] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008.

[GS] Pravin M Ghate and SD Shirbhadurkar. Speech synthesis using syllable for marathi language. *International Journal of Engineering and Research Technology*.

[Har72] William J Hardcastle. The use of electropalatography in phonetic research. *Phonetica*, 25(4):197–215, 1972.

[Har14] Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014.

[HB96] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE, 1996.

[Hen99] J Hennig. K-space sampling strategies. *European radiology*, 9(6):1020–1031, 1999.

[Hix71] Thomas J Hixon. An electromagnetic method for transducing jaw movements during speech. *The Journal of the Acoustical Society of America*, 49(2B):603–606, 1971.

[HJPA+03] Mark Hasegawa-Johnson, Shamala Pizza, Abeer Alwan, Jul Setsu Alwan, and Katherine Haker. Vowel category dependence of the relationship between palate height, tongue height, and oral area. *Journal of Speech, Language, and Hearing Research*, 2003.

[HKT+12] Hiroaki Hatano, Tatsuya Kitamura, Hironori Takemoto, Parham Mokhtari, Kiyoshi Honda, and Shinobu Masaki. Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five japanese vowels uttered by fifteen male speakers. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[HM11] Ian S Howard and Piers Messum. Modeling the development of pronunciation in infant speech acquisition. 2011.

[HPP17] Nicolas Hermant, Pascal Perrier, and Yohan Payan. Human tongue biomechanical modeling. In *Biomechanics of Living Organs*, pages 395–411. Elsevier, 2017.

[HS65] J. M. Heinz and K. N. Stevens. On the relations between lateral cineradiographs, area functions and acoustic spectra of speech. In *Proceedings of the 5th International Congress on Acoustics*, page A44., 1965.

[HVG+96] William Hardcastle, Béatrice Vaxelaire, Fiona Gibbon, Philip Hoole, and N Nguyen. Ema/epg study of lingual coarticulation in/kl/clusters. In *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data*, 1996.

[IAse] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Trans. IECE*, J62-A(4):217–223, 1979 (in japanese).

[JLL93] Keith Johnson, Peter Ladefoged, and Mona Lindau. Individual differences in vowel production. *The Journal of the Acoustical Society of America*, 94(2):701–714, 1993.

[JMD94] Mohammed A Jan, Ian Marshall, and Neil J Douglas. Effect of posture on upper airway dimensions in normal human. *American Journal of Respiratory and Critical Care Medicine*, 149(1):145–148, 1994.

[Kaw06] Hideki Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[KBG+11] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. Sailalign: Robust long speech-text alignment. In *Proc. of workshop on new tools and methods for very-large scale phonetics research*, 2011.

[KBRN11] Athanasios Katsamanis, Erik Bresch, Vikram Ramanarayanan, and Shrikanth Narayanan. Validating rt-mri based articulatory representations via articulatory recognition. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[KF+82] David P Kuehn, John W Folkins, et al. Relationships between muscle activity and velar position. *The Cleft palate journal*, 19(1):25–35, 1982.

[KFL+07] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121(2):723–742, 2007.

[KGBL08] Bernd J Kröger, Verena Graf-Borttscheller, and Anja Lowit. Two and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In *Interspeech, 9th Annual Conference of the International Speech Communication Association*, 2008.

[KKLN14] Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth Narayanan. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *Proceedings of the International Seminar on Speech Production ISSP*, pages 222–225. Citeseer, 2014.

[KKT+14] Hideki Kawahara, Tatsuya Kitamura, Hironori Takemoto, Ryuichi Nisimura, and Toshio Irino. Vocal tract length estimation based on vowels using a database consisting of 385 speakers and a database with mri-based vocal tract shape information. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[KKTH19] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824. IEEE, 2019.

[KLN11] Jangwon Kim, Sungbok Lee, and Shrikanth Narayanan. An exploratory study of the relations between perceived emotion strength and articulatory kinematics. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[KMAS+11] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011.

[kov] URL: `https://www.peterkovesi.com/matlabfns/index.html#edgelink`.

[KS02] Edith Kaan and Tamara Y Swaab. The brain circuitry of syntactic comprehension. *Trends in cognitive sciences*, 6(8):350–356, 2002.

[KTAH09] Tatsuya Kitamura, Hironori Takemoto, Seiji Adachi, and Kiyoshi Honda. Transfer functions of solid vocal-tract models constructed from atr mri database of japanese vowel production. *Acoustical science and technology*, 30(4):288–296, 2009.

[KTK+14] Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan. Usc-emo-mri corpus: An emotional speech production database recorded by real-time magnetic resonance imaging. In *International Seminar on Speech Production (ISSP), Cologne, Germany*, page 226, 2014.

[KWMPM00] Bernd J Kröger, Ralf Winkler, Christine Mooshammer, and Bernd Pompino-Marschall. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. In *Proceedings of the 5th Seminar on Speech Production*, pages 333–336, 2000.

[Lap99] Y. Laprie. Snorri, a software for speech sciences. In *Proceedings of Matisse 99 (Methods and tools inovations for speech science education)*, pages 89–92, London, April, 1999.

[LB11] Y. Laprie and J. Busset. Construction and evaluation of an articulatory model of the vocal tract. In *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Spain, August 2011.

[LBV+18] Mathieu Labrunie, Pierre Badin, Dirk Voit, Arun A Joseph, Jens Frahm, Laurent Lamalle, Coriandre Vilain, and Louis-Jean Boë. Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning. *Speech Communication*, 99:27–46, 2018.

[LET15] Yves Laprie, Benjamin Elie, and Anastasiia Tsukanova. 2d articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes. 2015.

[LETV18] Yves Laprie, Benjamin Elie, Anastasiia Tsukanova, and Pierre-André Vuissoz. Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2110–2114. IEEE, 2018.

[Lev92] Willem JM Levelt. Accessing words in speech production: Stages, processes and representations. 1992.

[lex] URL: `http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php`.

[LJW+12] Shu Liao, Hongjun Jia, Guorong Wu, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. A novel framework for longitudinal atlas construction with groupwise registration of subject image sequences. *NeuroImage*, 59(2):1275–1289, 2012.

[LLM+13a] Y. Laprie, M. Loosvelt, S. Maeda, E. Sock, and F. Hirsch. Articulatory copy synthesis from cine x-ray films. In *Interspeech 2013 (14th Annual Conference of the International Speech Communication Association)*, Lyon, France, August 2013.

[LLM+13b] Yves Laprie, Matthieu Loosvelt, Shinji Maeda, Rudolph Sock, and Fabrice Hirsch. Articulatory copy synthesis from cine x-ray films. 2013.

[Lon84] F. Lonchamp. Les sons du Français — Analyse acoustique descriptive. Cours de phonétique, Institut de Phonétique, Université de Nancy II, 1984.

[LS71] Björn EF Lindblom and Johan EF Sundberg. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4B):1166–1179, 1971.

[LSM+17] Falk Lüsebrink, Alessandro Sciarra, Hendrik Mattern, Renat Yakupov, and Oliver Speck. T 1-weighted in vivo human whole brain mri dataset with an ultrahigh isotropic resolution of 250 μm. *Scientific data*, 4(1):1–12, 2017.

[LSMN16] Sajan Goud Lingala, Brad P Sutton, Marc E Miquel, and Krishna S Nayak. Recommendations for real-time speech mri. *Journal of Magnetic Resonance Imaging*, 43(1):28–44, 2016.

[LSSS+15] Eleanor Lawson, Jane Stuart-Smith, James M Scobbie, Satsuki Nakai, David Beavan, Fiona Edmonds, Iain Edmonds, Alice Turk, Claire Timmins, Janet M Beck, et al. Seeing speech: an articulatory web resource for the study of phonetics [website]. 2015.

[LSVE14] Yves Laprie, Rudolph Sock, Béatrice Vaxelaire, and Benjamin Elie. Comment faire parler les images aux rayons x du conduit vocal. In *SHS Web of Conferences*, volume 8, pages 1285–1298. EDP Sciences, 2014.

[LZL+19] Yongwan Lim, Yinghua Zhu, Sajan Goud Lingala, Dani Byrd, Shrikanth Narayanan, and Krishna Shrinivas Nayak. 3d dynamic mri of the vocal tract during natural speech. *Magnetic resonance in medicine*, 81(3):1511–1520, 2019.

[Mad84] Ian Maddieson. *Patterns of sounds (Cambridge Studies in Speech Science and Communication)*. Cambridge university press, 1984. doi:10.1017/CBO9780511753459.

[Mae82] Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech communication*, 1(3-4):199–229, 1982.

[Mae90] Shinji Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, pages 131–149. Springer, 1990.

[Mae91] Shinji Maeda. On articulatory and acoustic variabilities. *Journal of Phonetics*, 19(3-4):321–331, 1991.

[mat] https://ch.mathworks.com/matlabcentral/fileexchange/20057-b-spline-grid-image-and-point-based-registration.

[MC90] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.

[Mer73] Paul Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[ML13] Shinji Maeda and Yves Laprie. Vowel and prosodic factor dependent variations of vocal-tract length. In *InterSpeech-14th Annual Conference of the International Speech Communication Association-2013*, 2013.

[MOST12] Paula Martins, Catarina Oliveira, Samuel Silva, and António Teixeira. Velar movement in european portuguese nasal vowels. In *Proc IberSpeech 2012 VII Jornadas en Tecnolog0ìya del Habla and IIIIberian SLTech Workshop*, pages 231–240, 2012.

[NAH97] Shrikanth S Narayanan, Abeer A Alwan, and Katherine Haker. Toward articulatory-acoustic models for liquid approximants based on mri and epg data. part i. the laterals. *The Journal of the Acoustical Society of America*, 101(2):1064–1077, 1997.

[NBG+11] Shrikanth Narayanan, Erik Bresch, Prasanta Kumar Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon Kim, Adam Lammert, Michael Proctor, Vikram Ramanarayanan, and Yinghua Zhu. A multimodal real-time mri articulatory corpus for speech research. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[NTR+14] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim,

Yinghua Zhu, Louis 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*, 136(3):1307–1311, 2014.

[NZK+13] Aaron Niebergall, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm. Real-time mri of speaking at a resolution of 33 ms: Undersampled radial flash with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, 69(2):477–485, 2013.

[ODZ+16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[OL05] Slim Ouni and Yves Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 118(1):444–460, 2005.

[OMST12] Catarina Oliveira, Paula Martins, Samuel S Silva, and António JS Teixeira. An mri study of the oral articulation of european portuguese nasal vowels. In *INTERSPEECH*, pages 2690–2693, 2012.

[OMT09] Catarina Oliveira, Paula Martins, and António Teixeira. Speech rate effects on european portuguese nasal vowels. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[OVB12] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on audio, speech, and language processing*, 20(4):1118–1133, 2012.

[PCS+92] Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

[PKSF17] Jamie L Perry, David P Kuehn, Bradley P Sutton, and Xiangming Fang. Velopharyngeal structural and functional assessment of speech in young children using dynamic magnetic resonance imaging. *The Cleft Palate-Craniofacial Journal*, 54(4):408–422, 2017.

[PLK+11] Michael Proctor, Adam Lammert, Athanasios Katsamanis, Louis Goldstein, Christina Hagedorn, and Shrikanth Narayanan. Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[PP97] Yohan Payan and Pascal Perrier. Synthesis of vv sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech communication*, 22(2-3):185–205, 1997.

[PPS+07] Vijay Parthasarathy, Jerry L Prince, Maureen Stone, Emi Z Murano, and Moriel NessAiver. Measuring tongue motion from tagged cine-mri using harmonic phase (harp) processing. *The Journal of the Acoustical Society of America*, 121(1):491–504, 2007.

[PPZP03] Pascal Perrier, Yohan Payan, Majid Zandipour, and Joseph Perkell. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *JASA*, 114(3):1582–1599, 2003.

[PZL+14] Michael Proctor, Yinghua Zhu, Adam Lammert, Asterios Toutios, Bonny Sands, Ulrich Hummel, and Shrikanth Narayanan. Click consonant production in khoekhoe: A real-time mri study. In *Khoisan Languages and Linguistics. Proc. 5th Intl. Symposium*, pages 337–366, 2014.

[Rab78] Lawrence R Rabiner. Digital processing of speech signal. *Digital Processing of Speech Signal*, 1978.

[RBBD14] Sophie Roekhaut, Sandrine Brognaux, Richard Beaufort, and Thierry 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.

[RFBM19] Matthieu Ruthven, Andreia C Freitas, Redha Boubertakh, and Marc E Miquel. Application of radial grappa techniques to single-and multislice dynamic speech mri using a 16-channel neurovascular coil. *Magnetic resonance in medicine*, 82(3):948–958, 2019.

[RLN+17] P-Y Rohan, Claudio Lobos, Mohammad Ali Nazari, Pascal Perrier, and Yohan Payan. Finite element models of the human tongue: a mixed-element mesh approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 5(6):390–400, 2017.

[RMS09] Friederike Roers, Dirk Mürbe, and Johan Sundberg. Voice classification and vocal tract of singers: A study of x-ray images and morphology. *The Journal of the Acoustical Society of America*, 125(1):503–512, 2009.

[RR05] A. Röbel and X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proc. of the 8 th Int. Conference on Digital Audio Effects (DAFx'05)*, Madrid, 2005.

[RSAF04] Laurent Romary, Susanne Salmon-Alt, and Gil Francopoulo. Standards going concrete: from lmf to morphalou. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 22–28. Association for Computational Linguistics, 2004.

[RSH+99] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[RSS96] Richard C Rose, Juergen Schroeter, and MM Sondhi. The potential role of speech production models in automatic speech recognition. *The Journal of the Acoustical Society of America*, 99(3):1699–1709, 1996.

[RTP+18] Vikram Ramanarayanan, Sam Tilsen, Michael Proctor, Johannes Töger, Louis Goldstein, Krishna S Nayak, and Shrikanth Narayanan. Analysis of speech production real-time mri. *Computer Speech & Language*, 2018.

[RWCL10] C Rumack, S Wilson, JW Charboneau, and D Levine. Diagnostic ultrasound, 2-volume set. *Missouri: Elsevier Mosby*, 2010.

[SAB+12] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012.

[SB05] Antoine Serrurier and Pierre Badin. A three-dimensional linear articulatory model of velum based on mri data. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[SB08] Antoine Serrurier and Pierre Badin. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on mri and ct data. *The Journal of the Acoustical Society of America*, 123(4):2335–2355, 2008.

[SBN+16] Yi Sun, Ole Brauckmann, Donald R Nixdorf, Arno Kentgens, Michael Garwood, Djaudat Idiyatullin, and Arend Heerschap. Imaging human teeth by phosphorus magnetic resonance with nuclear overhauser enhancement. *Scientific reports*, 6:30756, 2016.

[SD95] Maureen Stone and Edward P Davis. A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of The Acoustical Society of America*, 98(6):3107–3112, 1995.

[SD01] Christine H Shadle and Robert I Damper. Prospects for articulatory synthesis: A position paper. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[SDD+01] Maureen Stone, Edward P Davis, Andrew S Douglas, Moriel NessAiver, Rao Gullapalli, William S Levine, and Andrew Lundberg. Modeling the motion of the internal tongue from tagged cine-mri images. *The Journal of the Acoustical Society of America*, 109(6):2974–2982, 2001.

[SDM+04] Dieter Seghers, Emiliano D'Agostino, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Construction of a brain template from mr images using state-of-the-art registration and segmentation techniques. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 696–703. Springer, 2004.

[SHL+11] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, and J. Sturm. DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models. In *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.

[SHT96] Brad H Story, Eric A Hoffman, and Ingo R Titze. Vocal tract imaging: a comparison of mri and ebct. In *Medical Imaging 1996: Physiology and Function from Multidimensional Images*, volume 2709, pages 209–222. International Society for Optics and Photonics, 1996.

[SMK+17] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

[SMWC03] Stephanie M. Strassel, David Miller, Kevin Walker, and Christopher Cieri. Shared resources for robust speech-to-text technology. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, 2003. URL: `http://www.isca-speech.org/archive/eurospeech_2003/e03_1609.html`.

[SNA13] Danny D Steinberg, Hiroshi Nagata, and David P Aline. *Psycholinguistics: Language, mind and world*. Routledge, 2013.

[SP16] Andrew Szabados and Pascal Perrier. Uncontrolled manifolds in vowel production: Assessment with a biomechanical model of the tongue. 2016.

[SRE12] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671, 2012.

[SSF18] Pramit Saha, Praneeth Srungarapu, and Sidney Fels. Towards automatic speech identification from vocal tract shape dynamics in real-time mri. *arXiv preprint arXiv:1807.11089*, 2018.

[SST+17] Tanner Sorensen, Zisis Iason Skordilis, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Jangwon Kim, Adam C Lammert, Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, et al. Database of volumetric and real-time vocal tract mri for speech science. In *INTERSPEECH*, pages 645–649, 2017.

[ST15] Samuel Silva and António Teixeira. Unsupervised segmentation of the vocal tract from real-time mri sequences. *Computer Speech & Language*, 33(1):25–46, 2015.

[ST17] Samuel Silva and António JS Teixeira. Critical articulators identification from rt-mri of the vocal tract. In *INTERSPEECH*, pages 626–630, 2017.

[STI04] Hiroyuki Segi, Tohru Takagi, and Takayuki Ito. A concatenative speech synthesis method using context dependent phoneme sequences with variable length as search units. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[STTN17] Zisis Iason Skordilis, Asterios Toutios, Johannes Töger, and Shrikanth Narayanan. Estimation of vocal tract area function from volumetric magnetic resonance imaging. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 924–928. IEEE, 2017.

[Sty01] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on speech and audio processing*, 9(1):21–29, 2001.

[SVP17] Protima Nomo Sudro, CM Vikram, and SR Mahadeva Prasanna. Vowel onset point based characterization of velopharyngeal activity using imaging techniques. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–5. IEEE, 2017.

[Tak01] Hironori Takemoto. Morphological analyses of the human tongue musculature for three-dimensional modeling. *Journal of Speech, Language, and Hearing Research*, 2001.

[TC10a] Bradley E Treeby and Benjamin T Cox. k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields. *Journal of biomedical optics*, 15(2):021314, 2010.

[TC10b] Bradley E Treeby and BT Cox. Modeling power law absorption and dispersion for acoustic propagation using the fractional laplacian. *The Journal of the Acoustical Society of America*, 127(5):2741–2748, 2010.

[TCJ12] Bradley Treeby, Ben Cox, and Jiri Jaros. k-wave a matlab toolbox for the time domain simulation of acoustic wave fields user manual. *Manual Version 1. 0. 1*, 2012.

[TDSL19] Anastasiia Tsukanova, Ioannis K. Douros, Anastasia Shimorina, and Yves Laprie. Can static vocal tract positions represent articulatory targets in continuous speech? Matching static MRI captures against real-time MRI for the French language. In *International Congress on Phonetic Sciences, 5-9 August, Melbourne, Australia*, 2019. Accepted for publication.

[TEL17] Anastasiia Tsukanova, Benjamin Elie, and Yves Laprie. Articulatory speech synthesis from static context-aware articulatory targets. In *International Seminar on Speech Production*, pages 37–47. Springer, 2017.

[TGH+19] Hironori Takemoto, Tsubasa Goto, Yuya Hagihara, Sayaka Hamanaka, Tatsuya Kitamura, Yukiko Nota, and Kikuo Maekawa. Speech organ contour extraction using real-time mri and machine learning method. *Proc. Interspeech 2019*, pages 904–908, 2019.

[Thi98] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[THM+06] Hironori Takemoto, Kiyoshi Honda, Shinobu Masaki, Yasuhiro Shimada, and Ichiro Fujimoto. Measurement of temporal changes in vocal tract area function from 3d cine-mri data. *The Journal of the Acoustical Society of America*, 119(2):1037–1049, 2006.

[TMK10] Hironori Takemoto, Parham Mokhtari, and Tatsuya Kitamura. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. *JASA*, 128(6):3724–3738, 2010.

[TMO+12] António Teixeira, Paula Martins, Catarina Oliveira, Carlos Ferreira, Augusto Silva, and Ryan Shosted. Real-time mri for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 306–317. Springer, 2012.

[TN16] Asterios Toutios and Shrikanth 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*, 5, 2016.

[TV06] G Touré and C Vacher. Anatomic study of tongue architecture based on fetal histological sections. *Surgical and Radiologic Anatomy*, 28(6):547–552, 2006.

[UZV+10] Martin Uecker, Shuo Zhang, Dirk Voit, Alexander Karaus, Klaus-Dietmar Merboldt, and Jens Frahm. Real-time mri at a resolution of 20 ms. *NMR in Biomedicine*, 23(8):986–994, 2010.

[Vax93] Béatrice Vaxelaire. *Etude comparee des effets des variations de debit-lent, rapide-surles parametres articulatoires, a partir de la cineradiographie (sujets francais)*. PhD thesis, Strasbourg 2, 1993.

[vow] http://www.isle.illinois.edu/sst/data/mri/.

[VPPA09] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[Whi98] Ross T Whitaker. A level-set approach to 3d reconstruction from range data. *International journal of computer vision*, 29(3):203–231, 1998.

[WIT+05] Douglas H Whalen, Khalil Iskarous, Mark K Tiede, David J Ostry, Heike Lehnert-LeHouillier, Eric Vatikiotis-Bateson, and Donald S Hailey. The haskins optically corrected ultrasound system (hocus). *Journal of Speech, Language, and Hearing Research*, 2005.

[WL05] Samuel R Ward and Richard L Lieber. Density and hydration of fresh and fixed human skeletal muscle. *Journal of biomechanics*, 38(11):2317–2320, 2005.

[WLM+15] Jonghye Woo, Junghoon Lee, Emi Z Murano, Fangxu Xing, Meena Al-Talib, Maureen Stone, and Jerry L Prince. A high-resolution atlas and statistical model of the vocal tract from structural mri. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 3(1):47–60, 2015.

[WMWK90] John Westbury, Paul Milenkovic, Gary Weismer, and Raymond Kent. X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1):S56–S56, 1990.

[Wre00a] Alan A Wrench. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. *Phonus.*, 2000.

[Wre00b] Alan A Wrench. A multichannel articulatory database and its application for automatic speech recognition. In *In Proceedings 5 th Seminar of Speech Production*. Citeseer, 2000.

[WSRS+17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[WT13] Elliott S Wise and Bradley E Treeby. Full-wave nonlinear ultrasound simulation in an axisymmetric coordinate system using the discrete sine and cosine transforms. In *Ultrasonics Symposium (IUS), 2013 IEEE International*, pages 1374–1377. IEEE, 2013.

[WTK+20] Yi-Chiao Wu, Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki Hayashi, and Tomoki Toda. Non-parallel voice conversion system with wavenet vocoder and collapsed speech suppression. *arXiv preprint arXiv:2003.11750*, 2020.

[WWK16] Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*, 2016.

[WXL+15] Jonghye Woo, Fangxu Xing, Junghoon Lee, Maureen Stone, and Jerry L Prince. Construction of an unbiased spatio-temporal atlas of the tongue during speech. In *International Conference on Information Processing in Medical Imaging*, pages 723–732. Springer, 2015.

[WXL+18] Jonghye Woo, Fangxu Xing, Junghoon Lee, Maureen Stone, and Jerry L Prince. 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):520–531, 2018.

[WXS+19] Jonghye Woo, Fangxu Xing, Maureen Stone, Jordan Green, Timothy G Reese, Thomas J Brady, Van J Wedeen, Jerry L Prince, and Georges El Fakhri. Speech map: A statistical multimodal atlas of 4d tongue motion during speech from tagged and cine mr images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(4):361–373, 2019.

[XPS+17] Fangxu Xing, Jerry L Prince, Maureen Stone, Van J Wedeen, Georges El Fakhri, and Jonghye Woo. A four-dimensional motion field atlas of the tongue from tagged and cine magnetic resonance imaging. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101331H. International Society for Optics and Photonics, 2017.

[XRLH18] Jing Xiong, Jing Ren, Liqun Luo, and Mark Horowitz. Mapping histological slice sequences to the allen mouse brain atlas without 3d reconstruction. *Frontiers in neuroinformatics*, 12:93, 2018.

[XSG+19] Fangxu Xing, Maureen Stone, Tessa Goldsmith, Jerry L Prince, Georges El Fakhri, and Jonghye Woo. Atlas-based tongue muscle correlation analysis from tagged and high-resolution magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 62(7):2258–2269, 2019.

[YEG+02] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The HTK book. *Cambridge university engineering department*, 3:175, 2002.

[YPH+06] Paul A Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C Gee, and Guido Gerig. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[ZKP+12] Yinghua Zhu, Yoon-Chul Kim, Michael I Proctor, Shrikanth S Narayanan, and Krishna S Nayak. Dynamic 3-d visualization of vocal tract shaping during speech. *IEEE transactions on medical imaging*, 32(5):838–848, 2012.

[ZY96] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE*

transactions on pattern analysis and machine intelligence, 18(9):884–900, 1996.

Résumé

Dans cette thèse, nous avons utilisé les données de l'IRM du conduit vocal pour étudier la production de la parole. La première partie consiste en l'étude de l'impact que le vélum, l'épiglotte et la position de la tête a sur la phonation de cinq voyelles françaises. Des simulations acoustiques ont été utilisées pour comparer les formants des cas étudiés avec la référence afin de mesurer leur impact. Pour cette partie du travail, nous avons utilisé des IRM statiques en 3D. Comme la parole est généralement une phénomène dynamique une question s'est posée, à savoir s'il serait possible de traiter les données 3D afin d'incorporer des informations temporelles de la parole continue. Par conséquent, la deuxième partie présente quelques algorithmes que l'on peut utiliser pour améliorer les données de production de la parole. Plusieurs transformations d'images ont été combinées afin de générer des estimations des formes du conduit vocal qui sont plus informatives que les originales. À ce stade, nous avons envisagé, outre l'amélioration des données de production de la parole, de créer un modèle de référence générique qui pourrait fournir des informations améliorées non pas pour un sujet spécifique, mais globalement pour la parole. C'est pourquoi nous avons consacré la troisième partie l'étude d'un algorithme permettant de créer un atlas spatio-temporel de l'appareil vocal qui peut être utilisé comme référence ou standard pour l'étude de la parole car il est indépendant du locuteur. Enfin, la dernière partie de la thèse, fait référence à une sélection de questions ouvertes du domaine qui restent encore sans réponse, quelques pistes intéressantes que l'on peut développer à partir de cette thèse et quelques approches potentielles qui pourraient être envisager afin de répondre à ces questions.

Mots-clés: IRM, production de la parole, conduit vocal simulation acoustique, transformation des images, amélioration des données articulatoires, atlas spatio-temporel

Abstract

In this thesis we used MRI (Magnetic Resonance Imaging) data of the vocal tract to study speech production. The first part consist of the study of the impact that the velum, the epiglottis and the head position has on the phonation of five french vowels. Acoustic simulations were used to compare the formants of the studied cases with the reference in order to measure their impact. For this part of the work, we used 3D static MR (Magnetic Resonance) images. As speech is usually a dynamic phenomenon, a question arose, whether it would be possible to process the 3D data in order to incorporate dynamic information of continuous speech. Therefore the second part presents some algorithms that one can use in order to enhance speech production data. Several image transformations were combined in order to generate estimations of vocal tract shapes which are more informative than the original ones. At this point, we envisaged apart from enhancing speech production data, to create a generic speaker model that could provide enhanced information not for a specific subject, but globally for speech. As a result, we devoted the third part in the investigation of an algorithm that one can use to create a spatio-temporal atlas of the vocal tract which can be used as a reference or standard speaker for speech studies as it is speaker independent. Finally, the last part of the thesis, refers to a selection of open questions of the field that are still left unanswered, some interesting directions that one can expand this thesis and some potential approaches that could help someone move forward towards these directions.

Keywords: MRI, speech production, vocal tract, acoustic simulation, image transformation, articulatory data enhancement, spatio-temporal atlas