Articulatory speech synthesis

Anastasiia Tsukanova

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Articulatory speech synthesis

THÈSE

présentée et soutenue publiquement le 13 12 2019

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Anastasiia Tsukanova

Composition du jury

Président :	Mme. Boyer Anne	
Rapporteurs :	M. Lolive Damien	d'ENSSAT (Université Rennes), IRISA
	M. Perrier Pascal	de Grenoble INP, Gipsa-lab
Examinateurs :	Mme. Adda-Decker Martine	de Laboratoire de Phonétique et Phonologie UMR 7018, CNRS Sorbonne-Nouvelle
	Mme. Boyer Anne	de l'Université de Lorraine

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

Mis en page avec la classe thesul.

Table of contents

List of Figures

Acknowledgments

A PhD thesis is a journey that you take on your own but see through to the end only thanks to a collective effort of those around you. All parts of that huge contribution are important. Sometimes it is someone who gives advice on how to build your model or looks at your results and says if they agree with your conclusion. Sometimes it is someone who listens to the accounts of an elusive bug and suggests an idea or works their magic on an administrative issue you are facing on top of everything else. It may also be someone who does not give up on you until they convince you that you can really do it. Or someone who stays up until very late just to talk to you. Or someone who drags you away from work for a coffee break and makes you smile. Or someone who stops by and does the dishes. I have never heard of anyone who would manage to pull off a PhD without the dishes.

So, my thanks goes to my supervisor Yves Laprie, who took me through a very big chunk of my career and personal transformation and, despite all hurdles, made sure I did not have a bad time. He is an admirable example of someone who has unique expertise in the field, loves what they do and delivers that to the world with a sense of justice, kindness and humor. Thank you.

I would also like to thank Ioannis Douros, without whom this thesis would have probably ended up being defended around 2031 and looking very different. A big part of what made this work possible is owed to him.

Thanks to Benjamin Elie, whom I worked with on acoustic simulations, to Asterios Toutios for helping out with their articulatory contour tracking tool, as well as to other collaborators of the ArtSpeech project, especially Pierre-André Vuissoz, Karyna Isaieva and Shinji Maeda. Helpful advice also came from the discussions with Peter Birkholz, Edwin Maas and Ian Howard. I am grateful for the very well developed suggestions for improvement made by the reviewers of my thesis, Pascal Perrier and Damien Lolive, and for the comments by Martine Adda-Decker and Anne Boyer. Additionally, I really appreciated the possibility to talk about all things that have to do with articulation with other members of Multispeech who work on it: Théo Biasutto-Lervat, Sara Dahmani, Aghilas Sini, Manfred Pastätter, Slim Ouni. Finally, my gratitude and respect goes to all our interns and volunteers, who showed a lot of endurance through manual annotation, urgent coding and the hardships of data acquisition.

It is inevitable that sometimes one gets stuck with an issue in software, hardware or, if you are really out of luck, both (repeatedly). This is where my big thanks goes to Ajinkya Kulkarni and Amal Houidhek for their help with Merlin, Sara Dahmani for eLite HTS, Nicolas Turpault and Manuel Pariente for Grid5000, Denis Jouvet for his general guidance, and probably many others, including the people from technical assistance services at Loria, who at some point had to visit my computer on an almost daily basis. In general, I need to say that Loria and in particular our Multispeech team are great places to work at, full of people who make for

amazing office mates (special shoutouts to Imran Sheikh for being this welcoming at my first office, Aditya Arie Nugraha for his ability to share a good laugh, Dayana Ribas for transforming every place she appeared at, Amélie Greiner for our coffee breaks, Guillaume Carbajal for his famous P line and Lou Lee for our shared self-indulgent rants) and lab mates and colleagues (especially Mathieu Hu, Sunit Sivasankaran, Nicolas Furnon, Mauricio Michel Olvera, Élodie Gauthier, Raphaël Duroselle, Diego Di Carlo, Badr Abdullah, Sucheta Gosh, Anna Kravchenko, Ameur Douib, Mohamed Amine Menacer, George Krait and Armelle Brun). Also, a heartfelt thanks to Isabelle, Tarek and Caroline from the canteen and Aroussiak from the reception desk at Loria — they really made my days brighter.

Then it is the turn of my personal fortress.

A lot of warm support came from Quentin Brabant (who would be there and help me out with everything, from French to bureaucracy to dealing with my teaching to assembling furniture), Anastasiia Demikhovska, Cristina Pérez Rivas, Alona Malukhina, Aitor Egurtzegi, Felicia Oberarzbacher, Rania Mohammed, Violeta Cervantes, Hennadii Leibenko, Iekaterina Blagina, Yevhen Los, Véronique Truk, Irene Stoukou, Kristina Shimanska, Celine Minier-Bidart, Angela Moreno, Caitlin McGrath (whom I especially thank for all the last-minute proofreading and editing she did, and by last minute I do mean *last minute*), Nikolai Ziuzin, Anna Ievleva, Anna Merkulova and Gilyan Mandzhieva. Also, Ksenia Timashkova, both on a professional and a personal level. I feel that each of you brought in something crucial, something no one else would have been able to, something that is integral to me now. Thank you, you are the best. I hope that one day I will be able to pay you back.

My PhD brought me two incredible people: Iordan Iordanov and Tatiana Makhalova. Both are people of immense kindness and ultimate inner integrity. I wouldn't have known what I would have missed, had I not joined this program. Iordan, thank you for putting yourself out there and really getting to know those around you. I think I have improved in that department over the course of the studies, and my role model was you (also, you are the hero of Chapter 3 thanks to that emotional tea break at our office). Tania, there are so many things that we understand in each other that, I feel, no one else can exactly relate to. The feeling I was not alone was crucial to be able to finish.

Both had their word of thanks in the academic part: Ioannis Douros and Amal Houidhek. You belong here too. Thank you.

My Patronus goes to Harun Šiljak for being on the same wavelength, wave amplitude, wave skewness and wave everything as I am, for being there for me, for making me believe again, and for showing me that life does not end with a PhD.

A bear hug is given to Serafima Shcherbina, who, across all kilometers, years and circumstances, continues to be the most loyal and loving friend ever. When I get in a whirlpool, you help me remember which way is up. :O:

I would like to dedicate a special pirozhenka to my favorite piatnitsa partner Anastasia Shimorina, who is a superhero in disguise: she is the friend who will check your linguistic reasoning, write you code, translate things between all languages imaginable, organize you any event you can think of, treat you to her delicious cooking, welcome you into her place, mind and heart and then tell you she did not do anything worth of mentioning. Nastia, I know I am blowing your cover, but it needs to be done. You are so genuine and so unique. Thank you for everything.

Veronika Bolshakova. You are the spark, you are the spice, you are the dance. Without you,

everything is so bland and inert. You are the only person I know who can look in the dark, fully take in the monsters who live there and then undo them with a sarcastic remark. It is thanks to none other than you I know that one should not wait for life to accommodate to become the person they want to be; we are what we are now, and what matters now is both big and complicated things such as shared secrets and profound advice, and simple and seemingly small ones, such as who comes over for tea. All my tea is yours.

Finally, lots of love to my family, who have supported me in all ways they could, followed my journey through all the ups and downs and who only wished me the best. I am fortunate to know that there is a place in this world where someone is waiting for me, always. That place is home.

1

Introduction

1.1 Articulatory speech synthesis

As technology gets more and more incorporated in our lives and the amount of exceedingly heterogeneous, multimodal data circulating around each of us increases, it becomes essential to learn to switch between data modalities and convert one into another. One of such types of conversion is that of written text into the sound of speech: text-to-speech synthesis, or TTS.

In the modern society, this artificial production of human speech, along with a constellation of other speech technology tasks such as speech recognition and understanding and speaker recognition and verification and others, is very well adopted and deeply rooted in everyday use: we hear synthetic speech in our cell phones and other gadgets, company hotlines heavily rely on it, public announcements are synthesized. It may appear that the presently attained naturalness and intelligibility of speech synthesis does not leave much room for scientific pursuit. However, there still are quite a few challenges, among which one can identify low-resource and/or small-footprint speech synthesis, speaker or domain adaptation, expressiveness, treating the multimodal aspects of speech within the scope of a specific application or for particular gains and embracing speech synthesis research results within the umbrella domain of speech studies.

The last two points are where we can situate the work on articulatory speech synthesis: the task to synthesize not only the speech wave but also the movement of the articulators that could cause it.

1.2 Motivation

As pointed out above, at present TTS is rightfully ubiquitous, its applications ranging from the most mundane such as reading out loud your text message while you are busy driving to the more obscure such as simulating Ötzi's voice—the voice of a man who lived between 3400 and 3100 BC and was found on the border between Austria and Italy in 1991 [ACFS17]. However, there still are domains that yet are to fully explore its benefits.

One such domain is language learning. As is commonly believed, around half of the world's population speaks more than one language [Eur17, Rya13]. In the case of English alone, there are 379.0 million native speakers and 753.3 million learners [ESe19]. These statistics show a dire need to streamline language learning process so that human teachers, native and non-native speakers alike, can provide more impactful help to their students. There are aspects to

their work that cannot—at least at present—be done by a machine: building rapport between the teacher and the student, getting creative with teaching methods, providing original, non-mechanic feedback to students' work. It is crucial that teachers have enough time to dedicate their effort there. Unfortunately, they often find themselves tasked to prepare study materials and exercises, while technology should actually be already able to take over at least a part of this duty.

One of the problems to solve in this direction is, indeed, the conversion of text into speech and back, as suitable for the field. Nonetheless, the context of language learning dictates the need to take into account a deeper understanding of how human speech is produced: for example, when evaluating a learner's speaking, not all pronunciation mistakes should be treated alike since some attempts are phonetically closer than others even when acoustically they may be quite different; besides, when correcting those pronunciation mistakes, it may be necessary not only to give a correctly performed sound, but also to guide the student through the articulation of the sound they are learning. Unfortunately, the two main branches of approaches of TTS—parametric and concatenative speech synthesis—both are quite technical solutions with, as of now, no easy way to introduce any information on the way the speech organs could move.

Another domain with a window of opportunity is medicine. When investigating a speaking problem that is not easy to see or capture, there is a glaring absence of tools for a speech therapist to simulate the suspected dysfunction in the vocal tract and see whether it would lead to the phenomena they actually observe in the patient. Also, when planning a surgery that could affect the patient's speech, at present there is no completely reliable way for doctors to make a prognosis on its eventual quality, while naturally this question is a big concern for the well-being of the patient. Finally, it has been shown that for many patients, the treatment of such speech problems as misarticulation [BHS14, PMRC+14, PLM17], apraxia [MKF+10, KMG10, PBL13] or Broca's aphasia [KBC99] may be facilitated with biofeedback, generalisation and maintenance being especially promising in the case of children. Such treatment requires regular sessions on site, which makes intensive training challenging. With acoustic-to-articulatory inversion and articulatory speech synthesis, however, the sessions would not require anything but an easily portable piece of software or even online access to a cloud computer, thus giving this method an unparalleled boost.

On top of such practical applications, there also are deep research questions in speech studies that have remained open for decades at least. For instance, what are articulation units and how do we process them mentally? Do we have any concept of articulatory targets and what happens to them in fluent speech, when neighboring sounds influence each other and there is not enough time to complete speech movements as articulately and precisely as we could do for a standalone sound? What are the consciously controlled factors in articulation, and what is just the effect of the laws of physics in play? If we have a model conceptualized with a particular outlook on these questions and it produces sufficiently natural and intelligible results, articulation can serve as a bridge between the biomechanics of speech as a mechanical process and its linguistic aspects and give us a clue; and vice versa, if a model is built fully compatibly with how speech scientists understand speech production and it gives unexpected results, it is very reasonable to revisit the knowledge and look for alternative theories. The problem is, state-of-the-art TTS systems are built with no relation to any speech production theories and cannot help give a satisfactory answer to these questions.

All the points above is what makes me argue that even while the quality attained by present-

day TTS systems is unprecedented and speech synthesis can sometimes be on a par with a real speaker, we need to proceed further and augment the sound of speech with other modalities, specifically articulation.

It must be pointed out, however, that articulation mostly happens hidden from the view, which makes data collection difficult. Some methods are dangerous, some are too invasive; some are not, but make it too uncomfortable for the subject speaker to speak; some are not fast enough, and some manage to keep up with the speaker's movements, but only in some specific points. Whenever a method is efficient, it seems to also be expensive and come in with a herculean task of data annotation.

On top of that, treating articulation within most speech production theories induces systems with a highly convoluted pipeline that is hard to tune. Propagating errors can easily accumulate to the point where the synthetic speech is of underwhelming quality.

The result is that despite the need for adopting articulatory aspects in speech synthesis, the topic has been out of the spotlight in speech research community. Fortunately, with the advent of considerably more informative, detailed, versatile and voluminous articulatory data such as real-time magnetic resonance imaging (RT-MRI), the field is in its prime to tackle these issues.

1.3 Problem statement

The present dissertation aims to work towards a full-fledged articulatory speech synthesizer, capable of treating a variety of speech phenomena and faithful to the effects of influence of nearby sounds on each other, coarticulation.

An input of such a system is text. The generated output is speech and the synchronous articulatory information related to it.

The objective is to explore two kinds of approaches: one drawing on the idea of frozen articulatory targets, based on a rich system of rules governing transitions, and another on treating articulatory parameters statistically, with no explicit knowledge on their interpretation put inside the system; then, to make a link between the two.

These goals constitute a multi-step piece of work, which induces a number of sub-objectives:

- Identify an appropriate type of data for each of the objectives and prepare the data accordingly.
- Build the text-to-articulatory speech system in accordance with the principles of speech production and the approaches commonly undertaken in speech synthesis community.
- Evaluate the results visually, perceptually, acoustically and numerically — not only separately for the two approaches, but also meaning to compare them.
- Draw conclusions on the relation between the two approaches and compare their efficiency.

1.4 Contribution

This dissertation starts with the work on articulatory speech synthesis from static magnetic resonance imaging (MRI) data capturing context-aware articulatory targets, presented in Chap-

ter 3. The presented method is primarily rule-based, essentially solving the problem of how to transition between captured static context-informed vocal tract configurations.

Then Chapter 4 follows with deep neural networks (DNN)-based parametric articulatory speech synthesis based on real-time MRI (RT-MRI) data. Here, no concept of articulatory targets exists, and articulation is synthesized the same way as speech in parametric speech synthesis.

Chapter 5 builds a bridge between the two approaches to investigate the relation between the static and dynamic data used in the work of the previous two chapters.

Finally, Chapter 6 gives a final overview of the results of this work and the lessons to be drawn from it, as well as its potential extensions.

The publications made over the course of the PhD studies:

- **Tsukanova, Anastasiia**, Ioannis 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." *International Congress of Phonetic Sciences*. 2019.

- Douros, Ioannis, **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." *INTERSPEECH 2019*. 2019.

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

- Laprie, Yves, Benjamin Elie, **Anastasiia Tsukanova**, and Pierre-André Vuissoz. "Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech." *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018.

- **Tsukanova, Anastasiia**, Benjamin Elie, and Yves Laprie. "Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets." In *International Seminar on Speech Production*, pp. 37-47. Springer, Cham, 2017.

- Laprie, Yves, Benjamin Elie, Pierre-André Vuissoz, and **Anastasiia Tsukanova**. "Articulatory model of the epiglottis." In *The 11th International Seminar on Speech Production*. 2017.

2

Background and context

2.1 Speech production

Speech production can be seen as a manifestation of the compromise between the complexity of what we, as human beings, need to be able to express, and the limited capacities that are at our disposal. The processes involved in speech production are intricate just enough for the simplifications and shortcuts not to hinder achieving the purposes of communication. The two contradictory forces, one to simplify the underlying processes of speech and the other to have enough subtlety to deliver messages as nuanced as they were intended, are what drives language change at all of its levels.

This dictates what we expect to learn in speech production: we shall find evidence of a certain variation and fluidity that mean to reduce the strain of the process of speech, and that variation will constantly test the physical, biological and cognitive constraints faced both by the speaker and by their listener.

Generally speaking, one can identify several subprocesses in producing speech:

- *Conceptual thinking*: the thought process that predates speech and generates the speaker's intent;
- *Lexical and grammatical selection*: filtering the speaker's intent through the available means of the language;
- *Movement planning and execution*: preparing and executing coordinated commands for speech organs: to control the breath, operate the vocal folds and put the articulatory organs where they need to be to produce the desired output;
- *Movement correction*: processing mismatches between the intended, possibly updated, outcome of any of the layers above, and the speech that is produced; correcting them.

In running speech, all these subprocesses occur at overlapping times, and it is through carefully controlled experiments that we can disassociate them. Such experiments can be cognitive, neuroscientific, articulatory and acoustic in nature, and they deal with such parameters as feature timing (e.g. voice onset time—VOT), articulator positioning, phonetic feature presence or absence; the evidence can come in the form of timing differences and changes in the place and manner of articulation. One particularly illustrative case of changes is speech errors.

The contributions of the present thesis are situated at the third level, movement planning and execution. However, other elements also need to be examined and taken into account. The levels of conceptual thinking and of lexical and grammatical selection are discussed in Chapter 2.1.1; the level of movement planning and execution in Chapter 2.1.2; finally, movement correction in Chapter 2.1.3.

2.1.1 Accessing language resources

Speech on its own does not signify anything. It is a collection of sounds. What makes it meaningful is the experience of the environment, its relation to the speaker's inner world: what it makes them think and feel, and how what they say can influence and change it [SNA13].

How do we transform the sense of self and the desire to interact with the world in a certain way into the act of speaking? It all starts with the brain.

The brain is an organ in the head that serves as the center of the nervous system. It is divided into vertical halves called hemispheres and comes out of the brain stem, the passage to the spinal cord. The matter covering hemispheres is called the cortex, and it is what provides for higher cognitive functions. The evolution of the brain preferred the cortex to have a larger surface area while occupying less space, which is why this tissue is laid in folds and has grooves and wrinkles. We can identify some partitioning in the brain: first into lobes (frontal, temporal, parietal and occipital) based on physical features, and then we can single out specific areas that serve an identifiable function.

Generally speaking, in the majority of people the dominant hemisphere is the left, and it deals with language, logical and analytical operations and mathematics. Meanwhile, the right hemisphere operates emotions, recognizing faces, perceiving structures globally (without detailed analysis), music and non-linguistic sounds. However, when we take a more detailed look, we find that this division is not as clear-cut, and both hemispheres perform some functions associated with their counterpart [SNA13].

One important region for speech production is situated in the lower back part of frontal lobe of the dominant hemisphere (typically, the left). It is called Broca's area, named after a French pathologist and neurosurgeon Pierre Paul Broca (1824–1880) who studied it. Once the stimulus for creating an utterance is formed—already with or yet without the syntax, depending on the complexity of the sentence [KS02],—this is where speech is formulated [SNA13].

If Broca's area is damaged, it causes a type of non-fluent aphasia that is called Broca's aphasia. In most cases speech becomes telegraphic and disfluent, made up of content vocabulary words stacked together instead of being properly joined into syntactically correct sentences (for example, "*Carry, couch, Sophie*" instead of "*Sophie and I had to carry the couch.*") The patients feel they know what they want to say, but cannot get it out, neither orally nor in writing. It may even happen, such as in the historic case of Leborgne, that whenever the patient tries to speak, only a limited combination of sounds will come out (such as the word "tan"). Speech comprehension is partially impeded as well, despite the fact that this is not the area primarily responsible for it. [DPIZC07, SNA13]

Indeed, a region that proved to play a major role in understanding of speech is Wernicke's area, named after a German neurologist Carl Wernicke (1848–1905). Just as Broca's area is right next to the motor cortex, Wernicke's area, being situated in the upper back part of the temporal lobe and extending upwards into the parietal lobe, is closely connected to the auditory area in

the temporal lobe through fibers of the arcuate fasciculus. The sound of the word is transmitted from the ear to the auditory area, and then to Wernicke's area for processing [SNA13].

And back in reverse for speech production, it is thought that this is in Wernicke's area where the basic structure of the utterance is constructed before being passed to Broca's area discussed above [SNA13]. Damage to Wernicke's area, especially in the left hemisphere, causes Wernicke's aphasia, which is fluent. Patients tend to speak in dragged out sentences that have no meaning, and sometimes they also use made up or irrelevant words. Their utterances make sense to them, and they do not see their speech errors.

The two flows of information—for production and for perception—form two streams: the ventral stream (speech comprehension) and the dorsal one (speech production) [HP00, HP04, HP07]. They are asymmetric: speech comprehension essentially depends only on the auditory system [SVH19], while speech production also relies on somatosensory-motor systems, using a mapping from acoustic speech signals to parietal and frontal lobe articulatory networks. The function of the left and right hemispheres of these two flows differ: in speech comprehension, despite playing different roles (understanding the meaning and treating the intonation, respectively) [FA04, WECT19], both hemispheres are involved, and in speech production, the left hemisphere dominates heavily.

Effectively, after the lifelong training side by side and assisting each other all along the way, the deep levels of memory, speech production and speech comprehension processes get tightly intertwined.

2.1.2 Movement planning and execution

Once formed in Broca's area, the activation is transferred through the nerve fibers of the arcuate fasciculus to perform phonological encoding in left frontal cortical regions, including the operculum, insula, lateral pre-motor cortex and anterior supplementary motor area [BGHV01]. The motor cortex controls the muscles corresponding to the vocal tract: the tongue, lips, jaw, soft palate, vocal cords and others [SNA13]. Here the mechanics of speech is controlled: it is produced through managing all speech organs together with pacing the breath. It is remarkable that this system originally developed for breathing and eating rather than for consciously making any kind of meaningful signals, which brings an additional contribution to the complexity of the speech production organization in the brain.

Fig. 2.1 schematically illustrates a mid-sagittal section of an adult's vocal tract, with the organs involved in speech production. As said before, these organs move according to the control from the brain, but how is this control organized?

One of the central issues in speech production is the notion of an elementary speech movement. From the auditory perspective, the elementary unit is a phoneme: the smallest unit of sound capable of changing the meaning but not bearing any meaning of its own. With a certain agreement rate, boundaries between phones can be picked through the temporal analysis of the spectrum frequencies quite certainly. The time spans of transition from one phone to another that are truly unstable and impossible to identify one way or the other generally are very short (more on the physical point of view on speech and how it is carried out is to follow in section 2.1.2). This is not at all the case for movements. Let us take sequence /fa/ as an example (Fig. 2.2). In the spectrogram, the presence of fricative noise /f/ is very distinct from the regular frequencies of the vowel /a/; this is due to the threshold effect triggering the

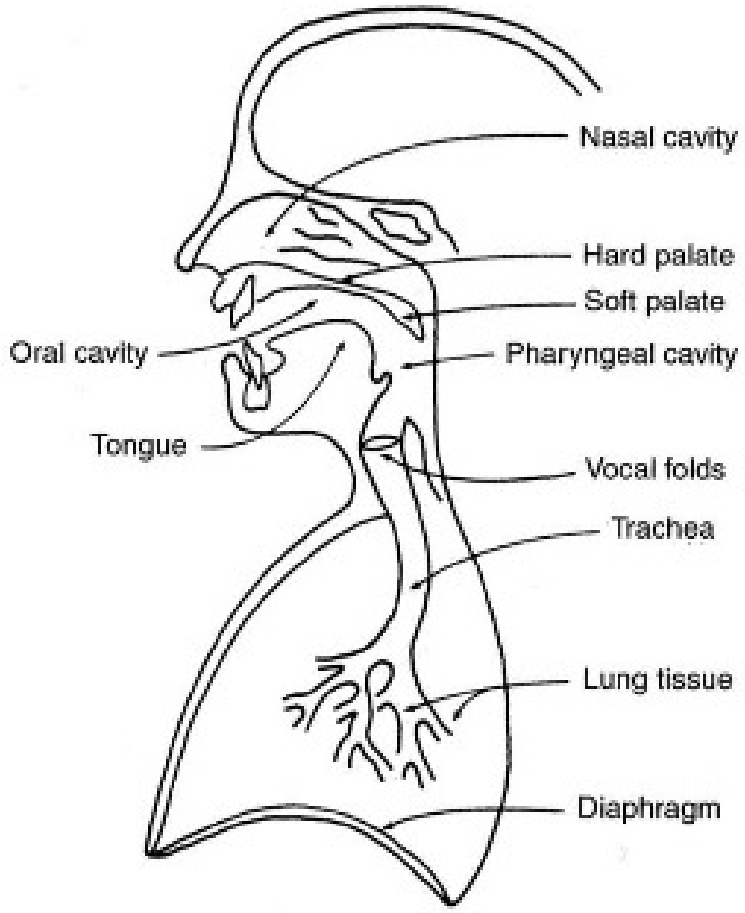


Figure 2.1: Human speech mechanism [Zem10]

turbulence in the vocal tract. However, within the actual vocal tract, the movement is smooth. It could be analyzed in two principal ways:

- This syllable can be seen as stored as two target configurations of the vocal tract: for /f/ and for /a/. Then the transition is just a gradual transformation of the first shape into the second.

- What is stored is the way we carry out this transition, the gesture for it on the level of individual articulators that are, however, necessarily coordinated. This gesture is executed just enough to produce the desired sensorimotor outcome.

Both views have given rise to a variety of theoretical and applied models discussed in Chapter 2.2. No matter how the speech organs are operated, the final result is that at every single moment of speech or preparation for it the speaker attains their particular configuration. The following subsection, Chapter 2.1.2, discusses what happens then.

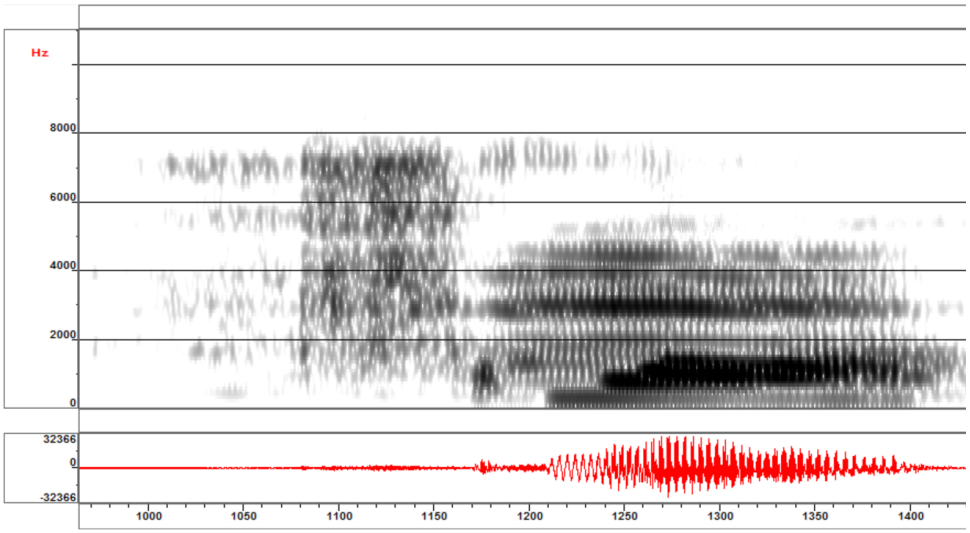


Figure 2.2: An example of a spectrogram: /fa/. The boundary between /f/ and /a/ can be seen at around 1170 ms: the fricative noise stops, formant bands appear, and the waveform exhibits the onset of voice.

Emitting speech sounds

The crucial elements for producing a sound of any kind are a source of an acoustic wave, a propagation medium, and the presence of this medium's boundary. For humans, this is their respiratory system (the source of pressure), the air, and the vocal folds (a vibrating element) along with the vocal tract that is able to produce constrictions. In terms of acoustic and electrical engineering, it means that we may describe the speech wave in terms of the source and filter characteristics: the human vocal tract is a sound-emitting filter system that responds to one or more sound sources, which can be written as the following equation:

$$| P(f) | = | U(f) | \cdot | H(f) | \cdot | R(f) |, \tag{2.1}$$

where f is the frequency, $|\cdot|$ is the module function, $|P(f)|$ is the sound pressure spectrum at a distance from the mouth opening, $|U(f)|$ is an amplitude versus source frequency characteristics, namely volume velocity spectrum, $|H(f)|$ is the frequency-selective gain function of vocal transmission, and $|R(f)|$ is the radiation characteristics at the lips converting volume velocity calculated on the mouth opening into sound pressure. This vocal tract interpretation is the foundation of the *source-filter theory* of voice production [Fan71a].

Respiration

The respiration system of a human involves organs such as the trachea, rib cage, thorax, abdomen, diaphragm, and lungs. The mechanics of breathing is largely explained by *Boyle's law*, which states that if a gas is kept at a constant temperature, pressure and volume are inversely proportional to one another and have a constant product; it means that we can regulate the pressure of the air in the lungs by expanding and reducing their volume, sending the air into and out of the lungs [Zem10].

At rest, the pressure within the lungs (alveolar pressure) amounts to the atmospheric one, and the diaphragm, which is the principal muscle of inhalation and an anatomical divider

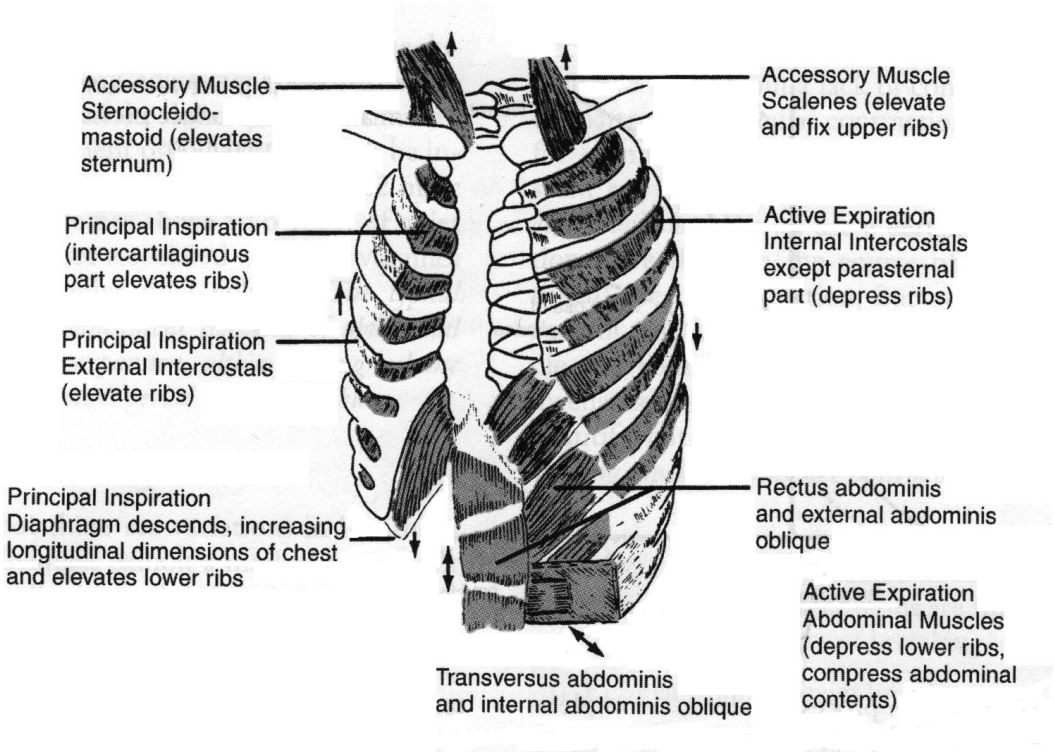


Figure 2.3: The principal muscles of respiration [Zem10]

between a thoracic and an abdominal cavity, is not tense, pertaining the form of an inverted bowl.

Then, the exterior air is drawn into the organs of respiration, i.e. the lungs, by contraction of the posterior and anterior muscle fibers that draws the central tendon downward and shifts it forward, thus expanding the chest cavity, lowering the diaphragm, and elevating the pressure in the abdominal cavity (see Fig. 2.3). With an increase of thorax volume, as the lungs press upon the walls of the thorax thanks to subatmospheric pleural fluid pressure, the pressure in the lungs becomes negative with respect to the atmosphere [Zem10]. The air goes down the respiratory tract, freely and directly passing the oral and nasal cavities, pharynx, larynx, trachea, and bronchi; the inhalation muscles gradually relax, activating the passive forces of exhalation.

Once the outside and inside pressures are in balance due to the natural physical limitations of the trachea and bronchial tree, relatively high pressure in the abdominal cavity tends to restore the relaxed shape of the diaphragm as well as the ribs and soft tissues. With this air inhaled, the resource for producing speech is available; it is used either for silent exhalation or for speech, and the flow volume will be proportional to the difference between atmospheric pressure and the pressure within the lungs.

To breathe out or actually speak, humans expel the drawn air by contracting the rib cage which decreases the volume of the thorax and consecutively increases the pressure in lungs (the greater the lung pressure, the louder and more high-pitched sounds come) and pushes the air out, up through the trachea, into the pharynx, throat cavity [Fla13]. The alveolar pressure falls from around 40 cm H_2O when exhaling passively or as high as 200 cm H_2O when adding a muscular effort to negative values at low lung volumes. As for speech production, it requires an

airflow brought upon by an alveolar or subglottal pressure in the range of 5 to 20 cm H_2O. Speech can occur at highly variable lung volumes (we may even speak "out of our breath", for example, when a sentence ends up being longer than intended first) [Zem10]. Most of the utterances are pronounced on expiration; ingressive sounds, for which the airstream flows inward through the mouth or nose, are rare [LM98]. So in order to avoid exhausting the drawn air amid an utterance, humans have to regulate their alveolar and subglottal pressure and check themselves, contracting the inspiratory musculature to make the expiration slower. It is this system of rational usage of inspiratory and expiratory muscles what allows us to speak, altering the vital process of natural breathing.

Phonation

When leaving the trachea, the air passes the larynx. Its cartilages hold two folds of ligament and muscular tissue which are called vocal folds.The opening between them is the glottis and serves as a gate for the airflow. Due to their mobility, the vocal folds are a source of highly variable resistance for the air flow which instigates the speech sounds. When the orifice between the vocal folds is closed, it means that there is a source of greater resistance on the way of the air flow, and at least some of the air will come back, raising the alveolar and subglottal pressures even higher. Consequently, the air flow will get only heavier, until it finally forces the vocal folds apart, letting the airflow pass through. Then, according to the Bernoulli's law, the local pressure falls, urging the cords close up again. With the flow reduced, the local and subglottal pressures amount to each other as in the beginning of this cycle. So, the vocal folds open and close rapidly on loop for a speaker to produce voiced sounds, i.e. to phonate. This defines the period of the oscillation forced onto the cords. In contrast, to produce an unvoiced sound, the vocal folds neither close together nor vibrate—they stay open instead.

The rate of vocal fold vibration is described as voice musical tone, or pitch (perceptually), or as fundamental frequency measured in Hz—cycles per second (physically). The basso voice corresponds to 60 Hz or lower or B_1 in the musical scale, and by raising the voice up to the soprano register one will reach the frequency of over 1568 Hz, or G_6 [Zem10].

For every individual speaker, the quality of the voice ranges with vocal fold vibration frequencies. There is a comfortable *middle* or *modal pitch range*. At its upper limits the quality of the voice suddenly changes into the *falsetto* register, also called *loft register*, or, possibly for female soprano singers, *laryngeal whistle*. In falsetto, the contact area in the vocal folds is much smaller, and the glottis turns into a tense and narrow slit that vibrates only at the edges. The mechanism of laryngeal whistle is the same as of falsetto, but with higher tension, pressure and resulting frequency. As for the lower limits of the modal pitch range, the voice changes there into *glottal fry* or *pulse register*, which gives the effect of a creaky voice. To produce it, the vocal folds are drawn together tightly, but let subglottal air bubble up between them in discrete bursts in a syncopated rhythm [MVL58].

Mathematical models of the larynx for speech simulation include a single-degree-of-freedom model by [FL68] where the vocal folds must move as a single mass toward and away from the midline (with one degree of freedom, hence the name) which can be a simple solution but does not describe behavior of the real larynx; two-degree-of-freedom models such as by [IF72] where the vocal folds are two masses instead of a single one, capable of an independent horizontal motion which is better but not devoid of artifacts and unrealistic consequences for the parameters; and the sixteen-mass model by [Tit73] that was to take into account the mucosa in the vibrating larynx and allow more degrees of freedom for the vocal folds.

The important parameters of voice production are as follows [Zem10]:

1. Maximum pitch range: how flexible is the voice pitch?
2. Mean rate of vocal fold vibration: what is the most comfortable, habitual pitch for the speaker, in relation to the pitch range?
3. Air cost: how long can the speaker phonate comfortably without running out of air?
4. Minimum-maximum intensity at various pitches: along the frequency range, how do the sound pressure level measurements change?
5. Periodicity of vocal fold vibration: what is the natural period of vocal fold vibration when other parameters are constant?
6. Noise: are there noisy areas in the sound spectrum? How are they related to intentional voice qualities such as hoarseness, breathiness?
7. Finally, resonance: how does the vocal tract resonate when the air is propagated from the larynx to the mouth opening?

The final point concerns the positioning of the further vocal tract rather than that of the larynx, which brings us to the next section.

Articulation

So, phonation involved vibrations of the vocal folds that can essentially be summarized in parameters of frequency, intensity, and duration. To obtain the speech sound, there has to be a resonator to receive the puffs of air from the larynx. While the fundamental frequency is defined by the rate at which the air column is driven into oscillations, it is the form and dimensions of the acoustic object what establish the resonating frequencies and in such a way determine the quality of the tone.

Resonation is what the vocal tract serves for in sound emission. The vocal tract starts just above the glottis and, from an acoustic point of view, is an acoustic tube, around 17 cm long for an adult male, with a varying cross-sectional area. The vocal tract consists of the oral tract and the nasal tract and ends with lips and nostrils respectively, from where the sound is propagated in the atmosphere. When emitting a voiced sound, the vocal tract receives quasi-periodic pulses of air. Since the glottal orifice is relatively small, its acoustic impedance is dominating, and unless there is a pronounced constriction in the further vocal tract, the glottis is the main source of turbulence. Otherwise obstacles on the way of the airflow, which are made by positioning the articulators that compose a source of widely ranging resistance to the air flow (from minimal, such as for open vowels, to neutral, such as for uttering a sound like "uh", and absolute, such as the moment of constriction at the lips to produce [b]), bring out vortices—their experimental evidence in the speech airflow was provided by [Tho86]. Then these findings were theoretically supported by [TT90] and [McG88]. Vortices can occur, for instance, due to changes in the velocity of the flow at the boundaries, flow disruption such as by adverse pressure from a cavity, flow separation, or appearance of rotational motion because of moving through the curved form of the vocal tract. When formed, a vortex can twist, stretch or spread further downstream [Mar94].

Intensifying, these effects will eventually lead to turbulent flow. The characteristic that helps predict whether the flow stays laminar or becomes turbulent is the Reynolds number, noted as Re:

$$\mathrm{Re} = \frac{\text{inertial forces}}{\text{viscous forces}} = \rho\frac{UL}{\mu}, \tag{2.2}$$

where ρ stands for the air density; U is a velocity scale; L is a linear scale such as the diameter of the vocal tract; and μ is the air bulk viscosity.

Low values of the Reynolds number are associated with laminar flows—the flows where the viscous forces dominate. A high value of the Reynolds number will indicate a turbulent flow with chaotic instabilities. These effects are necessary to produce frication, aspiration and whisper or contribute to the effect of a breathy or creaky voice.

To operate the airflow and differentiate the resulting sounds of speech, humans position their *articulators* so as to make a constriction of the flow at a particular place and in a particular way. This deliberate sound formation is called *articulation*; every phoneme has a place and manner of articulation associated with it, though some degree of freedom is allowed. The phonetic details and how they play out in the language will be covered below [LD12].

Fig. 2.4 shows an outline of the vocal tract. The vocal tract comprises five[1] resonating cavities: the buccal, oral, pharyngeal, and two paired nasal ones. Since they are interconnected, the division is made from the anatomical perspective. Articulation can be formulated in terms of operating the cavities: as the speaker articulates to produce speech, the cavities can grow or diminish in volume up to complete blockage, and this changes the acoustic properties of the vocal tract, namely the resonant characteristics. The result is a sound as intended by the speaker, with a correct energy distribution (how much energy concentrates at which frequencies—both aspects are perceptually important).

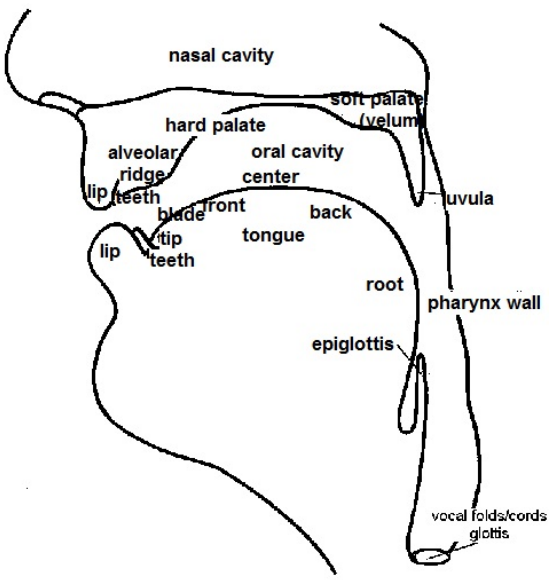


Figure 2.4: Human vocal organs [LJ14]

- The *buccal cavity* is the space extending from lips and cheeks to the teeth and gums. It is connected to the oral cavity through the space between the teeth and behind the last molars. Its volume varies between subjects, but is small.

- The *oral cavity* is bounded by the roof of the mouth, the teeth, the glossopalatine arch, and the muscular floor which is mostly the tongue. Due to the tongue's flexibility and high

[1] The figure can get higher if we factor in smaller cavities and areas that are not clearly separated from some of the cavities' proper, such as two small cavities in form of a deep depression, lateral to the aditus laryngis, that are situated at the bottom of the pharynx and called *pyriform sinuses*, or the *sublingual cavity* under the tongue. However, they do not seem to play a crucial role in speech production.

mobility of the lips, during speech the volume of this cavity varies greatly. This cavity communicates with the pharyngeal and nasal cavities through a port called pharyngeal isthmus, bounded by the anterior faucial pillars, soft palate, and the dorsum of the tongue.

- The *pharyngeal cavity* proper extends over the pharynx (see Fig. 2.5), which is a vertically aligned musculomembranous tube, oval in a transverse section (wider in the frontal plane and more narrow in the sagittal one) and reaching from the level of the sixth cervical vertebra (the posterior position) and the cricoid cartilage (the anterior one) to the base of the skull. The mucous membrane of the tube continues into the one of the nasal cavity. We define three major regions in the cavity of the pharynx: the *nasopharynx*, the *oropharynx*, and the *laryngopharynx* (see Fig. 2.6).

- The *nasal cavities* are two approximately symmetrical chambers with the nasal septum between them. Anterior nares are nostrils, the way from the nasal cavities to the exterior. Posterior nares are choanae, the way to the nasopharynx. The superior, middle, and inferior nasal conchae, arranged in a labyrinth-like way, along with their nasal passages comprise lateral walls of the cavities.

There are *passive articulators* that cannot change their position—the upper jaw, the hard palate, the teeth—and *active articulators* that are free to move: the lower jaw, the lips, the tongue, and the velum.

The *mandible* is a very important articulator—the only truly movable bone in the face. Not only does it differentiate vowels by the degree of openness, it also helps enunciate sounds better. The jaw mainly rises and falls, though it can also be protruded and retracted, or make a grinding motion. The jaw is set on the temporomandibular and ginglymoarthrodial joints.

The vocal tract ends with an orifice formed by the *lips*. They consist of muscular and grandular tissues and some fat. Of the two, the lower lip is faster and more mobile. To participate in speech production, the lips can close and open and protrude and retract.

The *tongue* is the most flexible vocal organ with a great degree of freedom, which plays such an important role in speech articulation that in many languages the word "tongue" has become synonymous to either "language" or "speech". In the tongue, we discern the tip, blade, and body, and the tongue body is divided into the front, center, and back further then. The tongue moves thanks to an intricate and complex system of extrinsic and intrinsic muscles, usually one or two muscles dominating in a particular motion, and other muscles gradually taking charge when it is their order. [Har76] sees most tongue movements as composed of just seven components: horizontal and vertical forward-backward and upward-downward movements of the tongue tip and the tongue blade, two parameters for the transverse cross-sectional configuration, and the form of the tongue dorsum plane—spread or tapered.

In the roof of the mouth, we single out a small ridge behind the upper front *teeth* called the *alveolar ridge*, *hard palate*, and *soft palate* (also called the *velum*). The velum can block the airflow from passing into the nasal cavity, which is necessary to produce an oral sound (then the velum goes up) or let the air through (then the velum goes down, and the sound is nasal). A fleshy extension at the back of the soft palate which hangs above the throat is called the uvula.

Behind and below the back of the tongue comes the pharynx, which can also be an articulator or be ignored by the language: for instance, French, on the materials of which the present dissertation is built, does not feature any pharyngeal sounds that would be significant on the

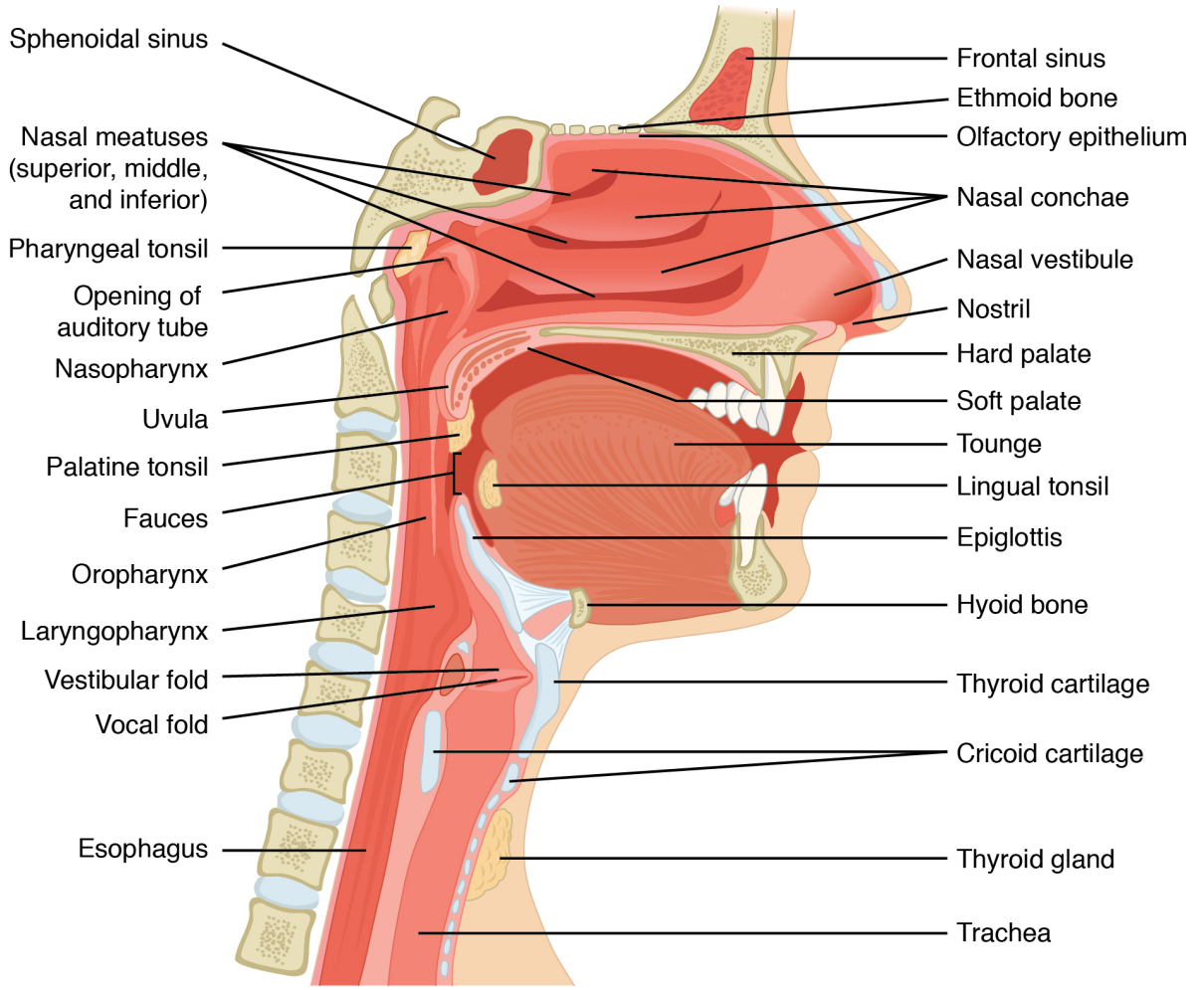


Figure 2.5: Upper airway [Ope13]

nominative level. Then there is a flap of cartilage behind the root of the tongue, which is depressed during swallowing to cover the opening of the windpipe—this is the epiglottis.

Places and manners of articulation

The two subdisciplines of linguistics which study sounds of human languages are called *phonetics* and *phonology*. While phonetics deals with the physical and physiological aspects of sounds, phonology treats sounds as parts of a particular language, disregarding the information that is linguistically irrelevant. Let us discuss the overall picture first, and then go further into the specifics of the phonetics of the language treated in this dissertation, French.

The smallest distinctive unit of speech is called a *phoneme*. It does not carry a meaning of its own, but can distinguish at least one word from another in the particular language where it belongs to. The manifold of all acoustic variations of a phoneme comprises its *allophones*.

Phonemes of a natural language are usually not in one-to-one relation with the written system of this language, and thus phoneticians—[Int99]—have introduced a special alphabet to note transcriptions.

Phonetic symbols are enclosed in square brackets [ ], and phonemes are enclosed in virgules

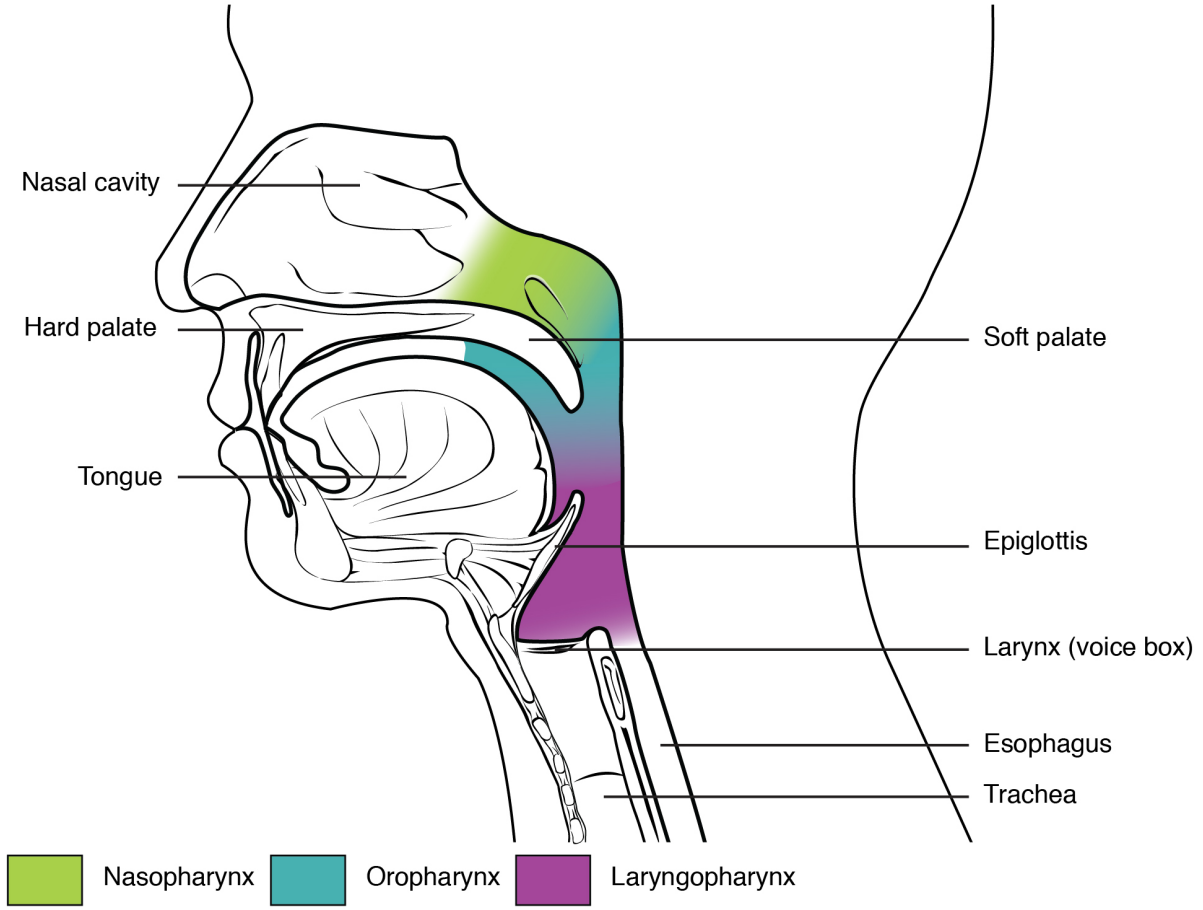


Figure 2.6: The division of the pharyngeal cavity into the nasopharynx, oropharynx, and laryngopharynx [Ope13]

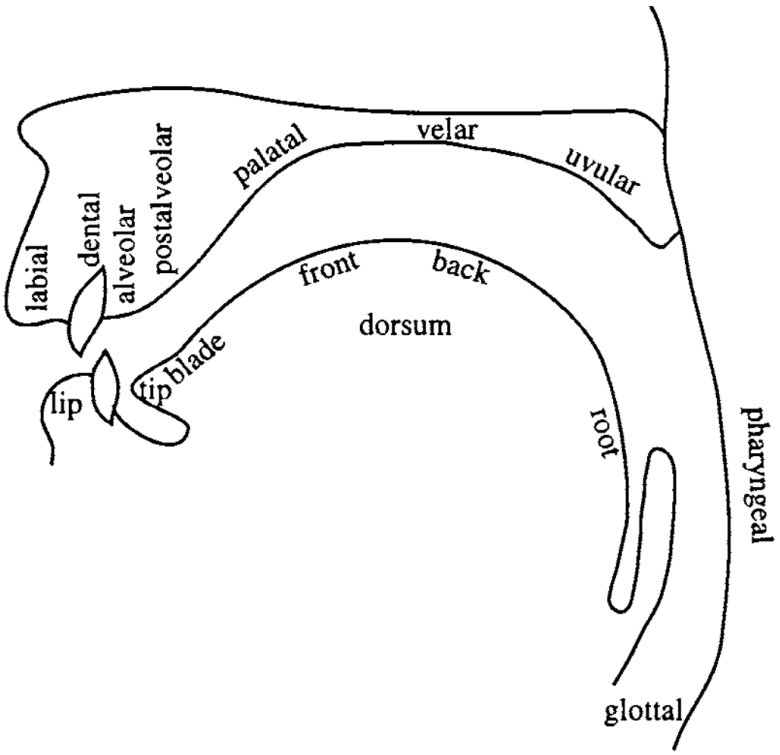


Figure 2.7: Mid-sagittal section of the vocal tract with labels for place of articulation [Int99].

//.

All sounds of natural languages are divided into vowels and consonants.

Vowels are articulated with an open vocal tract where the air flows virtually unimpeded. All organs of speech are tense, including walls of resonating cavities. The air stream is relatively weak.

On the other hand, *consonants* are pronounced by means of creating an obstacle on the way of the air stream. To get past the obstacle, the air stream has to be heavy. Only those articulators that are responsible for the place of constriction are tense, and the others are lax. Consonant production depends on fast and imperatively precise motions of articulators.

Usually the transition from voiced consonants to vowels in terms of the degree of constriction is gradual: the language's phonemic continuity makes sure that there are no abrupt jumps in the openness / closure range [LJ14].

Vowels can be classified [Int99]:

- *by height*: the lowest resonance of the voice—the first formant, associated with the vertical position of the tongue with respect to the roof of the mouth, or, alternatively, the degree how open the jaw is. The more open a vowel is, the higher is the frequency of F1. Vowels can range from *close* ones—when the tongue is close to the roof of the mouth—to *open* ones—when the jaw is low, open. There are seven degrees of vowel height.

- *by backness*: defined by the second formant of the voice, associated with the position of the tongue relative to the back of the mouth. The more front a vowel is, the higher is the

frequency of F2. Vowels can range from *front* ones—when the tongue is forward in the mouth—to *back* ones—vice versa. There are from five to seven degrees of vowel backness.

- *by roundedness*: defined by the third formant of the voice and indirectly associated with the rounded or unrounded lips.
- *by nasality*: vowels can be *oral* or *nasal*, depending on whether the velum is raised or lowered and whether the nasal tract is participating in the vowel production.
- *by movement of the tongue*, *by voicing*, *by secondary constrictions*, *by tenseness*...

The vowels' chart is given in Fig. 2.8.

Naturally, the fact that vowels can be arranged both in the articulatory and the acoustic spaces has lead to recognition of *cardinal vowels* as the extreme ones that all others can be compared to. [Jon56] suggested a set of eight vowels, and [CH68] updated this notion. Three vowels in this set, /ɑ/, /i/, and /u/, have articulatory definitions outside of the scope of any particular language on Earth and may be called *corner vowels*, and the others are arranged between them so that they divide the acoustic space into even-sized areas. Then vowels of all languages can be set in this vowel space, expressed through the *cardinal* ones.

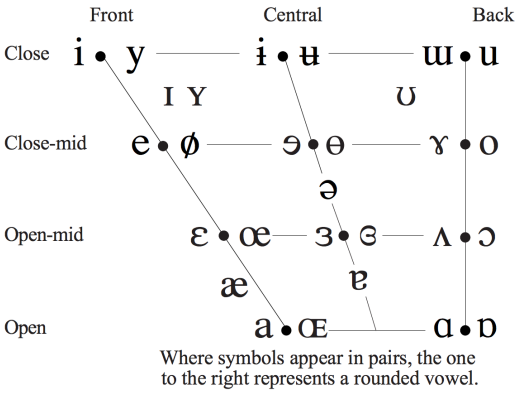


Figure 2.8: The vowels' chart based on their main features [Int99].

Consonants can be classified [LJ14, Lav94, Int99]:

- *by voice* into *voiced* and *voiceless*: the ones during which the vocal folds vibrate and the ones during which they do not;
- *by place* (see Fig. 2.7): at each articulator constriction, a speech sound can be formed, which may or may not belong to the language phoneme inventory. Consonants articulated by the lips are *labial* with further distribution into *bilabial*, *labiodental*, *dentolabial* depending on how the sound is produced (with both lips, with the low lip against the upper teeth, or vice versa), and others.

  Consonants articulated by the tongue are, in case of the raised tongue apex and blade, called *coronal*, and in case of the tongue dorsum, *dorsal*. Depending on where the tongue tip must be put, among others, we apply the corresponding terms: *linguolabial* (the tongue tip in contact with the upper lip), *interdental* (the tongue comes between the teeth), *dental* (the tip of the tongue or the tongue blade comes in contact with the back surface or bottom of the top teeth), *denti-alveolar* (the tongue touches the upper part of the back surface of the top teeth), *alveolar* (the tongue is in contact with the alveolar ridge), *postalveolar*, *retroflex*, *palato-alveolar* (the tip or the blade of the tongue comes in contact with the back area of the alveolar ridge). For the tongue dorsum, the applicable terms are *palatal* (when the front of the tongue articulates with the domed part of the

hard palate), *velar* (with the soft palate), *uvular* (with the very back of the soft palate and uvula).

When the sound is produced with the tongue, there is a further dichotomy depending on the tongue shape. If the air flows across the centre of the mouth over the tongue, the consonant is called *central*. If there is a constriction at the center of the tongue and the air parts to flow along the sides of the tongue, such a consonant is *lateral*.

Consonants pronounced at the pharynx are *pharyngeal*, produced by the faucal pillars moving together or raising the larynx.

Glottal articulation occurs directly on the vocal folds.

- *by manner* [LJ14, Lav94, Int99]:
  - *Fricative* consonants: produced by excitation of a noise by means of narrowing the vocal tract at the point of articulation without a complete obstruction of the airway. This obstruction generates a turbulent air flow, which is perceived as a slightly hissing noise. Common ways to make a fricative are to make the tongue approach the teeth or the alveolar ridge or make the lower lip approach the upper teeth—actually, in general, any other two articulators that can come close enough to each other.
  - *Stop* consonants: produced in three phases:
    * *Catch*: articulators come into the contact, making a complete closure in the vocal tract. This closure can be labial, alveolar, palatal, velar, and glottal.
    * *Hold*: even if the point of contact is not immobile, the articulators stay tightly locked and do not leak the air. The articulators are tense. The air is being accumulated. The pressure rises;
    * *Burst*: the speaker lets the pressure force the articulators apart, and all the drawn air gets momentarily released. This explosion is the most helpful perceptual cue to identify a stop consonant.

    Stops can be either *oral*—then their behavior is as described above—or *nasal*. Nasal stops are produced with a constriction somewhere in the oral cavity, after the velopharyngeal port. From the temporal point of view, they are just like oral stops. But since the velum is opened wide, the air is not accumulated as much, since it goes to the nasal tract and radiates from the nostrils.
  - *Affricate*: a stop immediately followed by a fricative.
  - *Approximant*: one articulator is close to another, but the narrowing is wide enough to avoid a turbulent airstream.
  - *Glides and semivowels*: the glides are dynamic sounds that are produced on the vowel they precede, and semivowels are very much alike vowels, but only with a greater degree of constriction.
  - Others: *trills, taps* or *flaps, clicks*...

Phonetics of the French language

As we said, the section above described the general picture of how the phonetic organization of a particular language may look like. This is important to consider to produce speech synthesis

systems that are not for one-case use and can be generalized for other languages. However, every language has its own specifics; an issue that may be negligible for a generic language in the grand scheme of things, can be crucial in this particular case. This is why it is fundamental to also go into details relevant for the language in question, French.

The phonetics of the French language has the following distinguishing characteristics:

- A relatively large number of vowels (fifteen);
- Four degrees of vowel openness: open, open-mid, close-mid, and close;
- Most vowels are open, and most vowels are rounded. In fact, lip protrusion can be extremely pronounced, such as in the case of /y/;
- There are both oral and nasal vowels;
- The vowels are articulated very clearly;
- Vowels retain their properties fully and are not reduced, French being a syllable-timed language;
- Before a pause, consonants are pronounced very clearly, often with a trace of /ə/. They are not unvoiced like in some other languages such as Russian, German or some accents of English;
- Assimilation of vowels by openness;
- Assimilation of consonants by voicedness;
- Absence of assimilation by place and manner of articulation;
- Liaisons (*Fr.*: liaison) and linking (*Fr.*: enchaînement)—two phenomena which make sure French speech remains connected when, based on the lexical choice alone, it would not be the case;
- Certain patterns in accentuation and rhythmics.

Phoneme inventory of the French language

There are fifteen vowels in French [Lon84, Cal89a], eleven oral and four nasal ones:

1. /i/: "qui", *Fr.* "who", /ki/—is a phoneme because it can be contrasted, for example, with "que", *Fr.* "that", /kœ/: close front unrounded vowel;
2. /e/[2]: "dé", *Fr.* "dice", /de/—contrasting with "dais", *Fr.* "canopy", /dɛ/: close-mid front unrounded vowel;
3. /ɛ/[3]: "fait", *Fr.* "done", /fɛ/—contrasting with "fée", *Fr.* "fairy", /fe/: open-mid front unrounded vowel;

[2] There is a tendency to neutralize the difference between /e/ and /ɛ/.

[3] /ɛː/, as in "fête", *Fr.* "holiday", /fɛːt/, is usually replaced by /ɛ/. However, there are rare pairs where some speakers still make a distinction in the vowel length: "mettre", *Fr.* "put", /mɛtʁ/, vs. "maître", *Fr.* "teacher", /mɛːtʁ/.

4. /a/: "**ta**", *Fr.* "your_*S_FEM*", /**t**a/—contrasting with "**te**s", *Fr.* "your_*PL*", /**t**e/: open central unrounded vowel;
5. /y/: "**tu**", *Fr.* "you", /**t**y/—contrasting with "**tou**s", *Fr.* "all", /**t**u/: close front rounded vowel;
6. /ø/: "**peut**", *Fr.* "can_*3P_SING*", /p**e**/—contrasting with "**pu**", *Fr.* "could", /p**y**/: close-mid front rounded vowel;
7. /œ/: "s**œur**", *Fr.* "sister", /s**œ**ʁ/—contrasting with "s**ûr**", *Fr.* "sure", /s**y**ʁ/: open-mid front rounded vowel;
8. /u/: "**court**", *Fr.* "short", /k**u**ʁ/—contrasting with "c**œur**", *Fr.* "heart", /k**œ**ʁ/: close back rounded vowel;
9. /**o**/: "**pôle**", *Fr.* "pole", /p**o**l/—contrasting with "**Paul**", *Fr.* name "Paul", /p**ɔ**l/: close-mid back rounded vowel;
10. /ɔ/: "**pomme**", *Fr.* "apple", /p**ɔ**m/—contrasting with "**paume**", *Fr.* "palm", /p**o**m/: open-mid back rounded vowel;
11. /ɑ/[4]: "**pâte**", *Fr.* "pasta", /p**ɑ**t/—contrasting with "**patte**", *Fr.* "paw", /p**a**t/: open back unrounded vowel;
12. /ɑ̃/: "**em**mener", *Fr.* "bring", /**ɑ̃**mə**n**e/—different from the word with an absent /ɑ̃/: "mener", *Fr.* "think", /mə**n**e/: nasal open back unrounded vowel;
13. /ɔ̃/: "**mont**", *Fr.* "mount", /m**ɔ̃**/—contrasting with "**mot**", *Fr.* "word", /m**o**/: nasal open-mid back rounded vowel;
14. /œ̃/[5]: "**un**", *Fr.* "a_*MASC*", /**œ̃**/—contrasting with "**an**", *Fr.* "year", /**ɑ̃**/: nasal open-mid front rounded vowel;
15. /ɛ̃/: "**fin**", *Fr.* "end", /f**ɛ̃**/—contrasting with "fine", *Fr.* "fine", /fin/: nasal open-mid front unrounded vowel.

Meanwhile, /ə/, while being present in the language, is not a phoneme. If it were one, adding or dropping it in speech would create new words, but it does not bring about semantic changes[6]: chemin /ʃmɛ̃ – ʃəmɛ̃/ (*Fr.* "way"), lentement /lɑ̃tmɑ̃ – lɑ̃təmɑ̃/ (*Fr.* "slowly"), and so on and so forth. For this reason, /ə/ is a stylistic variant of the phoneme /œ/ rather than a phoneme on its own [And82, Tra87], and its omission in words is elision.

To summarize the footnotes, there is a (at times regional) tendency towards neutralization in pairs /a/ vs. /ɑ/, /ɛ̃/ vs. /œ̃/, /e/ vs. /ɛ/, /o/ vs. /ɔ/, and /ø/ vs. /œ/ to the point when they can be indeterminable [FKJ06].

There are twenty consonants in the French language [Lon84, Cal89a], all of which are pulmonic egressive sounds. Twelve of them are obstruents (six plosives, six fricatives, all categorized into pairs of a voiced and an unvoiced constituent), and eight of them are sonorants (three nasals, two liquids, and three semivowels):

[4] /ɑ/ is often replaced by /a/, though /ɑ/ is preferred, for example, before /z/ or after /ʁw/ [Tra87].
[5] Fewer and fewer speakers distinguish between /œ̃/ and /ɛ̃/, especially in fluent speech.
[6] Though it should be noted that it does play a role in listening comprehension [FS97].

1. /p/: "c**a**pe", *Fr.* "cape", /kap/: oral voiceless bilabial stop;
2. /b/: "cra**b**e", *Fr.* "crab", /kʁab/ or /kʁɑb/: oral voiced bilabial stop;
3. /f/: "**f**confirmer", *Fr.* "confirm", /kɔ̃fiʁme/: oral voiceless labiodental fricative;
4. /v/: "ca**v**e", *Fr.* "cellar", /ka:v/: oral voiced labiodental fricative;
5. /t/: "**t**issu", *Fr.* "fabric", /tisy/: central oral voiceless laminal denti-alveolar[7] stop;
6. /d/: "ca**d**eau", *Fr.* "present", /kado/: central oral voiced laminal denti-alveolar[8] stop;
7. /s/: "**s**ymbole", *Fr.* "symbol", /sɛ̃bɔl/: central oral voiceless laminal alveolar dentalized[9] sibilant fricative;
8. /z/: "ga**z**", *Fr.* "gas", /ga:z/ or /gɑ:z/: central oral voiced laminal alveolar dentalized[10] sibilant fricative;
9. /ʃ/: "**ch**ômage", *Fr.* "unemployment", /ʃoma:ʒ/: central oral voiceless palato-alveolar labialized[11] sibilant fricative;
10. /ʒ/: "â**g**e", *Fr.* "age", /ɑ:ʒ/: central oral voiced palato-alveolar labialized[12] sibilant fricative;
11. /k/: "**c**cuper", *Fr.* "occupy", /ɔkype/: central oral voiceless velar stop;
12. /g/: "**g**lobal", *Fr.* "global", /glɔbal/: central oral voiced velar stop;
13. /m/: "**m**unir", *Fr.* "provide", /myni:ʁ/: bilabial voiced nasal;
14. /n/: "**n**asal", *Fr.* "nasal", /nazal/: laminal denti-alveolar[13] voiced nasal;
15. /l/: "**l**lumer", *Fr.* "light", /alyme/: lateral oral apical alveolar[14] lateral voiced liquid approximant;
16. /ʁ/: "**r**ien", *Fr.* "nothing", /ʁjɛ̃/: central oral voiced uvular liquid fricative[15];
17. /ɲ/[16]: "Bourguig**n**on", *Fr.* "Burgundian", /buʁgiɲɔ̃/: palatal voiced nasal;
18. /w/: "**ou**i", *Fr.* "yes", /wi/: the central oral labio-velar voiced approximant—semivowel glide, corresponding to the close vowel /u/;

[7] So, pronounced as [t̪].
[8] So, pronounced as [d̪].
[9] So, realised as [s̪].
[10] So, pronounced as [z̪].
[11] So, pronounced as [ʃʷ].
[12] So, pronounced as [ʒʷ].
[13] So, pronounced as [n̪].
[14] Carried out as [l̪]. Sometimes [l̪] or [l] occur. /l/ can come voiced or voiceless depending on its position in the word, but these differences do not carry any meaning.
[15] Other regional or contextual options: the voiceless uvular fricative [χ] and the uvular trill [ʀ].
[16] Currently /ɲ/ seems to be merging with /nj/ [GLM12].

19. /ɥ/: "huile", *Fr.* "oil", /ɥil/: the central oral labio-palatal voiced approximant—semivowel glide, corresponding to the close vowel /y/;

20. /j/: "yeux", *Fr.* "eyes", /jø/: the central oral palatal voiced approximant—semivowel glide, corresponding to the close vowel /i/.

Some dialects of French also feature /ŋ/. This sound also occurs in some loaned words such as "parking" and can be replaced by /ŋg/ or /ɲ/ [GGGG11].

The acoustic properties of French can be found in the work by [Lon84].

Phoneme duration and timing in the French language

The length of a speech sound depends on its neighboring phonemes, on the position within the utterance, and on the syntactic and semantic structure of the utterance [Cal89b].

Long *consonants* can be used for emphatic stress or in gemination.

When the utterance involves a co-occurrence of a consonant, the first consonant instance usually loses its third phase of production (the burst), and the second one loses its first phase of production (the catch). Then there is no pause between the consonants, but they belong to two syllables. The consonant from the syllable coda's intensity is attenuating, while on the syllable onset the intensity is rising.

In gemination of voiceless plosives, /p/, /t/, and /k/, the hold phase is voiceless. The pause before the burst is longer than in a non-geminated plosive, and the articulators are tightly locked during this pause.

To produce geminated voiced plosives, /b/, /d/, and /g/, the vocal folds start vibrating earlier: the closure is completely voiced. Hence the pause before the burst is not entirely voiceless—there is an insignificant glide.

Fricatives, liquids, and nasals—/f/ and /v/, /s/ and /z/, /ʃ/ and /ʒ/, /l/, /m/, /n/, and /ʁ/—are not silent even during the hold phase.

In French vowel contrasts, there is evidence that duration is not a reliable cue, and though in production it is normally held as per the rules of French phonology, the difference between vowel durations is not sensed in perception [GB88].

Prosody and Voice Stream Segmentation

The minimal articulatory unit is the *syllable*. Syllables are formed by vowels which can be accompanied by consonants before them (in the *syllable onset*) and after them (in the *coda*). If a syllable ends with a vowel, it is open (e.g. "répéter", *Fr.* "repeat", /ʁe.pe.te/); otherwise it is closed (e.g. "acteur", *Fr.* "actor", /ak.tœːʁ/). By the general principle, open vowels are used in closed syllables, and close vowels are used in open syllables [Sch21], and more precisely, syllable segmentation is done according to rules that differentiate between the number of consonants to be distributed between vowels, their classes, and position within the word.

The units of utterance segmentation are:

- *Rhythmic groups*: groups of words—actually, given the French connectedness of speech, groups of syllables, where syllables can overlap word boundaries and the number of syllables coincides with the number of pronounced vowels—having some sense as a whole and stressed on the last syllable [Gra50]. Rhythmic groups are separated from each other by changes in speech melody, rhythm, and duration of the stressed vowel;

- *Syntagms*: groups of rhythmic groups, giving a wider view on the units expressed via rhythmic groups. Syntagm boundaries are more free to be established by the speaker

than the ones of rhythmic groups, and they are marked by optional pauses and changes in speech rate and melody;

- *Breath groups* along with *intonational units*: even larger units of rhythmic organization, more or less coinciding with sentence boundaries. Breath groups are separated from each other by pauses that are used for starting a new breath cycle.

As mentioned above, from the rhythmic point of view, French utterances are very continuous, which is supported by two sandhi phenomena, *liaison* and *enchaînement*, that occur in word sequences that are closely linked by sense and eliminate boundaries between words in favor of merging them into sequences of syllables.

Enchaînement regroups the phonemes in an utterance into syllables in such a way that the last pronounced consonant of one word is attached to the initial vowel of the next word. This phenomenon does not affect the quality of the involved sounds, i.e. normally, there is no assimilation.

Liaison is divided into *vocalic* and *consonantal liaison* (the latter usually simply bearing the name "liaison").

Vocalic liaison happens when two similar vowels coincide in the flow of speech within a syntagm, resulting in one long vowel with a minor change of tone and intensity, and when there are two different vowels occurring one by one, resulting in a very fast transition from one to another with a temporal overlap between the first phase of the second vowel production and the third phase of the first one. The exception to vocalic liaison may be nasal vowels.

Consonantal liaison occurs at word boundaries too. French language has a great discrepancy between its written and spoken forms: most consonant clusters at the end of the word are not pronounced. However, liaison can preserve the speech flow, making the previously mute consonant link the words by means of a new syllable made from the consonant and the initial vowel of the next word. Liaison can be obligatory, optional, or impossible depending on the context and pronunciation style.

As a study by [FDR04] shows, productivity of liaison and enchaînement in French is relatively the same and keeps under the level of 6 occurrences per 100 words.

Prosody is an essential formal feature of a sentence that allows the listener to single it out from the voice stream, divide it into smaller semantic, rhythmic, and melodic segments. Prosodic cues organize the set of words in a sentence into the whole, make the syntactic relations between its parts clear; it is them what is responsible for the expressiveness and most delicate disambiguation in speech.

Prosodic cues are the same in most languages: stress, speech melody, voice pitch, pauses, register, and speech rate.

Stress is used to organize the utterance into segmentable units and to highlight its logical center. Stressed syllables are distinguished from unstressed ones by the voice intensity (*dynamic stress*, controlled by the tension of articulation and amount of the exhaled air), the pitch (*tonic accent*), and duration of the vowel (*quantitative stress*). The tonic accent is dominative in French along with the quantitative stress; voice intensity does not vary much from stressed syllables to unstressed ones.

Depending on the speech segment in question, the notion of stress can apply to words, phrases, syntagms, and utterances as a whole.

Word stress is bound in French: the accent always falls on the end of the word.

Phrasal stress strips most words in the phrase of their stress, leaving *rhythmic*, or *normal stress* (*Fr. "accent d'intensité"*), and *emphatic stress* (*Fr. "accent d'insistance"*). Normal stress defines the voice intensity pattern for the utterance when it is said with a neutral emotion. It is marked by pitch, intensity, and vowel duration and falls on the last syllable of each rhythmic group, being attended by secondary stress that falls, gradually fading, on every other syllable from the end of the rhythmic group. As for emphatic stress, it imparts the emotion behind the words of the speaker (*emotive stress*, usually on emotional words such as *Fr. "misérable", "admirable"...*—making the initial consonant, the first consonant, or the consonant from liaison long) or highlights her line of thought such as in making definitions, corrections, (*didactic*, or *intellective stress*—making the initial vowel long and articulated tenser, often preceded by a glottal stop /ʔ/, or, in case of the initial consonant, doing the same for the first vowel and making the initial consonant be articulated tenser). Emphatic stress does not have to be present in the utterance and cannot replace the regular stress [Fou59, LL70].

Then, syntagms also influence phrasal stress: it is the last rhythmic group that gets accented most, and all groups before it are marked by stress less and less as we move from the end of the syntagm. The greater the speech rate and the longer the syntagms, the less the rhythmic stress instances are pronounced.

Finally, there is a ranking within the utterance, based on the logic of organization of syntagms: the speaker highlights the syntagm that carries the central meaning in the utterance.

Speech melody is the main feature to establish to communicative type of the utterance—declarative, interrogative, imperative, and exclamatory sentences and certain discourse elements such as detached appositions, itemizing, marking an utterance that was cut short, asking for reassurance of the interlocutor, etc. (which is especially important in French, since declarative, interrogative, and imperative sentences can be built with exactly same sequences of words; however, since French is not a tone language, meanings do not depend on voice pitch)—and to identify syntagms and their relation to each other.

Voice timbre brings in the emphatic information on the utterance and depends on the additional tones and overtones inherent in a particular speaker.

Timing control and, in particular, *pauses* add to the utterance segmentation as defined by the stress and melody. The pause serves as a cue on how related the syntagms are and, additionally, is a means of emphasis.

Temporal variations within phonemes and phrases are related to *speech rate*: it can increase, for example, to let the listener identify a subordinate clause, or decrease.

[JF00] modeled prosodic features of the French language in utterances of various communicative types based on four speakers. It is also argued by [Fou01] that the close relation between the segmental and suprasegmental features in speech, articulation and prosody, highlights the necessity to move forwards to their joint analysis.

2.1.3 Adjusting the plan: self-correction

To speak, humans rely on the clues from motor activity of their vocal apparatus and on the gap between the desired acoustic result and the actual one. Motor feedback is processed mostly unconsciously: the receptors of muscles, tendons, and mucous membrane report on their condition to the brain and spinal cord, influencing new neural commands for the muscles of speech,

urging them for compensatory movement. Conversely, the speaker is more conscious of auditory feedback, and it is more difficult to compensate for its loss than for a problem with the motor feedback which can happen in case of a disease, disorder or a surgery. Interfering with the acoustic signal that reaches our ears leads to extreme speech degradation: prosody gets flat, and speech becomes mistimed and inarticulate [Zem10], which can be observed, for example, in children who lost their hearing at a very young age. Even studious training, such as the one for simultaneous interpretation, does not eliminate these effects fully.

2.2 Speech modeling and synthesis

Now when we have covered how speech is produced *in vivo* in general and in the case of French, we need to see the research that has been done to model the anatomy involved as well as the approaches to synthetic speech.

2.2.1 Speech production models

Since speech production is easy to be seen as a multistage process and each of the stages has plenty of room for exploration, the stages have oftentimes been studied independently, producing different models for word representations [LRM99], grammar encoding, phonological encoding and motor planning and control [MTS+10], the first three groups usually falling in the domain of psycholinguistics and the last of motor control. However, recently more and more research has been trying to bring the stages in a unified process [WH18]. Notable examples are [Hic12, Hic14, FD00, WH16], which try to use circumstantial evidence such as aphasia symptoms or specifically targeted experiments to reconcile the disparate fields and unite them within a single framework.

2.2.2 Speech synthesis

Let us see now how the problem of speech synthesis is approached in the state of the art and what were the ideas that brought that about.

The components of present-day text-to-speech synthesis systems can commonly be divided into two big parts.

At the front end, the system analyzes and processes the input text (tokenizes it, parses it and performs its structural analysis, disambiguates homographs, normalizes it according to the intended use of the system)—producing what we call linguistic specification—and converts it into a sequence of phonemes to produce, also marking stress. The result of this step is a set of linguistic features.

At the back end, the speech waveform is generated. However, the presence of two such clearly identifiable blocks is not obligatory (end-to-end speech synthesis).

The speech synthesis research has come a long way from knowledge- and rule-based techniques that exploited phoneme-specific acoustic parameters to two major families of approaches: concatenative speech synthesis that is based on data samples and parametric speech synthesis that works only at the level of one or another kind of parameter evolution, programmatically generating the sound of speech.

Concatenative approaches

Concatenative speech synthesis appeared in the 1980s and the 1990s, seeing the synthesized speech as a concatenation of processed acoustic units taken from a pre-precorded speech database. These units generally are diphones, which are the signal from the second half of one phone to the first half of the other. This way, covering the transition time in a single chunk allows for capturing coarticulatory phenomena: the effect of the phone on its neighbor. As for the cutting point, the phone center, from the acoustic perspective it is the most stable region and therefore is most suitable for concatenating with a segment from another recording. To cover all phoneme pairs, let alone all their positions in the rhythmic phrase and other factors such as intonation patterns and duration, quite a substantial database is necessary. If it is not sufficiently large, the units can be reduced to half-phones. As the database grows, the units can be made larger and larger: phone clusters, syllables and entire words. When the system needs to deal with a narrow domain with little lexical and/or syntactic variation, for example, train station announcements, the units can grow up to segments and phrases (limited domain synthesis).

Diphone synthesis

One popular concatenative approach is diphone synthesis. Within it, the system database contains a single instance of each particular unit available for concatenation (in some implementations, though rarely, a few, to account for different pitch values and speaking rates). Then the waveforms of the instances involved in the phrase to be synthesized are concatenated at points where the sample waveform has the same value to ensure continuity. Then the waveform can be smoothed.

The size of a suitable database depends on the language. While theoretically one would need the number of all uniphones squared, not all combinations actually occur, thus typically requiring between 800 and 2000 diphones (for example, [KEB08] found that Italian used 851 diphones and English 1763, while the number of uniphones was 66 and 39, respectively; in Polish, there are 37 uniphones and 1008 diphones [Bac10]). In comparison to more advanced concatenative techniques, such a database is of a modest size. Nevertheless, preparing it can still be challenging. In the case of French, [DPP+96] indicates it took them one month to record a corpus with a properly monotonous intonation, segment it, equalize the energy levels at the beginning and the end of segments and normalize pitch.

Concatenating and processing such samples does not require a big computational power, so this method can be the solution for mobile devices.

Speech synthesized by this method generally is highly intelligible but not very natural due to its absence of variation in pitch and speech rate. To solve it, as mentioned above, one can record more samples, conditioning diphones on different pitches and speeds; however, this increases the database, thus removing the primary advantage of the low footprint of this method. Another solution would be to manipulate the concatenated samples with time-scaling and pitch shifting, which needs to be done carefully for multiple reasons: first, since it is the articulators what humans move faster or slower, not the vocal folds, the signal needs to be separated into source and resonator frequencies before any speed adjustment; second, voiceless consonants do not have much or any voice contribution and therefore need to be protected from time and pitch shifting to remain natural.

To manipulate the selected samples and ensure a smooth transition between them, one uses speech signal processing algorithms. There are three classes of their representations of speech

[BSH07]: time-domain algorithms such as the pitch-synch overlap-add method (TD-PSOLA [MC90]), linear prediction coding (LPC) and frequency domain-based algorithms.

In TD-PSOLA, probably most commonly implemented method of the three, first we identify pitch periods and mark some identifiable glottal event in each period in the waveform (for example, glottal closure—the point of maximum excitation). Then, during runtime, each segment is windowed with a Hanning window around the pitchmark. Then the extracted windowed signals can be shortened to increase pitch and padded with zero amplitudes to lower pitch; duration can be adjusted by repeating and omitting frames. Then the frames are simply added one after another. As this method does not take care to ensure the homogeneity of the spectral shapes or values, this imposes an important limitation that one can perform only small waveform modifications, using only well-adapted for each other units. Otherwise the result will contain audible glitches.

Unit selection

The major concatenative method is unit selection [HB96]. Unlike diphone synthesis, it operates with a large database, where there are multiple copies of diphones as well as larger units (diphthongs and phoneme clusters with a complex transition), each with its own intonation, speed, intensity, position in the speech segment and the part of speech of the word or words it belongs to. When synthesizing with the units from such a database, every diphone is to be picked among numerous options. To make the choice, we use target specification: before generating the waveform from a selection of units from the database, the system predicts the required prosody (explicitly, with a model, or implicitly, through identifying segments typically exhibiting significant prosodic effects: phrase boundaries, punctuation, etc.), speech rhythm and intensity for the input text. Furthermore, after the input text passed the front end, we also have the phonetic context and its phonological classes, parts of speech, phrase and sentence type and size. This way one can pick a sequence of units matching those supplementary factors as closely as possible.

It must be noted, however, that the concept of a "close match" is not straightforward. There are multiple factors to weigh in:

- Between composing a sequence from several shorter units or a longer one, the longer one will probably end up being more natural. Such a longer unit can equally be introduced into the system as a sequence of consecutive shorter units.

- Furthermore, the results will be more natural if one has a sequence of passably suitable units amounting to the value of target cost function, say, X, rather than a sequence where all units but one are perfect, and all the cost X is concentrated in that single unit, resulting in a noticeable distortion.

- Another point to consider is that even if each of the candidate units is reasonably adapted to the target specification on its own, together they will not produce a naturally sounding result unless the neighboring units are sufficiently similar in terms of acoustics, prosody, voice source and other attributes, meaning that they together have to have a low join cost.

This is why the weighting the target and join components of this decision tree is typically fine-tuned according to the naturalness results.

Then, to mask the transition between segments and establish a common pitch, intensity and speaking rate pattern, speech processing algorithms such as TD-PSOLA are applied as explained in the part on diphone synthesis above.

Overall, unit selection synthesis analysis proves to be highly natural thanks to the direct re-use of recordings of a real speaker. Manipulation with its pitch and speaking rate, however, do not provide a sufficient degree of control over the continuity of the signal, thus rendering the synthetic speech less intelligible. However, this naturalness-intelligibility payoff is still found practical in real applications.

Another point to consider is the size and the computational load of the system: the unit database size and the algorithms of selection and processing, both of which can exceed the available hardware capacity. To improve them, one can build a strategy to prune the database or at least the unit search space and/or replace some components of the system (the front end, the model of prosody, intensity and duration, the algorithm of unit selection) by statistical models. In the latter case, the system is called hybrid.

Employing statistics and machine learning does not have to be restricted to certain sections of a speech synthesis system. In fact, this is another highly productive avenue of speech synthesis research, which brings us to the following section.

Parametric approaches

In parametric speech synthesis, acoustic parameters (spectral envelope, information about the source, usually the fundamental frequency, and noise-like components for obstruents) are modeled as time series, stochastically generated as such from the linguistic specification, and then used in a speech production model—a vocoder—to obtain a waveform of speech. This approach has instigated a very productive area in the field of speech synthesis. Until recently, the state of the art was typically achieved when generating acoustic parameters with hidden Markov models (HMMs), trained by force alignment from an annotated speech corpus [TNT+13]; as the field, like many others, underwent the deep learning revolution, neural speech synthesis has reached the state of the art.

HMM synthesis alone is unable to provide the continuity of the flow of speech: the Markov assumption, incorporated in it, brings in a tendency to centralize speech: the spectral vectors are all emitted close to the average. To solve this and add the necessary dynamic features of speech, the system has to incorporate not only the static coefficients, but also the differences and second-order differences between them. Then, in contrast to automatic speech recognition where triphone models seem to provide already enough context, synthesis requires both phonetic and prosodic factors—a longer-term context or access to the whole phrase or even the sentence. An arising critical problem of the data sparsity is solved by parameter tying and constructing speaker-specific adaptations for generic speech synthesis systems.

2.2.3 Multimodal speech synthesis

There has also been a productive direction of research aiming to add other information to speech synthesis output. This can deal with how the generated speech will be delivered, for example, talking heads meaning to animate an avatar so that it produces a natural output—e.g. [BLDO]; or, aside from the audiovisual aspect, focusing also on the expressivity such as [DCGO19].

Or, alternatively, it may concern with going deeper into how speech is produced: how it is premeditated (brain-machine interfaces and neural speech [GBW+09, GB11, RMC19]) and/or how executed ([BHG+16] for the link with articulation).

If we concentrate on articulation, we enter the domain of articulatory speech synthesis: a method of synthesizing speech by managing the vocal tract shape on the level of the speech organs. The vocal tract can be modeled with geometric [Öhm66, BJK06a, Sto13], biomechanical [LSF12, AHM+15] and statistical [Mae90a, HM11] models. The advantage of statistical models is that they use few parameters, speeding up the computation time. Their disadvantage is that they follow the data a priori without any guidance and do not have access to the knowledge of what is realistic or physically possible. Because of this, to produce correct configurations, they need to be finely tuned.

One crucial notion to synthesizing the movement of the vocal tract is the smallest unit or some other elementary component of articulated speech. From the auditory perspective, it is the phoneme: the smallest semantically distinguishing unit. Meanwhile, from the articulatory point of view, each and every sound we utter is a result of the source, formed by the vocal folds, and the filter, formed by the vocal tract. The shape of the vocal tract is a compound effect of coordinated articulatory movements. Given that each articulator seemingly follows its own timing and has considerable degrees of freedom in space, it is easy to see how the question of organization of speech movements is no trivial task.

A major branch of modeling these motions consists in decomposing speech into articulatory gestures (articulatory phonology [BG92b], task dynamics [SGBR88, NMHJ+12], motor primitives [MIGG99, RVSN16]). However, when dealing with overlapping gestures, we face the disadvantage that, at least as of now, the ground truth is not available; every muscle contraction and every bend at a joint can be integrated into various gesture elements, and it is up to the model only to stay consistent about determining where the boundaries of those gestures are.

From this standpoint, it is interesting to consider an alternative view on what guides speech phenomena: targets. They can be found in a virtual task space (task dynamics [SK87, SM89]; [SMB18]), or they can literally be specific positions of the vocal tract [LQSN17]. It is this line of thought that the present work follows, chosen due to the benefits of the methodological clarity and high applicability, e.g. in articulatory speech synthesis [Lin91, TSS+16, TEL17], which in turn will be capable of serving as further evidence for the underpinnings of speech production [Per17, Eng00, RTP+18], for example, for purely theoretical questions of timing and articulation [PP15, PP14, MPH+17] to more psycholinguistically oriented studies such as [HB19].

[LQSN17] employed targets as the vocal tract configurations attained at the middle of the duration of each phoneme, which is when it is most stable. However, that study recognized the need to allow for contextually modified targets to capture coarticulation. [Bir13a] worked in that direction and differentiated targets according to their vocalic context, following the previous ideas of [Öhm67].

The following subsections are going to be dedicated to the specifics of what goes into treating articulation in speech synthesis.

Articulatory data

Speech is such a dynamic process involving so many different structures in the human body that there is a great need of dynamic and precise data capturing techniques, preferably without

any harm to the subject and tampering with the process of natural speech production.

Aerodynamic measurements were one of the earliest methods; their aim was to analyse the pressure of the air flow when the subject is speaking.

Electromyography is a muscle activity recording technique. It can collect responses from a range of speech organs except for those that are inaccessible, usually by means of hooked-wire or surface electrodes. Hooked-wire ones are inserted into the body of the muscle, causing a minor discomfort for the subject and possibly affecting the way they speak. Surface electrodes are non-invasive and easier to apply [Har99].

Photography can be used to capture articulatory data for visible or partially visible articulators.

Radiography: *X-rays*, *X-ray microbeam*, *cineradiography*, *computed tomography* (CT). All of them can use X-ray to capture the configuration of the vocal tract. The soft tissues appear grey, and even if CT manages to capture them more clearly, the edges of the tongue shapes are not sharp enough. However, the radiation exposure makes these methods unsafe for the subject [BH07].

Magnetic resonance imaging (MRI): uses a magnetic field and radio waves to image a section of tissue. The three-dimensional space is compressed into two dimensions, which may become a source of error for small objects that will be treated as if they were in the same plane. For instance, the epiglottis may be condensed into one single slice, resulting in blurry edges and misinterpretation of its size and shape. Furthermore, MRI is not able to capture solid tissues, and such articulators as the teeth are invisible. Just like in computed tomography, the subject usually has to assume the supine position, which affects the configuration of the articulators and the dynamics of speech—though vertical MRI machines are available, too. MRI can be used to capture dynamic speech—the usual approach is to repeat the same utterance over and over again and then join the scattered images taken at different times into a whole utterance, or use Fast Spin Echo for images of a poorer quality, but better imaging rate (4–24 captures per minute). In comparison to CT, there is no ionizing radiation, but nevertheless long-term biological and clinical safety for the subject's health remains to be proven [KSKM13]. With the advance of technology, real-time MRI (RT-MRI) is being adopted more and more ([THM+06, ATB+09, UZV+10, NBG+11, NZK+13, ELVO16, TN16, SST+17, RTP+18, LZL+19]).

One major question raised with the use of any kind of MRI is its segmentation, i.e. determining the contours of the articulators. This has been addressed with geometrical, statistical and deep learning models ([NTR+14, TN16, TSS+16, SST+17, STTN17]).

Palatography: an early technique to study tongue placement across the palate and teeth; it has developed into *electropalatography* and is now able to make captures in conversational speech rather than in short sequences. It is a simple technique: easy to operate and relatively non-invasive [GN99]. It was continued to improve over the years with variation in materials, the number of sensors and chips [PB15, SB16].

A range of *point tracking techniques* is available too. Their advantage is a fast sampling rate and selectivity: it is possible to track exactly that point in the tissue that is of interest. However, it is not possible to apply enough trackers to obtain the whole picture without hindering the speech, while imaging techniques can provide true multi-dimensional data.

To summarise, there still is no completely safe and informative method for collecting dynamic articulatory data. Either the method captures a very particular behavior of the subject, or the data are comprehensive and of high quality but come only in small amounts and at much lower

frequencies than the frequency of 100 Hz that is deemed to be necessary to capture speech phenomena.

Vocal tract modeling and synthesizing speech

So, speech synthesis augmented with articulatory information tries to stick to the natural speech mechanism as close as possible. Nevertheless, there is variability in how close that is. Biomechanical approach solves the Navier-Stokes equations to estimate deformation of all (prominent) vocal tract muscles and organs and estimate the corresponding aero-acoustic phenomena. It has a benefit of full control over the articulators, but suffers from a heavy computational load and lack of required anatomical and physiological data [LSF12, AHM+15].

Meanwhile, articulatory speech synthesis is based on using a simplified articulatory model to imitate human speech, controlling the articulators in a feasible way. It is able to replicate the anatomical and physiological phenomena without delving deep into the structures that are involved—it concerns only temporal evolution of the vocal tract geometry, which turns out to be enough to synthesize speech of ample quality.

To alleviate the computational load and make the known theoretical knowledge applicable in this case, the direct numerical simulation of speech phenomena is still limited to individual cases that often are overly simplified, and even in such cases the complexity is so high that the computations take dozens of hours [Mae90a]; it can take days to compute the flow dynamics at the vocal folds [JHdA+13]. So, another approach is to develop models only for the most essential physical phenomena.

[PVC+96] and [EPZ+11] provide theoretical foundations that allow us to define what phenomena are most important in phonation; they are supported by experiments of [RPVH+07] and [SSDW+01]. The constructed larynx models are locally incompressible (with a low Helmholtz number), quasi-stationary (with a low Strouhal number) and with an insignificant viscosity (with an intermediate Reynolds number). It is not necessary to specify all details about the flow motion to produce a turbulence-induced sound [HM05, Kra05, YNW19].

The vocal tract configuration, from the glottis to the lips opening, can be encoded by means of area functions that approximate the vocal tract by acoustic tubes of varying size. The area function can be estimated without any knowledge of separate articulators [Fan71b]. [Sto13] generates area functions as a transition from one vocalic area function to another, with consonants superimposed as constrictions attained during this vowel-to-vowel transition. This approach, roughly based on Öhman's [Öhm66] vowel substrates and consonantal perturbation, can be used to synthesize some fixed phrases but is difficult to generalize for use in text-to-speech (TTS) systems. [SMB18] approximates area functions as a one-dimensional model of the vocal tract based on just six points.

Another approach would be to construct the sagittal section of the vocal tract (the two-dimensional case, such as in the work by [Mer73]) or the vocal tract proper (the three-dimensional case, such as in the work by [BJ03]) as the area bounded by a bunch of primitives. Both of the mentioned studies do not fit the form of the vocal tract too precisely. Instead of geometrical primitives, the vocal tract can be shaped out from medical images by means of an articulatory model. Many articulatory models have appeared in research, ranging from as simple as three-parameter ones [Fan60] to as sophisticated as hundreds-parameter ones [GWTPP06]. One of the most famous models is the one by [Mae90b], comprising a model for the lips, a model for

the tongue, and a model for the larynx. All three articulators' configurations are decorrelated from the influence of the lower jaw position and encoded by means of the principal component analysis into vectors of varying lengths. To keep track of the encoding, one stores the vertical opening and protrusion of the jaw. [Mae90b] indicates that three tongue parameters can describe 96% of the variance in the test images, and all three articulators are encoded in seven parameters in total (see Figure 2.9).

First applied to vowels only, this model was followed by models to account for consonants too—in two dimensions [LB11a, LVC14, MZF19] as well as in three dimensions [BBR+02]. (Obviously, the two-dimensional models have to be accompanied by some kind of spatial estimation at a further stage in speech synthesis; since, again, full-scale 3D calculations take excessively long computation time—often hours to calculate a 10-ms-long speech signal—the more feasible alternative appears in limiting the number of dimensions.)

As for another articulator that was left untreated by Maeda, the velum, there are not many models for it. [SB08] made a three-dimensional model of the velum based on static 3D MRI and CT images, which raises the question of being actually able to capture the high variability in the dynamic process of speech. [LET15] propose a PCA-based velum model built on an X-ray film of 15 short French sentences; this model is able to capture 70% of the total variance by means of two components.

When the shape of the vocal tract is established, the vocal tract has to be divided into narrow tubes that are "strung" on the vocal tract's central line [ML13] which can be computed by various algorithms. The areas A (in cm^2) can be computed from the heights (d, in cm) of the estimated tubes and two speaker-specific parameters that are to be determined empirically:

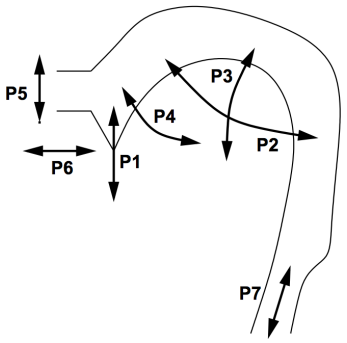


Figure 2.9: The seven parameters of Maeda's articulatory model: *P7* is the *larynx* height; *P2*, *P3*, and *P4* encode the *tongue*—its body, shape, and tip respectively; *P1* stores the *jaw* opening; *P6* and *P5* define the position of the *lips*—their protrusion and opening respectively.

$$A = \alpha d^{\beta} \tag{2.3}$$

To control the dynamics of the modelled vocal tract, one can use the notions of *gestures* (motions of formation of a particular constriction over time) and *targets* (the vocal tract configurations that the speaker aims to reach). Examples of studies that aim for simulating the underlying mechanisms of speech production involve the task dynamic model by [SM89] (accompanied by the work on how to operate it: [NMT+12]), the gesture-based dynamic model by [BG92a] and gestural dominance model by [BJK06b], and the targets-based model by [Bir07]. Furthermore, the dynamic characteristics of the system can be encoded together with the parametric representation of the vocal tract, such as in work of [WHS19].

3

Articulatory speech synthesis from static MRI data

3.1 Introduction

One of the principal notions that one has to face when dealing with speech synthesis and processing is coarticulation: the influence of neighboring sounds on each other, stemming from the fact that we cannot simply append one shape of the vocal tract after the other and always have to try to find the most efficient articulatory trajectory that would still leave our message comprehensible.

For example, when we pronounce both /ba/ and /by/, what is definitive is that the lips need to close to produce the stop /b/. However, there is no restriction on *where* they need to do that. So, a natural transition is to give the lips the shape that is required for the coming vowel: to protrude them for /y/ and not for /a/. This way, the moment the burst of /b/ is released, production of the vowel starts.

This influence is most pronounced in the direction of anticipation: when the sounds planned in the near future acquire some features of those that currently are in production. The other direction of coarticulation, carryover, where the past sounds have lingering effects on the present ones, is less prominent and is mostly attributed to passive inertia [KN99].

The greatest impact is typically brought by vowels, which become syllable nuclei. This is especially true in the case of French, where vowels need to be articulated quite precisely and, unless it is a /ə/, cannot be reduced, unlike some other languages like Russian.

This largely outlined the approach of this study: to produce speech and articulatory movements from phonetic transcription, relying on a set of basic vocal tract configurations and focusing on the way how vowels influence consonants before them.

We were interested in exploring the potential in using quite little, and yet sufficient, static magnetic resonance imaging (MRI) data and implementing one of the few existing attempts at creating a full-fledged articulatory speech synthesizer with a comprehensive control over the articulators, that would be capable of reproducing the vast diversity of speech sounds. This lead us to dealing with these building-block vocal tract configurations by following the steps of [Bir13a], applying that methodology to French and extending the scope to cover practically the entire set of French phonology.

3.2 Objectives

The objectives of this study were as follows:

- To represent the available MRI data with an existing articulatory model (joint work with Yves Laprie);
- To extrapolate this limited dataset so as to estimate the missing samples, generally following the approach of [Bir13a] (carried out by myself);
- To develop a set of rules defining how, given an utterance to produce, to pick necessary samples from the extended library of basic articulatory configurations, how to adjust them to set up an efficient transition both at the level of each of the articulators and the entirety of the vocal tract, and how to transition between them in a set time while also managing the parameters of the source and pressure in the subglottal and supraglottal cavities (carried out by myself—the main contribution of this part);
- To post-process the obtained vocal tract configurations to make sure that they follow the phonetic rules relevant to the French language (carried out by myself);
- To process those configurations with an existing acoustic simulation unit (joint work with Benjamin Elie and Yves Laprie);
- To analyze and evaluate the resulting utterances (joint work with Yves Laprie and Benjamin Elie).

3.3 Building an articulatory speech synthesis system

As set up by the objectives, the system is made up of three major components: the database with the "building blocks" for articulating utterances, the joint control algorithm for the vocal tract and the glottal source, and acoustic simulation. The primary concern of this work are the first two components.

3.3.1 Dataset

The data were a static-MRI subset—the subset that was available at the time of the study—of the ArtSpeechMRIfr dataset [DFF+19] recorded at Nancy Central Regional University Hospital, France, under the approved medical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053). The scanner in use was General Electric Signa HDxt 3T (GE healthcare, Chicago, Illinois, United States). We used 3D FGRE (TR $= 3.12$ ms, TE $= 1.084$, FOV $= 26 \times 26$ cm, flip angle $= 10$ degrees) 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. The subject (subject A, subsequently referred to as S_A) is male, 35 years old, 182 cm tall and 74 kg. Overall it made for 97 images.

Examples of mid-sagittal cuts of these data are shown in figures 3.1 and 3.2.

Figure 3.1: Mid-sagittal slice of the static 3D images of subject S_A for several of the French vowels.

Figure 3.2: Mid-sagittal slice of the static 3D images of subject S_A for some of the French consonants. Consonant was pronounced in context of the following vowel.

I selected the mid-sagittal slice in those images. These data captured articulation without phonation: the speaker was instructed to show the position that he would have to attain to produce a particular sound. For vowels, that is the position when the vowel would be at its clearest if the subject were phonating. For consonant-vowel (CV) syllables, that is the blocked configuration of the vocal tract, as if the subject were about to start pronouncing it. The assumption is that such articulation shows the anticipatory coarticulation effects of the vowel V on the consonant C preceding it. There were 13 vowels, 72 CV syllables and 2 semi-vowels in the final dataset. This covers all main phonemes of the French language, but not in all contexts. Each consonant was recorded in the context of the three cardinal vowels and /y/, which is strongly protruded in French. Some intermediate vocalic contexts were added so as to enable the vowel context expansion algorithm to be checked.

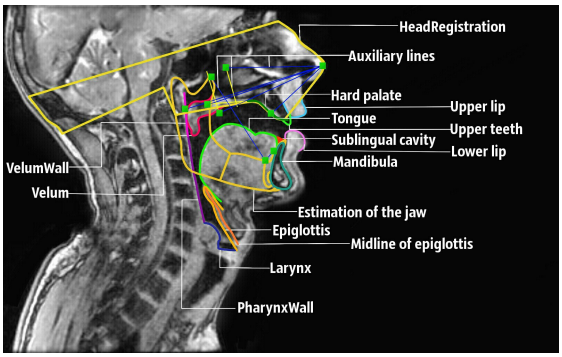


Figure 3.3: An example of dataset image annotation (/a/).

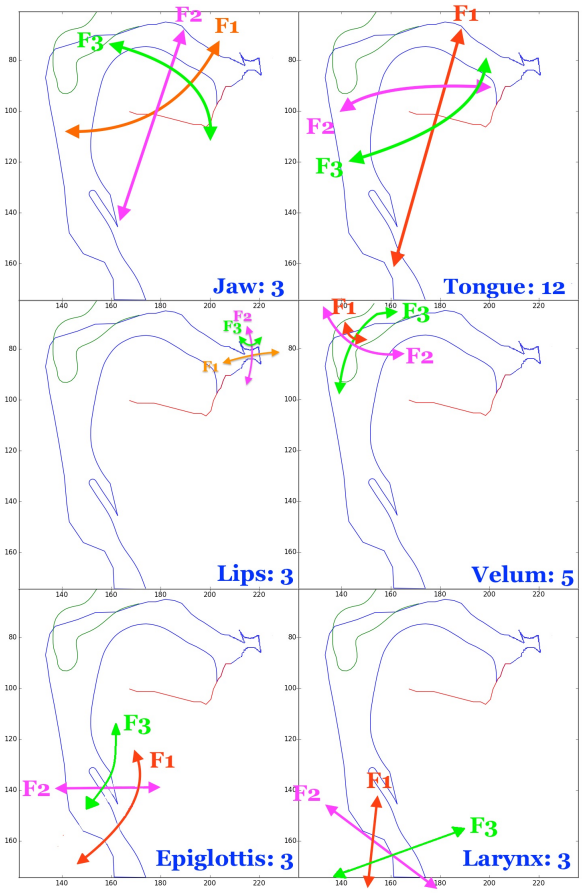


Figure 3.4: The PCA-based articulatory model: curve change directions encoded in the first three factors of each articulator (the jaw, the tongue, the lips, the epiglottis, the larynx).

Expanding the dataset

Since the collected French phonemic dataset was limited, I needed to expand it to cover other contexts as well. We used the notion of the cardinal vowels—/a/, /i/, /u/ and /y/,—assuming that /a/, /i/, /u/ and /y/ represent the most extreme places of vowel articulation, and since then any other vowel articulation can be expanded as a combination of its /a/, /i/ /u/ and /y/ "components". Having captured the C+/a/, C+/i/, C+/u/ and C+/u/context for all consonants C and all non-cardinal vowels V on their own, I was able to estimate the missing C+V samples:

- I projected the vowel V articulatory vector (from $\mathbb{R}^{29}$) onto the convex hull of the /a/, /i/, /u/ and /y/ vectors.
- Assuming that the linear relationship between the C+V vector and the C+/a/, C+/i/, C+/u/ and C+/y/ vectors is the same as the one between V and /a/, /i/, /u/ and /y/, I estimated C+V from C+/a/, C+/i/, C+/u/ and C+/y/ using the coefficients from the projection of V onto the convex hull of /a/, /i/, /u/ and /y/.

I also estimated the neutral C configuration, the one without any anticipatory effects, as the average of C+/a/, C+/i/, C+/u/ and C+/y/.

Finally, I assumed that the voiced and unvoiced consonants did not have any differences in the articulation.

Articulatory model

After manually annotating the captures as shown in Figure 3.3 we applied a principal-component-analysis (PCA)-based model on the articulator contours [LB11b, LVC14, LET15]. We paid special attention to the interaction between articulators and the relevance of deformation modes. Moreover, articulators other than the jaw, tongue and lips are often neglected and modeled with insufficient precision, whereas they can strongly influence acoustics at certain points in the vocal tract. Here are two examples. The position of the epiglottis, which is essentially a cartilage, is likely to modify the geometry of the lower part of the vocal tract by adding an artificial constriction disturbing all the acoustics. It is therefore important to model its deformation modes and interactions with other articulators correctly. In the same way, the velum plays an important role both in controlling the opening of the velopharyngeal port, and in slightly modifying the oral cavity to obtain resonant cavities that give the expected formants of vowels. The acoustic tests we have carried out show in particular that the velum makes it possible to better control the balance between the two cavities necessary for the realization of /u/ and /i/.

Regarding the tongue, PCA was applied on the contours delineated from images. Deformation modes are likely to be impacted by delineation errors. In the case of the tongue, these errors are marginal, or at least give rise to deformation modes coming after the genuine deformations whose amplitude is bigger. On the other hand, the width of epiglottis and/or velum is small on the images, and the errors of delineation, whether manual or automatic, are of the same order of magnitude as genuine deformations. Consequently, PCA applied without precaution will mix both types of deformation. To prevent the apparition of these spurious deformation components the epiglottis was approximated as a thick curve, and only the centerline of epiglottis was analyzed. As a matter of fact, the centerline was determined after delineation of all the epiglottis contours, and the width was set as the average width of all these contours in the upper part where the two epiglottis edges are clearly visible (see Figure 3.5). The height of the upper part (where both contours are visible) is adjusted by hand to fit the contours extracted from images. The centerline is approximated as a B-spline and represented by its control points P_l ($0 \leq l < M$ where M is the number of control points) in the form of a two-coordinate vector, and the reconstruction of the epiglottis from the centerline amounts to draw a line at a distance of half the width from the centerline.

The influence of delineation errors is very similar for the velum, which is a fairly fine structure not always well marked on MRI images because it moves quickly. As for epiglottis we used the centerline and a fairly simple reconstruction algorithm. However, PCA was not applied directly on the control point of the splines because the velum can roll up on itself. This particularity does not lend itself well to the direct use of PCA, which results in the emergence of linear components not appropriate in this case. The centerline is therefore broken down into a series of segments of the same length. Each segment articulates with its predecessor and the first point is fixed. The centerline is then defined as the vector of angles between two consecutive segments. In this way PCA can be applied effectively to velum and gives rise to relevant deformation modes.

The architecture and general organization of the articulatory model are based on the dependency links between the articulators. The main articulator is the jaw which is represented by

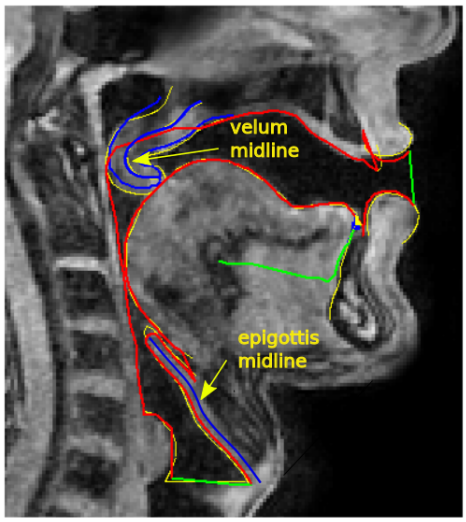


Figure 3.5: Epiglottis and velum centerlines reconstructed by the model (solid blue lines). Reconstructed vocal tract is represented by solid red lines. The vocal tract input and output are represented in solid green lines. All these contours are superimposed onto the contours (represented as solid yellow lines) delineated from the image.

3 parameters to get a complete and accurate control. Its geometrical contribution is subtracted from the tongue contours before the application of PCA because tongue is directly attached to the mandible. The tongue is represented by 12 parameters in order to obtain a sufficient precision for the realisation of consonant constrictions. The lips are represented by 3 parameters. Unlike the tongue, the interactions between lips and jaw are more complex. For this reason we subtract the correlation between jaw and lips before applying PCA. The larynx is considered to be independent of the jaw and is represented by 3 parameters to control ist orientation and vertical position. In the same way the velum is considered as an articulator independent of the others. It is analyzed as explained above and is represented by 5 parameters. The epiglottis is the articulator that is subject to the greatest number of influences: the jaw via the tongue, the tongue itself and the larynx. These influences are subtracted by applying a multiple regression to the epiglottis centreline before applying CPA. Analysis of the variance shows that the various influences on the epiglottis account for most of its deformations. Its intrinsic deformation are represented by 3 parameters.

In total these parameters form a vector from $\mathbb{R}^{29}$ (see Figure 3.4 for major parameter contributions to the articulator shape). Since the model uses PCA, the zero configuration should correspond to the central position as identified in the dataset, and small changes in the parameter space within a certain neighborhood of zero should correspond to small changes (in terms of distance and shape, not in terms of the resulting acoustics) in the curves. A clipping algorithm is used to solve problems of collision between articulators, i. e. essentially between the tongue and palate. So the model's behavior is not entirely linear.

3.3.2 Strategies for transitioning between the articulatory targets

The dataset provided static images capturing idealistic, possibly over-articulated, targets for consonants anticipating particular vowels, whereas the goal was to be also able to deal with consonant clusters and consonants that would not anticipate any vowel at all—for example, due to their ultimate position in a rhythmic phrase. So, in our context, to establish a transitioning strategy would mean three things:

- Choose the building blocks: identify the articulatory target for each phoneme in a phrase. It can either be what was captured in the dataset (a vowel or a consonant assuming vocalic anticipation), an estimation of what the dataset was missing (missing phonemes, such as voiced consonants, missing contexts or the absence of any context), or only a subset of articulatory parameters corresponding to the critical articulators for the particular target. A consonant cannot anticipate multiple phonemes, nor can vowels anticipate anything due to the restrictions of the dataset at my disposal.
- Decide when — and whether — the articulatory target should be attained.
- Decide how to generate the articulatory positions between the target ones.

Our basic assumption was that by default, consonants anticipate the next coming vowel. However, it would be unrealistic to assume it happens in all cases. This is why I imposed several restrictions on the anticipatory effect:

- Temporal: no coarticulatory effect if the anticipated phoneme is more than 200 ms ahead;
- Spatial: if there is any movement scheduled between the anticipated vowel, the phoneme in question negates the effect. For example, consider such sequence as /lki/: after /l/, the tongue needs to move backward to produce /k/ before coming back forward for /i/. In this situation, my algorithm does not allow the /l/ to anticipate the coming /i/. Algorithmically, it is done with associating every place of articulation to a number and checking the difference between those numbers;
- Categorical: it is not possible to anticipate a vowel more than 5 phonemes ahead, and this restriction becomes stricter if it applies across syllable boundaries.

For vowels, there is also a model of target undershoot.

Having established the articulatory targets, the question is how to transition between them. We have tested out three strategies for interpolation between the target vectors:

- Linear: the interpolation between the target vectors is linear, with sharp turns at the knots;
- Cosine: smooth transitions;
- Piecewise 1-d monotonic cubic Hermite interpolation that has smooth transitions, the magnitude of each transition section bounded by its corresponding interpolation knots;

- Complex: transitions are done with the previous cubic Hermite interpolation, but the timing varies by the articulators. The critical ones reach for their target position faster than the others, while those articulators whose contribution to the resulting sound intelligibility is not as large move slower (for example, the tongue can be in a number of positions for the sound /b/, but the lips have to come into contact). Furthermore, the articulators composed of heavier tissues (such as the tongue back) move slower than the light and highly mobile ones (such as the lips).

Which of the strategies is more realistic from the articulatory standpoint, is a question that can only be answered with the help of dynamic data. In their absence, I had to restrict myself to the analysis of the resulting acoustics. The choice of the strategy did not, however, seem to make a difference in the quality of the audio output.

3.3.3 Obtaining the sound

Each vocal tract position was encoded in an area function. They were obtained by the algorithm of [HS65] with coefficients adapted by Shinji Maeda and Yves Laprie. These parameters only depend on the position in the vocal tract between the glottis and the lips. The transition from the sagittal view to the area function has given rise to several works which contradict each other slightly ([SLMD02] and [MJB12]) and it is therefore clear that the determination of the area function will have to take into account the dynamic position of the articulators in the future.

Then, having obtained the translation into area functions, the constrictions were corrected with the knowledge of the phoneme in production: a stop, a fricative or a vowel. This way I was able to ensure that all stops attained closure at the place of their articulation, all fricatives did not close too much or open too wide, and all vowels had enough space for the air to pass. From the development perspective, each place of articulation was associated to a section in the 40-tube area function representation: the lips at $[37, 39]$, the teeth at $[36, 38]$, the alveolar ridge at $[34, 36]$, the palate at $[32, 35]$, the velum at $[17, 33]$, and the uvula at $[21, 29]$. Then, whenever the place of articulation needed to have a constriction at that time or, on the contrary, could not have too close a constriction according to the timing rules, the area functions were corrected: closed vowels were not allowed to have a constriction of less than 0.25 cm^2, mid-close less than 0.3 cm^2, mid-open less than 0.35 cm^2, open less than 0.4 cm^2; oral and nasal stops were enforced to have a complete closure, 0, at their place of articulation; fricatives were not allowed to have a closure of less than 0.1 cm^2. Velopharyngeal opening was verified and corrected as well: if it was an oral sound, velopharyngeal opening was corrected to 0; if it was a nasal sound with an opening less than 0.5 cm^2, the opening was reset to that minimal value.

Then we used an acoustic simulation system implemented by [EL16a] to obtain sound from the area functions and supplementary control files: glottal opening and pitch control.

Glottal opening was modeled by using external lighting and sensing photo-glottography (ePGG) measurements [HM08]. Within the model I implemented, glottal opening is a relative value from 0 to 1, 0 corresponding to most closed (as in vowels) and 1 corresponding to as open as possible. The key value to attain during the production is 0.0 for a vowel, 1.0 for a voiceless fricative or stop, and 0.7 for a voiced fricative or stop. The other phoneme classes do not have key target values. Then, empirically, for every scenario (the opening value increasing from one peak, such as 0.7, to another, or decreasing from one peak to another, or moving toward or from

0), I added auxiliary points to match the shape and temporal behavior of the few examples in the data that were at my disposal.

There was no need to model voicing (high-frequency oscillations of low amplitude superimposed onto the glottal opening waves) since the vocal folds operated by the glottal chink model [EL16a, EL16b] are self-oscillating.

3.4 Evaluation

Each step in the system was evaluated on its own, and afterwards the synthesis results were evaluated visually, acoustically and perceptually. Since the objective of the work was rather to have a fully functional algorithm that produces reasonably realistic movements and sounds rather than to obtain high-quality speech, a more rigorous evaluation, such as a quantitative comparison to the dynamic data on articulatory trajectories, is still an avenue of future work.

3.4.1 The articulatory model and the trajectories

One peculiarity of the dataset and therefore of the model was the fact that it used only the sagittal section of the speaker's vocal tract. While full three-dimensional models can capture the full geometry of the vocal tract with such phenomena as lateral phonemes (e.g. /l/), two-dimensional models get the benefit of faster computation time and overall simplicity, but irreversibly lose the spatial information.

In general, the articulatory model captured vocal tract positions correctly or with no critical errors, though some part of its success is definitely owed to the post-processing stage where area functions are corrected, since on its own the model did not impose much control over constrictions. This way, control became two-fold: the articulatory model operated at the level of articulators, and the post-processing set of rules on the resulting vocal tract geometry.

As for the movements, we can say that they were reasonable and the coarticulation-affected targets guided the articulators to the positions necessary to produce a particular utterance. One key point here is the timing strategy. Rule-based timing strategy seems to be very rigid for the dynamic nature of speech; it would be more natural to follow speech production processes in humans and to guide the synthesis with the elicited sound or the speaker's expectation—based on their experience—on what this sound will be.

3.4.2 Glottal opening control

The algorithm for the glottis opening successfully allowed to distinguish between vowels and consonants. Distinguishing between voiced and voiceless consonants, though, stays a point for improvement, as well as well-coordinated control over the glottis and the vocal tract to avoid acoustic artifacts.

3.4.3 The synthesized sound

Vowels and stops were the most identifiable and correct, although sometimes some minor adjustments in the original data were necessary to obtain formants close to the reference values. When compared to human speech, the formant transitions within the suggested strategies sometimes

occurred too fast and sometimes too slowly; again, this highlights the utmost importance of realistic timing strategies. Figure 3.6 shows an example of the synthesis when it is guided by real timing: /aʃa/ as produced by the system and as uttered by a human. The high-frequency contributions in /ʃ/, not appearing in the human sample, are due to the acoustic simulation. The noise of /ʃ/ is at the correct frequencies, but with a bit different energy distribution, probably because of differences in articulation or in the area functions. There is also an acoustic artifact between /ʃ/ and /a/, which means that more work is necessary on liaising the vocal tract and the source control.

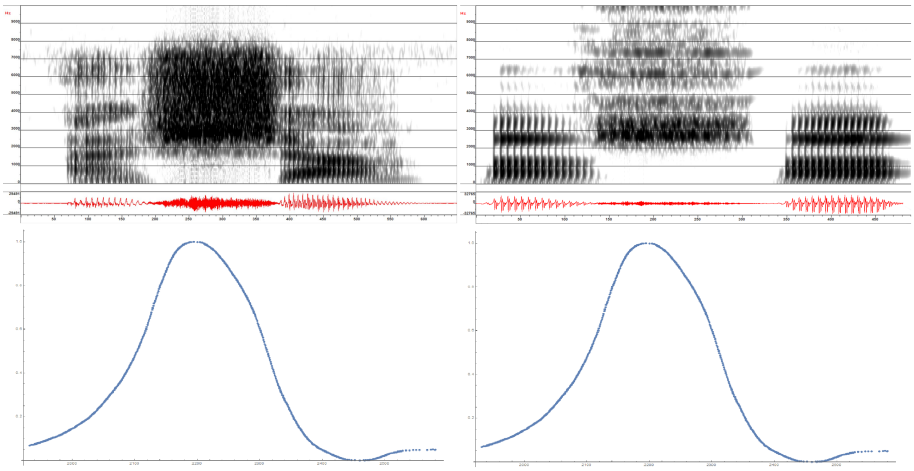


Figure 3.6: An example of a human's utterance of /aʃa/ (left) and its synthesis (right) along with the glottal closure control as copied from the EPGG data (below). Phoneme durations are aligned.

3.5 Conclusion

3.5.1 Overview of results

Regarding speech as a process of transitioning between context-aware targets is an interesting approach that can be connected with the mental processes of speech production: to allow the others to perceive the necessary acoustic cue, the speaker needs to come close enough to the associated articulatory goal. The important difference between a real speaker and the algorithm is the fact that the algorithm solves a static problem, laid out in full; it needs to hit particular targets in a given order. As for humans, we solve a dynamic problem, and coarticulation is not something we put in its definition; rather, coarticulation is our means to make the problem of reaching too many targets in a too short period of time solvable.

The statistically derived articulatory model encodes complex shapes of the articulators in only 29 parameters, sometimes struggling at the constrictions because of the inherent—and intentional—lack of control over the resulting geometry of the vocal tract. This drawback is addressed with a post-processing stage, which is a compromise (though brutal) solution.

The shapes of the articulators change in time according to the produced trajectories of the vocal tract, and those are phonetically sound. They evolve in synchrony with other speech

production parameters: F0, subglottal and supraglottal pressure. Together, the timing and values of the system need to be in a delicate balance so as not to produce any artifacts down the synthesis pipeline. We find that they are sufficiently well tuned for vowels and stops; as for fricatives, the interplay between the place of articulation, pressure control and the temporal evolution is so intricate that it essentially boils down to modeling each fricative separately.

I conclude that, within the laid out setting—comprehensive but static and little in quantity data of the vocal tract, a statistical model to encode it, and a set of rules to manipulate it—this work represents a thorough exploration of the approach. The results show that such data and methods are not unsuitable for building an articulatory speech synthesizer; however, given the limitations of the approach, a more promising course of carrying on with this work would be incorporating patterns from some actually recorded dynamic data rather than a continued search for a better set of rules through theoretical modeling, trial and error. More on this in the following section.

3.5.2 Future work

As was said above, for the approach to work well, a closer, intertwined and well-aligned interaction between all its components—vocal tract configuration, voicing and pressure control—is necessary. Considering the uncountable amount of timing strategies and possible transitions, it is reasonable to learn them from dynamic data; for example, approaches by [ELVO16] could be used for that. Employing annotated dynamic data such as EMA or RT-MRI and augmenting the articulatory model used in this study with temporal control, aligned with voicing control as extracted from simultaneous audio, should considerably improve the results. Additionally, exploring a connection between dynamic data and the present system would allow to properly evaluate the generated trajectories and the sound.

Another approach to boosting the performance of the system would be to automate the search for better rules: set a cost function for how good a generated articulation sample and its audio are, have several initial sets of synthesis parameters, generate the output with them, evaluate the result, adjust the parameters, re-generate the output, make conclusions about the new adjustment of the parameters, etc., until the improvement disappears or is negligible—comparable to how this relating to the result was treated in [WSB17]. This, like taking patterns from actually observed dynamic data, is also a very productive idea, but difficult to implement in a setting with multiple disjoint components rather than a single environment as in VocalTractLab [Bir13b].

4

Articulatory speech synthesis from real-time MRI data

The previous chapter explored articulatory speech synthesis where the aspect of articulation was treated as an extrapolation of the available static data with no reference of its dynamics. The obvious inherent pitfall of that approach was the temporal aspect: managing the duration, the speed and the trajectory of articulatory gestures.

To address that, this chapter is dedicated to articulatory speech synthesis with all its components derived from real-time data.

4.1 Objectives

The objective was to address both articulatory and acoustic aspects of speech by taking into account all the effects actually observed in use by a human speaker during normal, natural and unrestrained speech; more specifically, to build a joint model to synthesize both articulatory and acoustic parameters that would correspond to a given phonetized textual input.

That included picking appropriate data and methods with a consideration for the difficulties of treating articulation, applying those methods to the data and evaluating and analyzing the results. Specifically, the objective was to prepare RT-MRI data as a rich source of articulatory information that can be collected at a good enough rate without tampering with the natural process of speech and use both its visual and audio components jointly to drive an articulatory speech synthesizer derived from patterns in the real behavior of the speaker rather than a set of developed rules.

4.2 Methods

Once the data and methods were chosen, the pipeline of the work was as follows:

- Prepare the data: build the corpus, carry out the acquisitions, check that the data are valid;
- Preprocess the data: perform phonetic annotations and obtain linguistic specifications, synchronize them with the individual frames;

- Do image segmentation or extract articulatorily relevant parameters from the frames;
- Train and test a joint acoustic-articulatory DNN-based parametric model to obtain an articulatory speech synthesizer;
- Evaluate and interpret the results.

The following sections of this chapter are going to present the details.

4.2.1 Data preparation

Given the long-term objective to synthesize speech based on real-time, actually observed acoustic and information-rich articulatory phenomena, it was necessary to get a good geometric description of the whole vocal tract and to get running speech which would exhibit how the global geometry of the vocal tract evolves over time during speech production; RT-MRI articulatory data seemed to be a good fit for that.

While it is relatively easy to find audio-articulatory databases in English [NBG+11, SST+17], there are no similar free data in French. This prompted off the creation of ArtSpeechMRIfr [DFF+19], the dataset used for this work.

Dataset conceptualization and creation

A dataset necessary for an articulatory speech synthesizer needs to meet several criteria:

- For the synthesizer to be robust, the corpus must have a good coverage of all phonetic contexts which can appear in French. Coverage is of a higher priority than the number of instances per each phonetic combination.
- For the synthesizer to be natural, and for us to be able to evaluate the naturalness, the data needs to be natural.
- For the eventual ramifications of articulatory speech synthesis for speech science, one may add utterances with some phenomena of theoretical interest, such as the articulation of sounds with a complicated physics of production, articulatory adjustment phenomena and intra-speaker variability.
- The eventual results that come from using the dataset need to be in a form that is comparable to previously existing work.

On top of that, the nature of RT-MRI acquisitions imposes a few more restrictions:

- The corpus planned for recording cannot be too long, both for the comfort of the speaker who is subjected to a long period of immobility in a tight space with a loud noise, and for the cost and limited availability to use the machine. The solution to split the corpus into several parts to be recorded over several sessions is suboptimal, as it could lead to a great inter-session variability in terms of the speaker's head position and speaking style;
- The speaking tasks need to be easily understandable by the speaker, since communication between the control station and the person in the machine is noisy and uncomfortable.

Regarding phonetic coverage, let us consider the classical way of constructing a corpus for speech synthesis, which is to collect sentences from a vast written corpus, for instance, of a newspaper. The greater the collection of sentences, the better its coverage. The problem with this approach, however, is that each sentence contributes only a very limited number of new phonetic contexts. A corpus vast enough to cover enough speech phenomena to serve as a basis for a speech synthesizer would have thus required a very long recording, more than a dozen hours of speech, which, as explained above, is not an option for MRI.

A good solution for our case was to specifically prepare a set of sentences for the subject to read out, so that the coverage and balance of phonetic combinations became exactly as necessary. The set of sentences for ArtSpeechMRIfr, generated by Yves Laprie from a phonetized version of French Morphalou lexicon which provides 620.000 flexed forms [JMPSA06, LSG04].

Yves Laprie used several levels of criteria to guide the manual composition of sentences with the available phonetized lexical entries. On inserting a new sentence, the first level is to count its number of double vowels V_1V_2 for all vowels, the number of CV with the consonant C being one of /p, t, k, f, s, ʃ, l, ʁ/ and V of /i, a, u/ or—additionally—/y/, the number of VC where C is in the list /l, ʁ, n, m/ as a coda and V in /i, a, u, y, e, ɛ, o, ɔ/, consonant clusters C_1C_2V with C_1 in /p, t, k, b, d, g, f/, C_2 in /ʁ, l/ and V in /a, i, u, y/, and the number of instances of 15 specific complex consonant clusters (at least a sequence of 3 consonants, between two vowels). Except for those complex clusters, all the contexts in question almost always appear within words to avoid the effect of prosodic boundaries.

Such corpus composition covered the core of mandatory phonetic contexts. Then, new additional contexts were introduced, in particular, contexts with vowels not present in the set of the cardinal vowels /i, a, u/ extended with /y/. The instances of VCVs were broken down by grouping vowels into classes from the open to the close (/i, e/, /ɛ, a/, /u, o, ɔ/, /y, ø/, /œ, ə/ and nasal vowels /ɑ̃, ɔ̃, ɛ̃, œ̃/). This provided a second level of phonetic coverage evaluation.

In total this corpus was made up of 79 sentences offering a very good coverage of all possible phonetic contexts in French. The disadvantage of the sentences being artificially constructed is their highly odd semantics (e.g. “Vous dactylographiez sa soupe sirupeuse au lit”—*Fr.* “You type his syrupy soup in bed”) and, in the case of two words, non-French inclusions (“cartoons”, “squaw”). However, syntactically they are correct, and a native speaker should not have any problem reading them.

Additionally, for comparison purposes, an additional set of syllables and sentences coming from previously existing corpora (the frequently used in phonetics "La bise et le soleil..." sequence [Int99] and a randomly selected subpart of the corpus in [MCTO11]) was added, as well as a sequence of syllables mirroring and extending our static corpus (Chapter 3.3.1). Also, to study the complex articulation of trills (that are not present in standard French), experimental sequences of /aβa, uβu, iβi/ and /ara, uru, iri/ were added.

As for speech naturalness, one could see how reading out artificially constructed sentences that would never be voiced in any real context is probably not the best representative example of natural speech data. For this reason, we wanted to record some spontaneous speech, which is less controlled than reading out loud and also manifests more effects like overlapping gestures and phoneme elision.

I created prompts for each subject to follow for a minute per prompt, in a random order. They covered everyday topics: “What do you like in your work?”, “Speak about your last trip anywhere”, “Speak about a film or a book that had a lasting impression on you” (20 topics in

total, see Table A.1). In the end, at the recording time, despite having hesitations in speech, both subjects had enough to say to fill the allowed minute.

Finally, during the acquisition, we had an opportunity to record additional data with S_B. This lead to the creation of utterances recorded in non-sagittal slices at an angle. This part, however, was filtered out in my work.

Technical settings

The data were recorded on a Siemens Prisma-fit 3T scanner (Siemens, Erlangen, Germany) at Max Planck Institute in Göttingen, Germany. We used radial RF-spoiled FLASH sequence [UZV+10] with TR = 2.02 ms, TE = 1.28 ms, FOV = 19.2×19.2 cm, flip angle = 5 degrees, and slice thickness is 8 mm. Pixel bandwidth is 1600 Hz/pixel. Image resolution is 136×136. The acquisition time varied from 34 sec to 90 sec, mostly about 60 sec. We followed the protocol described in [NZK+13]. Images are recorded at a frame rate of 55 frames per second with the algorithm presented in [UZV+10] (more on the frame rate will follow in Chapter 4.2.1).

The subjects are two adult male French native speakers speaking French. Subject A, S_A, is the same as in the static dataset from Chapter 3, now with the following measurements: male, 35 years old, 182 cm tall and 74 kg. Subject B, S_B, is male, 32 years old, 180 cm tall and 65 kg.

Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner by using a FOMRI III optoacoustics fibre-optic microphone. The subject wears earplugs to be protected from the noise of the scanner, but is still able to communicate orally with the experimenters via an in-scanner intercom system.

Since the sound is recorded at the same time with the MRI acquisition, there is additional noise of the machine in the audio signal. In order to denoise it, we used the algorithm proposed in [OVB12]. Since the noise was so strong, disruptive and present also in the frequency bands of speech, the denoising algorithm ended up in removing some energy from speech formants too, which was especially noticeable in nasal sounds. The resulting signal is considerably more intelligible than the original, but the noise is still present.

Transcribing the corpus

Per each wav recording, I produced a text file with the text of what the speaker says there. The tokens with non-standard pronunciation were stored in a separate simple text file along with their phonetic transcription in SAMPA [W+97] (such as "j_pense S @ p a~ n s" reflecting the fact that, as it is common among native speakers of French, S_B devoiced his /ʒ/ from "je" /ʒə/ before a voiceless fricative /ʃ/)—/ʃə/.

The transcription procedure was based on the guidelines for [SMWC03]:

- Numbers are written in words;
- Acronyms (that are pronounced letter by letter) have to be written letter by letter: "*Je suis allé en prépa scientifique, en P C S I*";
- Punctuation marks: *, . ? ! "*
- Disfluent speech:

  - Filled pauses ("*euh*", "*hein*");
  - Other hesitations:
    * Long syllables: marked with colon ("*donc :*");
    * Pauses: marked with the plus sign ("*+*" or "*++*" if it is very long);
  - Fusions: marked with an underscore ("*ch_ pratique*" to mean "je pratique" but realized as /ʃpʁa.tik/, "*ça va_ êt' Internet*" to mean "ça va être Internet" but realized as /sa.vɛ.tɛ̃.tɛʁ.nɛt/);
  - Partial words: marked with an apostrophe, if it seems to be an intentional way to increase the speed of speech while retaining intelligibility, or a hyphen, if the word is just abandoned, which is often followed by a restart — see below ("*bah je rent' assez tard*", "*dans l- -- dans la jungle*");
  - Restarts: marked with double hyphens ("*pas de petit-déjeuner mis à part, euh – mis à part du café*");
  - Mispronunciation, non-standard words: asterisk before the word, no space ("*je me demande si ça sert à *queqchose*");

- Unintelligible speech: the closest guess, if there is any possible, put inside double parentheses ("*((mais bon))*" or "*en tant qu'un amateur, ((cvraiqu)) j_ vais : dans des musées*");
- Interjections: treated as any other lexemes ("*bah*", "*hm*", etc.);
- Actions: "*{BR}*" for breath, "*{CG}*" for cough, "*{LS}*" for lip smack, "*{LG}*" for laughter, "*{NS}*" for a loud background noise (aside from the noise of the machine which is always present).

This procedure posed some limitations. In particular, since the bilabial trill /β/ and the alveolar trill /r/ do not belong to the French language, they are not recognized by French SAMPA. Therefore, they had to be mislabeled: such sequences as /aβa/ and /ara/ had to be transcribed as /aba/ or /ava/ and /aʁa/.

A phenomenon that turned out to be impossible to represent within the framework of SAMPA was stops with a long block phase, occurring in sequences like "*crabes bagarreurs*", /kʁabˈ.ba.ga.bœʁ/: the first /b/ has no audible release and is followed by the next /b/. Their transcriptions were decided upon on an individual basis.

Phonetic labeling

To translate the transcribed text into its phonetic transcription, I tried out two tools: Astali [FMJ15] and eLite HTS [RBBD14].

The benefit of Astali is the fact that its phonetic labeling is more flexible, giving me a way to take into account the non-standard phonetic tokens I took note of in the supplementary phonetizing dictionary files. Its disadvantage, however, is the fact that its output was a Praat TextGrid file [BW18] with no linguistic information, while I needed a more complete coverage to train an articulatory-acoustic model.

The output of eLite HTS is in the HTS format [Zen06], and it gives every phoneme a richer piece of information: it includes the phoneme identity itself, phonetic context (two phonemes before and two after, thus forming a quinphone together), position in the syllable (forward and backward), stress, accent and the number of phonemes in the syllable (now, in the previous syllable and in the next one), the vowel in the current syllable as well as text parsing around the phoneme and utterance-level information.

All in all, it was more reasonable to keep using eLite HTS. However, upon a closer look it turned out that the tool had a bug: despite passing to it input texts in a correct encoding, it processed them in ASCII, resulting in an incorrect treatment of all French diacritics: whenever a character was not recognized—let us say, *é* that should be phonetized as /e/,—it was phonetized as /a/. Besides, the tool could not deal with apostrophes or other symbols, thus mishandling even more cases.

To fix phonetic labeling of eLite HTS, I employed a set of text replacement rules so that the instances that were known to impede eLite HTS were replaced with character sequences that it would handle correctly (see Table 4.1). Naturally, these substitutions hindered the parsing unit of eLite HTS, so for any given text I ran eLite HTS twice: once to obtain the correct phonetic information but incorrect syntactic one and once vice versa.

To combine the outputs, there was an issue that the correct phonetic labels did not have to have a one-to-one correspondence to the syntactic ones. Sometimes it would be expected to insert, delete or substitute elements in one sequence of labels or the other. Thus I matched the sequences through Levenshtein's distance [Lev65]. I put the cost of substitution to 1 and the costs of deletion and insertion to 2, so that a correction of the label is more likely to happen than a copy or an omission.

This processing fixed all mistakes identified when reviewing labels, but also created some new. For example, the character sequence "*ai*" can be phonetized correctly: *faire* /f ɛ ʁ/; however, when treating the word "*système*"—normally phonetized as /s i s t ɛ m/—and replacing the character *è* with the sequence *ai* ("*systaime*"), which avoids using accentuation but should be phonetized the same, eLite HTS produces an inexplicable phonetization /s i s t ɛ i m/, i.e. the sound /ɛ/ is correct and should cover both letters "*a*" and "*i*" but is followed by an extra /i/. Having no open source code of the tool, it is difficult to determine the cause of this error.

Finally, e-Lite HTS assumed that every file started with a phoneme, while actually the beginning of every recording, albeit short, was silence. So I prepended every output of e-Lite HTS with an extra phoneme *sil*.

The next step was to perform state force alignment to estimate the boundaries of the phonetic labels. I used HVite from HTK [YEG+02] with Merlin as frontend [WWK16]. For each speech file—in my first run a whole single-acquisition recording (one minute, typically)—HVite loaded the corresponding label file, expanded it to create an alignment network and produced a phonetic transcription according to the output probabilities of the HMMs in the network.

Then, to improve phoneme alignment and reduce subsequent computational load, I split each recording into sentences according to the timing of state force-aligned label files and each textual transcription according to the French sentence tokenizer `sent_tokenize` of NLTK [LB02] and manually checked the cut-off points for the audio; it turned out that often, especially with spontaneous speech and syllable by syllable enunciation, the force-alignment algorithm put sentence

Character combination	Replacement
*, (,), ”	empty string
- at the end of a word (marks the speaker cutting their flow of speech)	empty string
: (marks a long syllable)	empty string
?, !	.
', ; '	empty string or a blank space; if possible, replace the word before the apostrophe with its complete form (*que*, *lorsque*, *jusque*, *presque*, *le*, *la*, *je*, *me*, *te*, *se*, *ce* (especially important, to avoid phonetizing it as /k/), *de*)
rés or *dés* at the beginning of the word + vowel	*rer* or *der* + white space + *z* + vowel
other instances of *és* + vowel	*er* + white space (to avoid pronouncing the *r*) + *s* + vowel
é, possibly followed by a silent *e* or a combination of *d*, *s* and *t* at the end of the word	*er* + white space (to avoid pronouncing the *r*)
è	*ai* (to obtain the sound /ɛ/)
à, *ù*	*a*, *u* respectively
ê	*ai*
â, *û*, *ô*, *î*	*a*, *u*, *o*, *i* respectively
ë	*e* at the end of the word (e.g. *ambiguë*), otherwise *ai* (*Noël*)
ï	*i* + white space (e.g. *naïve* → *na ive*)
ü, possibly followed by an *e* at the end of the word	*u*
ö or *ü* (words of German origin retaining the original orthography, not present in the corpus)	*oe* and *ai* respectively
ç	*s*
ñ (words of Spanish origin retaining the original orthography, not present in the corpus)	*gn*
œ	*ae*
œ	*e*

Table 4.1: Replacing characters in eLite HTS inputs so that they get phonetized correctly.

boundaries incorrectly. In such cases the text was corrected to correspond to the audio, even though this did not represent a full sentence then. This way the entire corpus gets transformed into a collection of short text and audio phrases, barring the files where phonetization of e-Lite failed due to internal errors. Then I regenerated phonetic labels from the sentence transcriptions and performed another run of state force aligning the sentence audios with those labels.

After all the improvements, the quality of phonetic annotation became quite reasonable provided that the label file listed the correct phonemes for the sound sequence. Unfortunately, this was not always the case because of the absence of processing non-standard pronunciations (e.g. the elision in *“*queqchose*” /kɛk.ʃoz/ instead of “*quelque chose*” /kɛl.kə.ʃoz/), errors in

splitting into single-sentence recordings or errors in phonetization.

Synchronizing audio, video and phonetic labels

Despite the fact that MRI acquisitions are conducted in a way that is as controlled as possible, with stable settings and a fixed position of the subject, there are factors that will inevitably vary from one acquisition to another. The most impactful of them is the temperature. The machine itself and the air inside it—as well as the body of the subject—can heat up and cool down along the acquisition time, according to the periods of continuous use or breaks. This can cause variation in image acquisition rate that can go up to 1%. In our case, the expected rate is 55 Hz, which means an image every 18.18 ms, and a one-minute long acquisition should have 3300 images. A variation of 1% would mean between a frequency between 54.45 and 55.55 Hz, that is, taking from 18.00 to 18.36 ms to make one capture.

Alone, a shift of ± 0.18 ms is negligible for articulatory events that are captured by RT-MRI. The only place where such a short period of time would be sensitive is the vocal folds, but their activity is not registered anyway. However, when this varying time per image accumulates over the course of the acquisition, aligning the audio track, the duration of which stays as set in advance before the acquisition, with the resulting video can lead to a considerable shift up to around 600 ms per minute, since we could be dealing with up to around 33 images that "go missing" or are unaccounted for. This duration can correspond to around ten phonemes in a stream of speech, thus misplacing the phonetic labels, especially at the sequence end.

To accommodate to that, once I split the audio and the transcriptions of the corpus into sentences, I analyze the corresponding sequences of images per each acquisition. Due to the setup, the first few images do not correspond to any sound due to the gap of 70 ms from the start of recording the image to the start of recording the audio. As for the images in total, just as predicted, their number varies from one take to another. In the treated cases, 95 acquisitions had more images than anticipated (8.39 more images on average, a standard deviation of 5.70), 11 had fewer (5.64 fewer images on average, a standard deviation of 6.09), and 3 had exactly the anticipated number of images. In total it makes 6.74 more images on average, a standard deviation of 7.15.

A failproof method to aligning images to audio timing is to find a phonetic label with a very short and recognizable visual cue, for example, /p/ requiring the lips to close, at the beginning and the end of the sequence. The images corresponding to this label will then be identified unambiguously, and having a clear correspondence between time stamps at the two ends of the sequence and the images should extrapolate to other images quite correctly. This is, however, an approach that is better to apply manually rather than through automatic detection of lip closure because of the many possibilities how it can go wrong if one does not also check the neighboring vocal tract shapes and how aptly they reflect neighboring phonetic labels: for example, the lips may or may not also close for periods of silence, or the lips may be recognized as open because of the contact being too fleeting or because of a processing error.

Thus, to associate every image to an HTS label, I made the assumed frame rate flexible, corresponding to the number of available images for the given audio, taking care to check that the estimated frequency stays within the predicted limits of 1% variation. This way I identified a 81811 ms-long sequence with 326 images missing (that is, no captures for some 6000 ms).

In the end I obtained a correspondence between any frame of a video from the corpus, the

time in the audio and its associated phonetic label. As force-aligned labels are provided every 5 ms, technically this means skipping a few labels with each new frame. In reality, however, phonemes do not change as frequently, so labels repeat, and skipping a few does not create a problem.

The next step, which is going to be covered below, was to process the frames and extract articulation-relevant information from them. The timing of these articulatory parameters revealed certain issues with the synchronization that was presented in the current section.

Processing the frames

An RT-MRI capture, as shown in Figure 4.1a, has a lot of areas that are irrelevant for studying articulation. To perform rigorous automatic processing of the images, we need to process the image so that it only contains the area of the vocal tract and we are able to extract articulatorily relevant information from it.

When choosing methods to do so, since the purpose of working with RT-MRI data was to have a complete information on the position of the articulators of the speaker, the preference was given to those that lose as little precision in the shapes of the articulators as possible—rather than, for example, a preference for the speed of computation or the interpretability of the resulting images by a human.

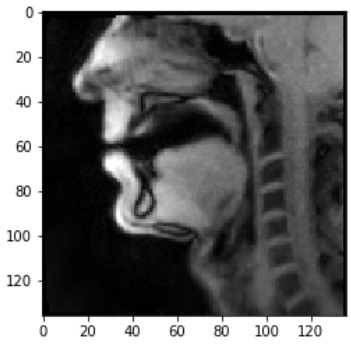


Figure 4.1a: An example of an original RT-MRI frame, before processing.

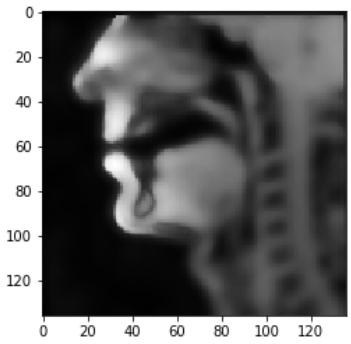


Figure 4.1b: Smoothing the frame in Figure 4.1a with a bilateral filter.

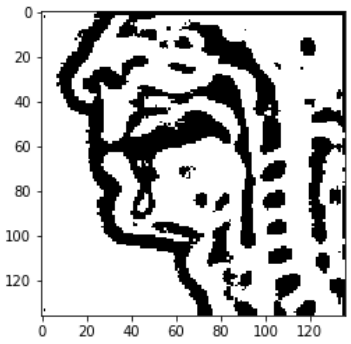


Figure 4.1c: Applying adaptive thresholding to the smoothed frame in Figure 4.1b.

RT-MRI frames are rather noisy, so the first step taken by Ioannis Douros was to reduce noise through smoothing [Sze10], an operation where each pixel is substituted by the result of some kind of operation on its neighborhood, most commonly by a weighted sum of the pixels around it. This idea works well due to the fact that typically, images change their pixel values gradually over space, and averaging a neighborhood can help get rid of the outliers, the pixels that are less correlated with the filter's center. However, a problem here is that boundaries of objects (in our case, the outlines of the articulators) represent edges, and the assumption of a slow change fails. We may average away the line that was previously sharp and lose some essential articulatory information.

This is why smoothing was chosen to be performed through applying a bilateral filter [TM98], which was developed so as to preserve edges: in the regions where pixel values are quite similar

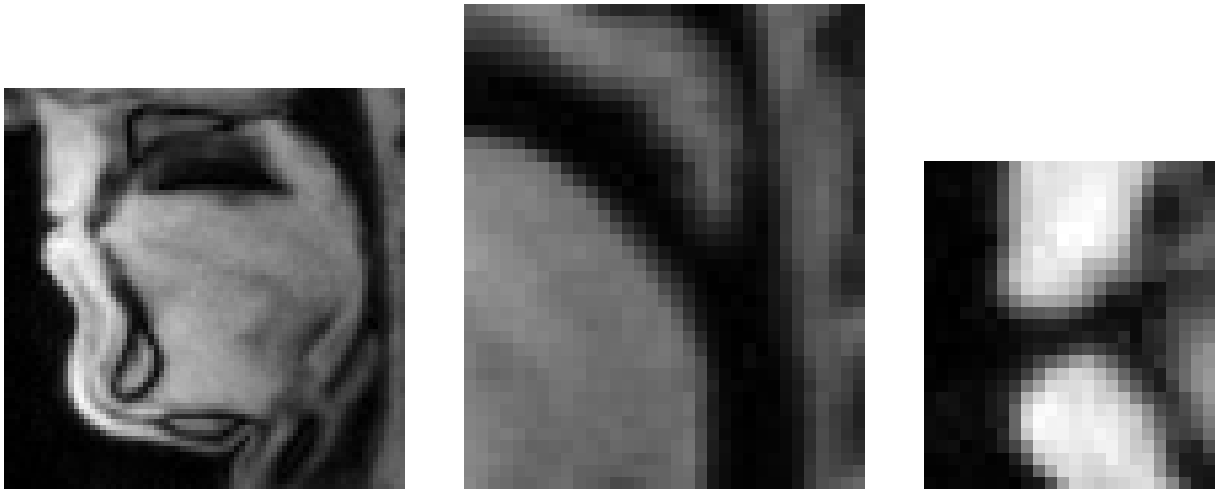
to each other, it acts as a standard domain filter, but it will not cross a sharp boundary in this averaging. The diameter of each pixel's neighborhood was $d = 9$ (the original image size being 136×136 pixels; such a value of d corresponds to a very large filter) and both filter sigmas in the color and coordinate spaces equal to 75, which is a moderate change. This parameter setup is acceptable for offline applications that need heavy denoising rather than the speed of computation. An example of smoothing a typical RT-MRI frame from the database (Figure 4.1a) is given in Figure 4.1b.

Once the image was smoothed, we wanted to simplify it for further treatment. This can be done with thresholding. The idea of thresholding a grayscale image is to replace all pixel values above a certain threshold with one color (white), and all pixel values below with another (black). However, a peculiarity of MRI captures is that pixel intensity is not the same in different regions of the frame, depending on the MRI coils—loops of conductive wire around the core, used to create a magnetic field. The nose and the frontal regions of the vocal tract—the lips, the tongue tip—are visualized with a much greater intensity, as almost white shapes, than the back regions (the pharynx, the larynx), where we can only see blurred gray outlines. For such images, simple thresholding with a globally assigned threshold value is not a suitable method; it is best to have a local threshold value per each luminosity region. This approach, chosen by Ioannis Douros, is called adaptive thresholding. A `blockSize`×`blockSize`-region can be characterized by its average value, but to use it as a threshold means to keep in too much noise. A better solution is to set the threshold value $T(x, y)$ based on the neighborhood's (a 11×11=rectangular around point (x, y) in our case) cross-correlation with a Gaussian window, minus an adjustment constant $C = 2$. Flooding the space outside of the head of the speaker with the white color, we obtained images as shown in Figure 4.1b.

To facilitate articulatory information extraction, it was important to limit the study to a single area of interest at a time. We used the tip of the nose as one point of reference, which is the middle of the leftmost points on the speaker's head. In cases when the nose tip strayed too far from where it was on average in its sequence (more than 3 pixels in either of the dimensions), since the speaker's head was not supposed to be moving, I assumed it to be an error and reset it to a sequence-specific default value. From there, I identified the values to hard-code windows containing (a) the vocal tract, Figure 4.2a (used for the work in the next chapter), (b) the velum (used for the work in this chapter), Figure 4.2b and (c) the lips, Figure 4.2c (used for the work in this chapter). Theoretically the relative location of these windows with regards to the nose tip could depend only on the speaker's anatomy and be therefore speaker-specific, but since the head angle varied from one acquisition session to another, it had to be session-specific. Additionally, in the individual images where I could automatically identify that the algorithm was failing to extract the features (see the details in the following subsections) due to a slightly misplaced window, the window was shifted in the direction that was expected to solve the problem.

Extracting features

Ideally, an articulatory speech synthesizer should work with articulatory information that is as rich as possible. There also is a very strong requirement for this information to be highly precise so that it remains interpretable and sound from the point of view of the physics of speech production. The approaches to get this information can be manual, semi-automatic and



(a) (b) (c)

Figure 4.2: Vocal tract (4.2a), velum (4.2b) and lips (4.2c) windows for the RT-MRI frames.

automatic.

For a given mid-sagittal MRI frame, static or dynamic, a trained human with the knowledge of phonetics and of the anatomy of speech organs can draw the contours of almost all the organs that are known to play part in speech production. The training is necessary to learn to be consistent, for example, in the choice where the lips begin and end or how the bottom part of the tongue is annotated when it is visible, and absolutely essential to correctly handle difficult cases, such as:

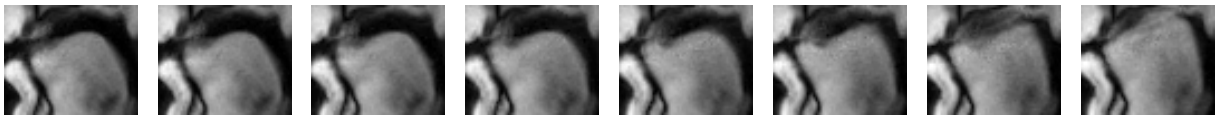
- The tongue tip is one of the fastest articulators, participating in rapid, precisely controlled wide-range motions to produce numerous phonemes, such as the voiceless alveolar stop /t/. The tongue dorsum is less so, but thanks to jaw displacement it can be agile too. These movements result in ghosting, for example, an image with two outlines of the tongue or a physically impossible shape, or in merging the boundaries of the tongue with a neighboring articulator (see Figure 4.3 for an example).

- It may appear that the uvula becomes inflated and increases in volume, which really is either blurring (Figure 4.4a) or the uvula rolling up onto itself (Figure 4.4b). Besides, the velum is represented with a blurry shape of only a few pixels' width. Sometimes the phoneme that is being produced needs to be taken into account to correctly identify whether the velum touches the pharyngeal wall or not: if it is a nasal sound, there should be a contact to close the velopharyngeal port, and if it is an oral one, no.

- The epiglottis is thin. Being set deep in the vocal tract, far away from the frontal side that MRI visualizes with a greater intensity, the precise shape of the epiglottis can be difficult to discern (Figure 4.5b). Besides, it can press up to the tongue, thus hiding its edge (Figure 4.5a).

- As a resonating system, the vocal tract is characterized not only by its width at any point, but also by its length. One end is clear to see, the lips; meanwhile, the other end, at

the vocal folds, can only be guessed upon (Figure 4.6), since the position of the larynx is extremely unstable in visualization, both due to its anatomical structure (very little tissue to see in the mid-sagittal frame) and due to its position at the rear of the vocal tract.



/e/ /s/ /s/ /s/ /s/ /ə/ /k/ /k/

Figure 4.3: The tongue in a fragment of an RT-MRI sequence /s(ə)k/: note the double blurred outline of the tongue tip to produce the alveolar fricative /s/ and the blurred shape of the tongue dorsum when making a constriction for the velar stop /k/; in the latter case, it is also difficult to see where the palate ends and the tongue begins.

(a) Rapid motion of the velum and therefore blurring (in the middle): S_B pronouncing /ʁ/.

(b) "Swelling" that is actually the velum's rolling up on itself: every third frame in S_A's production of /uʁi/.

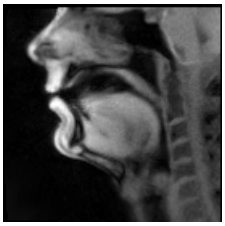
Figure 4.4: Two main causes of the velum contour to be difficult to annotate: rapid motions and an intricate geometry.

Aside from being this challenging, manual annotation also is quite taxing and extremely time-consuming. Coupled with the copious amounts of data in the RT-MRI corpus (around two hours of speech), all the points above ruled out the option of making annotations by hand.

Another solution would be semi-automatic: to have a baseline annotator and use a human to correct the curves. The problem with this approach was that we needed the data to be up to a very high standard, sound from the perspective of physics and with a great degree of consistency. This lead to the need to perform at least minor corrections in every single image, and evaluating the annotator's performance showed that corrections did not get to be done any easier or faster than drawing the contours oneself.

A more automatic approach would be to perform manual preparatory work on some RT-MRI images and run an existing contour extraction and identification method, for example, annotate

(a) The epiglottis pressing up to the tongue: S_B pronouncing /a/.



(b) The epiglottis being barely visible in the picture: S_B pronouncing /ã/.

Figure 4.5: Two main causes of the epiglottis contour to be difficult to annotate: merging with the tongue and being barely visible.

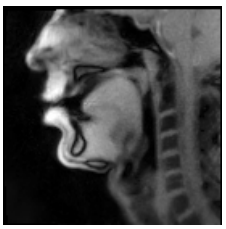


Figure 4.6: The outlines of the larynx can be very difficult to discern, which can lead to a wrong estimated position of the vocal folds and a wrong vocal tract length: S_A pronouncing /k/.

a part of the frames and use it to train a statistical or deep learning model. Despite the challenges posed by high variability in the vocal tract shape and configuration when producing the extensive variety of speech sounds, differences in speakers' anatomy and the connectivity of the tract airway to other channels of air through lip opening, velum opening or larynx [RRUC13], this has been a prolific direction of research, both with articulatory data that was more in use in past decades and those that are being explored now, in particular, RT-MRI captures. Some notable examples are methods that start off a predefined curve for an articulator or a predefined vocal tract shape and try to avoid deforming it too much when transitioning between frames, such as a geometrically constrained snake model that also relies on the information from landmarks, contact points and pronounced curvature areas [SWF+18], or models relying on the vocal tract appearance [AE17] or the appearance of some vocal tract components [TN15]; unsupervised tracking methods, usually providing less informative outlines, such as [BN08, LRP+13, PBKN10]; and supervised methods, relying on annotated examples and/or landmarks, such as [EB11, RRUC13, KG18, TGH+19].

In this direction I explored the image segmentation method of [NTR+14]. Its idea is to create templates—correctly annotated frames,—run the segmentation algorithm, identify frames with significant delineation errors, correct their contours, save the new contours as a template and rerun the algorithm. An example of a template that I fitted to our data is shown in Figure 4.7.

Nevertheless, image segmentation proved to be a hard problem, particularly exacerbated in our case by the need to have results that would be very precise. Having weighed in the expertise and the estimated long hours of work necessary to identify images to serve as templates and to treat them, it was decided to proceed with fully automatic approaches, limiting ourselves to less rich information while retaining precision.

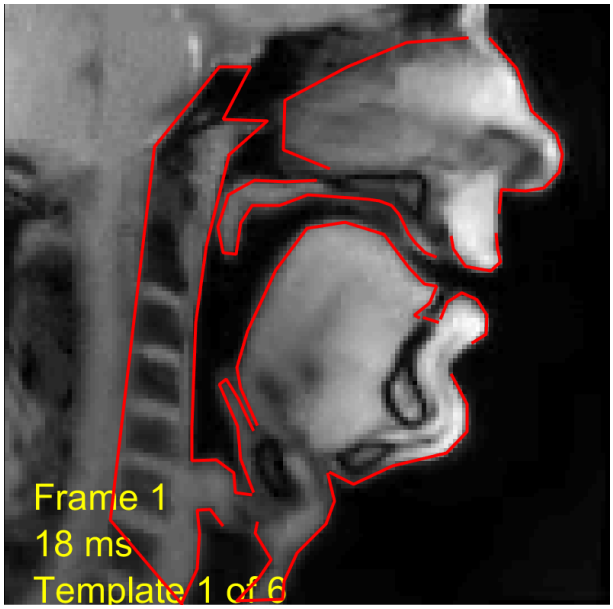


Figure 4.7: An example of a segmentation template fitted to the ArtSpeechMRIfr RT-MRI data (S_A, /ə/).

The articulator that was the easiest to process was the lips. For speech production in French, the most significant attributes about them are the following values:

- Whether the lips are open or closed, i.e. whether there is a labial obstruction of the air passage in the vocal tract.
- If they are open, how wide, i.e. how large is the area at the outer end of the vocal tract.
- If they are closed, how tightly they are pressed together, i.e. whether there is any pressure building up behind their closure.
- Whether they are protruded or not, i.e. whether the vocal tract is made longer due to their location on the anteroposterior axis.
- Whether they are rounded or not, i.e. the shape of the vocal tract at the outer end.

All features but the last one are reflected in the mid-sagittal frames, and this information can be reliably extracted from a window containing the lips.

Another significant articulator that still can give us some useful information even without the complete knowledge of its entire contour is the velum. It controls whether the air goes through the oral and/or the nasal cavity when producing speech: when the velum is down, the velopharyngeal port is open and the airflow can go through the nose producing a nasal sound, and when it presses up, this passage is blocked and the air is contained in the oral cavity, making the sound oral.

Furthermore, the velum can come in contact with the tongue. Here we should mention two relevant places of articulation, the one involving the large body of the velum and the other the smaller uvula dangling at the end of the latter, and three manners of articulation: approximant

(e.g. the voiced labio-velar approximant /w/), fricative (e.g. the voiced uvular fricative /ʁ/) and stop (e.g. the voiceless velar stop /k/).

Therefore, the most important attributes of the velum are:

- Whether there is a passage between the velum and the pharyngeal wall;
- If there is, how wide it is (determines the intensity of nasalization);
- Where the contact between the velum and the tongue is (whether it is a full contact between the velum and the tongue or only the uvula and the tongue—determines the place of articulation);
- The temporal characteristics of this contact (determining the manner of articulation: a prolonged narrowing of the vocal tract makes an approximant, a constriction that, because of turbulence, rapidly changes between a wide and a narrow one or even a contact, makes for a fricative, and a prolonged contact with a subsequent burst makes a stop);
- The spatial arrangement of the airflow: whether it is completely obstructed, or there is a constriction, or the airflow is blocked at the center of the vocal tract by the uvula, but there still is a passage for it at its sides, thus creating a lateral sound.

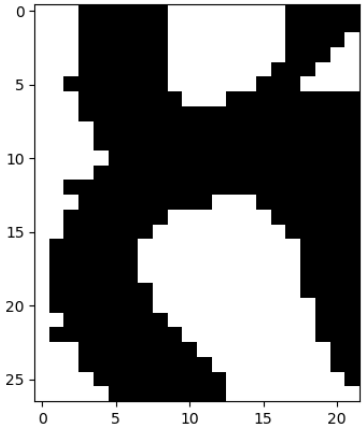
Most of these aspects are, at least to some degree, represented in RT-MRI data, the limiting factors being the absence of other planes but the mid-sagittal one (no information regarding the potential lateral ways for the airflow) and a relatively low frame rate (blurring the frication between the articulators together). For the purposes of articulatory speech synthesis the available amount of information should still be useful, so we proceeded with extracting it.

The procedure we followed was a result of a collaborative effort of Ioannis Douros and myself: Ioannis Douros conceived the general idea to narrow down our view to specific windows and to use seed points, color fills and contour detection tools to look for contacts between the organs in those windows (see details below); then I worked completely on my own on producing articulatory parameters (picking and developing methods for processing contours, readjusting the windows for better results, treating errors, unforeseen cases and minor noise in the image, giving informative feedback from the computation process, treating the calculated values for consistent temporal behavior, analyzing and evaluating the results).

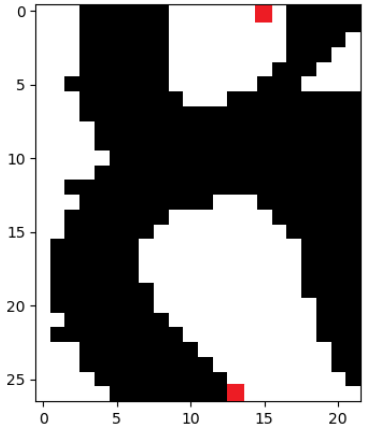
Extracting the articulatory parameters for the lips

Given a frame window containing the lips as in Figure 4.8a, first we discard (i.e. fill with black) the leftmost white area of the image that corresponds to the space outside the head of the speaker. Then we look for seed points of the lips (Figure 4.8b): the upper lip pixel is a white $(x, 0)$ pixel that is closest to the expected $(15, 0)$ and belongs to an area of at least 8 pixels; the lower lip seed could be the first white pixel from the left-bottom corner of the image rightwards, except that the first white pixel normally corresponds to the white space outside the speaker's head, so it should be the first white pixel seen after a span of black pixels that surround the speaker's head, with the same condition on the area size.

If the algorithm fails to find a seed point for the upper lip, it is assumed that the window was placed too low, and the process is repeated with the window one pixel higher; the symmetric process is done for the seed point of the lower lip and the window shift by one pixel down.



(a) Raw filtered image



(b) The upper and lower lip seeds

Figure 4.8: A sample image to extract articulatorily relevant information from the lips. The leftmost white area corresponds to the end of the vocal-tract-related part of the image and is thus irrelevant. The two elongated shapes in the center are the lips. There is the tongue tip approaching the upper lip at the right. The seed points are indicated in red.

The seed points can be used to suppress all contours detected with the border-following algorithm of [S+85] but those that correspond to the lips. Figure 4.9, from left to right, shows the various cases we may encounter: If we find two smaller contours, the lips are apart; if we find a tall contour stretching over the entire image height, the lips are closed. A complicating issue may be the tongue touching either one of the lips or both, which typically happens outside speech segments as the speaker licks their lips, but can nevertheless produce unrealistic values or contours.

This way, the fact whether the lips were open or closed was the first piece of articulatory information extracted from the images.

If, however, any of the contours assumed to be the lips touches the left border of the image, it is assumed that the window was placed too much to the right and needs to be shifted to the left in order to include the outermost points of the lips. The window is shifted to the left, and the previous steps are repeated.

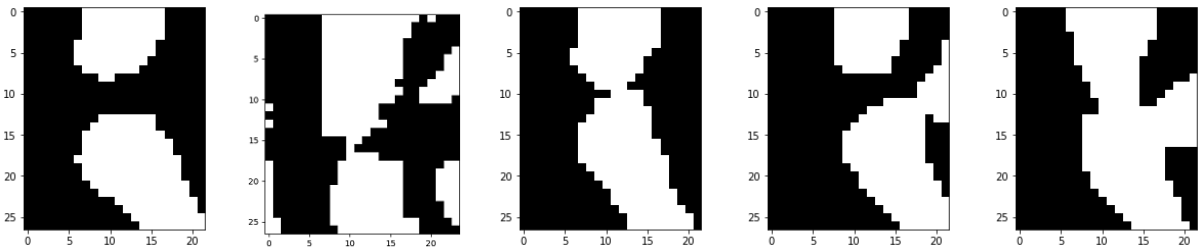


Figure 4.9: Cases to be treated when extracting the articulatory information from the lips, from left to right: open lips, fleetingly closed, closed, the tongue touching the lower lip and both the lips and the tongue being in contact.

Then, whenever there were two contours, we could find the distance between them (the smallest euclidean distance between their pixels borders—see Equation 4.1 below), which served as the value of their opening—the second articulatory parameter extracted from this window: Equation 4.3.

$$\mathrm{eucl}(A, B) = \sqrt{(\mathrm{displ}_x(A, B))^2 + (\mathrm{displ}_y(A, B))^2}, \tag{4.1}$$

where displ_{axis} is displacement along a given axis:

$$\mathrm{displ}_{axis}(A, B) = \begin{cases} |A_{axis} - B_{axis}| & \text{if } A_{axis} = B_{axis} \\ |A_{axis} - B_{axis}| - 1 & \text{otherwise} \end{cases} \tag{4.2}$$

$$ls\_dist = \min(\mathrm{eucl}(U, L)), \begin{cases} U \text{ being a pixel on the border of the upper lip contour,} \\ L \text{ the lower lip} \end{cases} \tag{4.3}$$

Figures 4.10a–4.10c marks how the distance between the lips is computed.

The meaning of having a distance between pixel *borders* rather than pixel coordinates themselves can be shown on the following example: if the lowest point of the upper lip is $(10, 10)$, and the highest point of the lower lip is $(10, 11)$, the traditional euclidean distance between them will be 1, while the contours are obviously in contact, sharing a common pixel border.

Having distances in pixels is a solution that is convenient for computations; naturally, all the values can be translated back into mm, since the distance between any two adjacent pixel centers (i.e. the value of 1 pixel in our implementation) is 1.412 mm.

As for the case when the lips were in contact, the ls_dist distance was set to 0, and we calculated the surface of their contact ls_cont as follows—see Figure 4.10e:

$$ls\_cont = \min(\mathrm{eucl}(O, I)), \begin{cases} O \text{ being a pixel from the outer contour of the joined lips,} \\ I \text{ the inner one,} \\ O_y, I_y < 0.8 \times h, h \text{ being the height of the window} \end{cases} \tag{4.4}$$

The vertical limit of the contact points being necessarily in the upper 80% of the image means to make sure that the thinning of the lower lip at the bottom of the image, the one that needs to be included in the window in case the lips open wide, does not get counted as the contact surface.

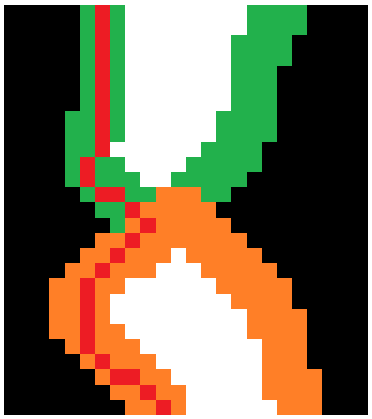
Then the contact surface was used to divide the joint area of the lips in contact into two separate lips touching, thus providing us with full lip contours in all cases.

These first three articulatory parameters are not independent: whenever the lips are open, the distance between the lips ls_dist is positive and the contact surface ls_cont is zero; other times, when the lips are closed, the distance between them is automatically zero, and their contact surface is not negative.

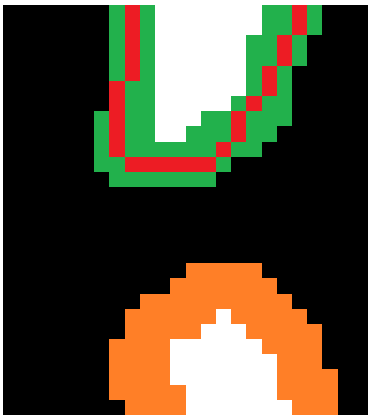
Then I calculated lips protrusion values up_l_pr and lw_l_pr—see Figures 4.10f:

$$up\_l\_pr, lw\_l_pr = RL_x - LL_x, \begin{cases} RL \text{ for the rightmost,} \\ LL\text{for the leftmost point of the upper or lower lip resp.} \end{cases} \tag{4.5}$$

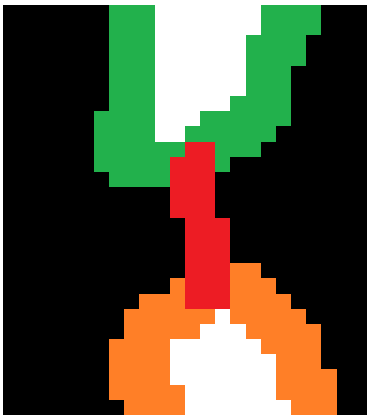
Assuming the window is stable enough and crosses the lips at approximately the same height, this difference should be informative enough about how far the lips protrude.



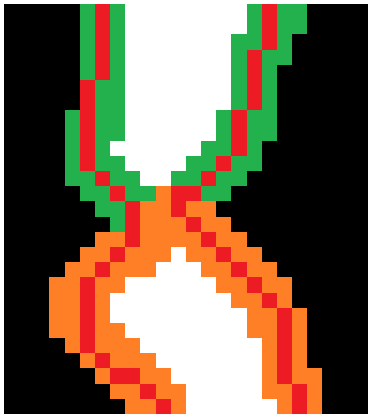
(a) The leftmost contour is tall — the lips are closed, ls_dist = 0



(b) The leftmost contour is short — the lips are open, ls_dist to be computed as the distance between the lips; ls_cont is 0.

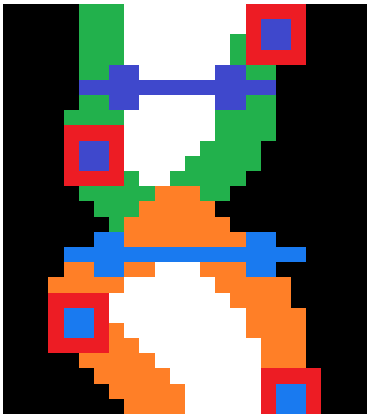


(c) ls_dist, the distance between the lips, is the shortest distance between the contours of the upper and the lower lip.



(d) The lips are closed. ls_cont is computed from the inner and outer contours.

(e) ls_cont is the distance between the outer and the inner contour, disregarding the lower 20% of the image (whitened out) that could produce a smaller distance.



(f) up_l_protr and lw_l_protr are the difference between the x coordinate of the rightmost and the leftmost points on the lips.

Figure 4.10: Examples marking how the lip parameters are computed. The upper lip is green, the lower lip is orange, the key elements are highlighted in red or blue.

Extracting the articulatory parameters for the velopharyngeal area

The next step was to extract the articulatory information of the velum. As discussed above, the velopharyngeal area of the frames has less constrast and brightness, which means that smoothing and thresholding it is less reliable and may bring in errors: feeble shapes and outlines can get discarded, and spaces can be recognized as tissue. Besides, the area has more articulators moving than it was in the case of lips, which makes it more challenging to reliably recognize

them automatically. Figure 4.11 shows what a processed window may contain, and Figure 4.12 shows how the presence of movement in the sequence can bring in some odd shapes.

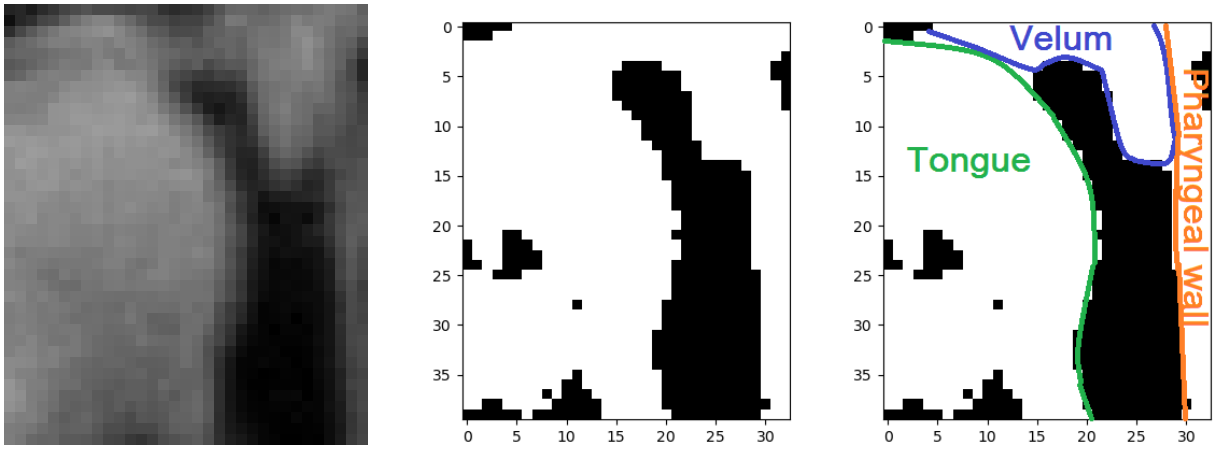


Figure 4.11: An example of a processed velum window: the initial window (the tongue and the velum in contact to produce /k/), the smoothed and filtered image, and its annotation.

Figure 4.12: Due to noise and movement, the extracted contours may be of unpredictable shapes: a sequence of frames where the loss of contact between the tongue (green) and the velum (violet rather than blue elsewhere in the chapter, to indicate that it is a joint contour of the velum and the pharyngeal wall that alone is marked orange) creates a highly irregular contour of the tongue (one frame before the last). The contour annotation error in the middle image (assigning the label "tongue" to the joint tongue-velum contour past the point of contact), however, is not related to this disformation but rather to the failure to recognize the two articulators as being in contact. The tongue-velum distance arising from this incorrect annotation was discarded at the subsequent consistency verification stage.

The biggest shape is the tongue, but even its recognition is not straightforward, since the layout of the muscle fibers creates darker areas in the body of the tongue that get processed as cavities. Their number, exact location and size can vary between one image and another.

At the right we have the border of the vocal tract that is the closest to the back: the velopharyngeal wall. Depending on the window, partly on the movement of the speaker but especially on whether the nose tip was detected correctly, the window may or may not also contain some of the cavities behind it that actually are filled with the vertebrae. This sets up a choice to be made when setting the limits to the window: including more space to the right

will guarantee that the entirety of the pharyngeal wall is present in the window at the expense of introducing there a lot more irregular and unstable shapes that will require processing, while making sure that the vertebrae do not get into the window

The major focus of our attention, the velum, is an articulator whose shape is particularly problematic to capture automatically. It may appear unattached to the more front area of the vocal tract; its shape, due to it being very thin and mobile, can be distorted, which can be even more aggravated by the heavy processing of the image. As for the window boundaries, depending on the speaker's head precise position, sometimes shapes from the nasal cavity are fully visible in the original large image, and sometimes not. Coupled with the fact that the velum may appear disjoint from the more frontal part of the vocal tract, it definitely rules out having a safety margin in the window to make sure that the entirety of the velum is contained in it, since the shapes in the nasal cavity would become indistinguishable from the shape of the velum. As the velum can curl, this, in turn, means that it may cross the upper boundary of the window multiple times (potentially, enter the window, leave it and reenter).

This explains why it may be difficult to recognize and process the outlines of these articulators even when they are clearly apart as in Figure 4.13. The task becomes even more difficult when they come into contact. Due to the velum curling, the velum and the tongue can touch multiple times within the same image, creating very irregular outlines. Past the contact, closer to the lips, there may or may not be space between the velum and the tongue; same goes for the space before the contact between the tongue and the pharyngeal wall, closer to the glottis. A situation when all three articulators are in contact can appear in many different ways, with spaces and not.

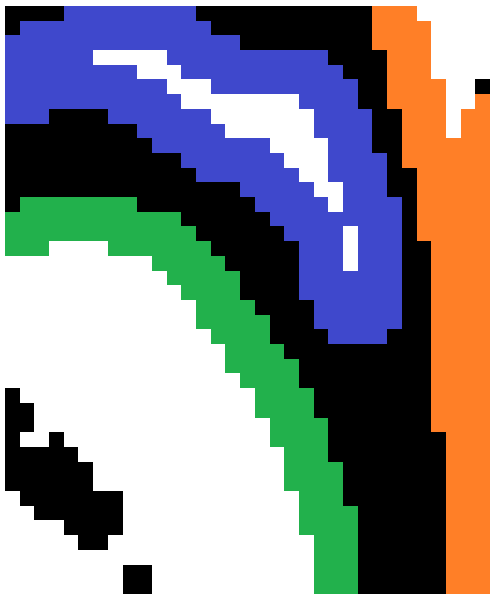


Figure 4.13: An example of automatically extracted contours in the window of the velum. Green stands for the tongue, blue for the velum, and orange for the pharyngeal wall.

Finally, contour detection, such as the chosen border-following algorithm of [S+85], is reliable when the entirety of the outline is contained within the image. When part of an object touches the border multiple times, coupled with inconsistencies such as the velum appearing to hang in

the air it becomes slightly unpredictable whether a shape will be considered part of one object or multiple ones. So, I had to break down all large border-crossing contours into small chunks that entered the image on the image border, followed a boundary and left the image. This was necessary to avoid falsely joined organs when they are not, actually, in contact, but created many more candidate contours to identify as the velum, the tongue or the pharyngeal wall.

First, I found seed points of the three articulators in question:

- The initial seed point of the velum was the closest white pixel (i.e. with tissue) to the upper left corner of the image as long as it belonged to a sufficiently large white area, and then it was shifted at most 7 pixels down as long as the color stayed white.
- The initial seed point of the tongue was the closest white pixel to the middle point of the left border of the image as long as it belonged to a sufficiently large white area, and then it was shifted up to 6 pixels to the right as long as the color stayed white.
- The initial seed point of the pharyngeal wall was the closest white pixel to the down right corner of the image if it belonged to large enough an area, and then if was shifted to the left as much as possible to find the border of the pharyngeal wall, the maximum shift being up to the middle of the image. No direction of search was preferred, as in case the entirety of the wall was in the image, it was possible to also find parts of spaces corresponding to the vertebrae, and they sometimes could appear disjoint from the rest of the wall.

If the seeds were found outside the areas where I roughly expected them to be, the corresponding images were saved as potentially problematic.

Then, to treat cases of contact, the following procedure was used to identify reference points for spaces between the organs:

- The point between the tongue and the velum: the first black pixel encountered on the line from the tongue seed point to the velum one; if it fails, try 15 pixels to the right. This way the point should be closer to the tongue surface than to the velum. Normally the tongue seed point is high enough that there are no black areas created by the muscle fibers on the way from this point to its velum counterpart;
- The point between the tongue and the pharyngeal wall: the last pixel in the first span of black pixels encountered on the line from the pharyngeal wall seed point to the tongue one.
- The point between the velum and the pharyngeal wall: first, a new pharyngeal wall point is found as the leftmost white pixel sharing the same white area with the pharyngeal wall seed point when looking from $(w-1, 8)$; then, the point is the last pixel in the first span of black pixels encountered on the line from the new pharyngeal wall seed point to the velum one—so, it would be closer to the velum than to the wall, but will not enter any small pockets created by the velum's curling.
- In a similar fashion, the newly constructed upper pharyngeal wall seed point could serve to find a point between the tongue and the pharyngeal wall that would be in the upper part of the image, closer to the velum.

Then, using the organ seed points, I identified organ contacts by coloring the white area around one seed point and checking whether the other seed point in question changed its color too. If no, the articulators were not in contact. If yes, there were two possibilities: either there is a true contact between them, or both of them touch the third articulator, but not each other. In the latter case, I identified a sub-window with the potential place of contact and repeated the process of coloring a point from one articulator and checking the color of the point of the other.

If it appeared that the tongue was in contact with the pharyngeal wall and the pharyngeal wall seed point was not at the very right of the image, most likely it was an error of the pharyngeal wall seed identification—see Figure 4.14.

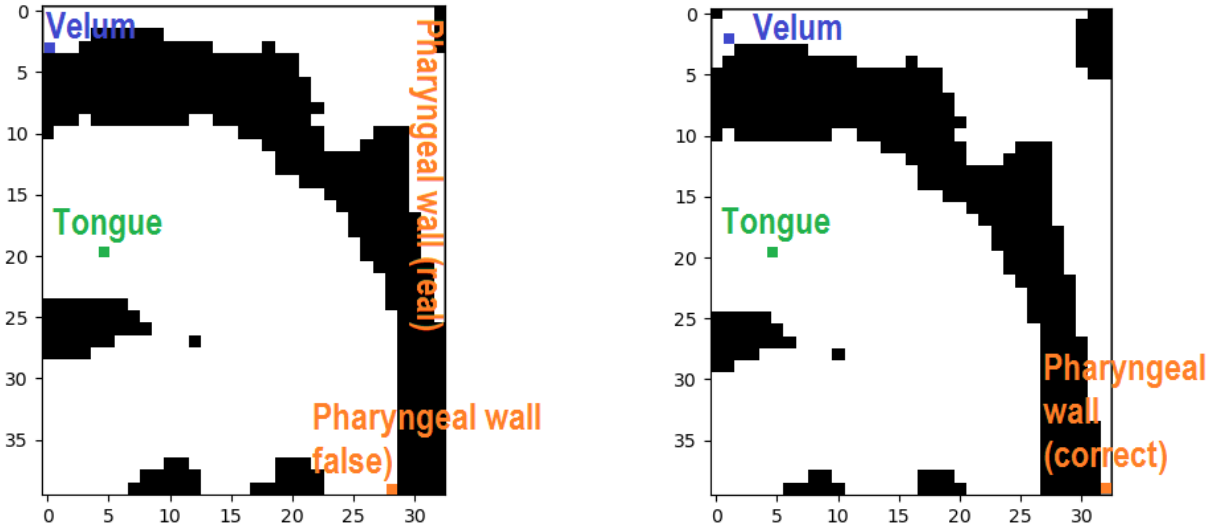


Figure 4.14: An example of an erroneous pharyngeal wall seed causing the tongue to be recognized as in contact with the pharyngeal wall: initial wrong guess (left), shifted correct window (right). The marked points are the seed points for the tongue, the pharyngeal wall and the velum.

Then I suppressed all shapes that did not contain a single seed point and applied the same border-following algorithm of [S+85] to find the contours present in the image. I cut them so that every contour would either be entirely contained within the image proper or begin at the boundary, go through the image and end somewhere else at the boundary.

There typically were numerous contours with irregular shapes.

If an organ was known to stand apart from the others, I was looking for one contour for it.

If an organ was known to be in contact with another one, I was looking for two contours: one for the part before the place of contact (this place having an *a priori* unknown location), the other for the part past the place of contact, closer to the lips. For example, if the tongue and the pharyngeal wall are in contact, the first contour would include the lower part of the pharyngeal wall, near the glottis, up until the contact point, followed by the lower part of the back of the tongue, and the second contour would include all the upper part of the pharyngeal wall until the place of contact below, followed by the outline of tongue dorsum.

If there was a chain of articulatory contacts, for example, the tongue touching the velum and

the velum touching the pharyngeal wall, but not the tongue touching the pharyngeal wall (as in Figure 4.11 when producing /k(u)/), I looked for the contour of the chain: tongue-velum-wall, velum-wall-tongue, velum-tongue-wall.

If all organs were touching each other (as in Figure 4.15), there was nothing to compute as taking note of the presence of contact was enough.

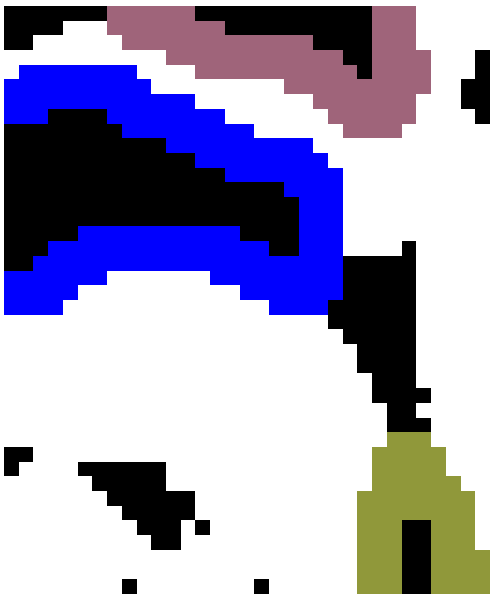


Figure 4.15: All articulators in contact with each other—no need to calculate the distances. Blue marks the contour that was labeled as the velum; purple the contour that was labeled as the velum and the pharyngeal wall; bottle green, the tongue and the pharyngeal wall.

First I made all possible guesses for which organ each contour could be.

`Special case`

In the special case when articulator X was in contact with Y, Y with Z, but not X with Z (see Figure 4.16), I searched for the joint X-Y-Z chain contour as the contour that passed close to all reference points (the upper part of the pharyngeal wall, the velum seed point and a reference point for the space next to the back of the tongue and the end of the velum) or was so large that it occupied more than 75% of the image and then broke down that chain into three contours: the tongue, the velum and the pharyngeal wall. The splitting points were found as the extreme points of the contour and the closest points to the reference and seed points.

No other case needed the contours to be split.

`Pharyngeal wall`

The contour of the pharyngeal wall, when it is not in contact with any other organ, needed to be sufficiently right, sufficiently tall, and its bounding rectangular needed to contain the pharyngeal wall seed (again, I used $eucl$ distance from Equation 4.1).

If the pharyngeal wall was in contact with the tongue, it had two options.

- One option for it was to be the contour whose bounding rectangular contained the identified point in the space between the pharyngeal wall and the tongue (the down part, closer to the glottis, before the contact point). This contour also needed to be contained in the lower right part of the image.

(a) The tongue in contact with the velum, the velum with the pharyngeal wall, but not the tongue with the pharyngeal wall.

(b) The velum in contact with the tongue, the tongue with the pharyngeal wall, but not the velum with the pharyngeal wall.

(c) The tongue in contact with the pharyngeal wall, the pharyngeal wall with the velum, but not the tongue with the velum.

Figure 4.16: Assigning contours in the special case of two articulatory pairs in contact, but not the third pair. To extract a correct distance between the two articulators that do not touch, it is necessary to cut the joint contour in three parts. As usual in this chapter, green stands for what was labeled as the tongue, blue as the velum and orange as the pharyngeal wall.

- The other option was to be the contour whose bounding rectangular contained the same point as above but shifted 30 pixels up and 5 pixels left (the upper part of the wall from the velopharyngeal port downwards to the contact point and the rest of the tongue from there).

If the velum touched the pharyngeal wall, the contour of interest could be the contour whose bounding rectangular contained the space point between the velum and the pharyngeal wall (see above).

Tongue

The contour of the tongue, when it was not in contact with any other organ, needed to encompass a large enough area, and that area had to contain the seed point of the tongue.

If, however, the tongue touched the velum, there were three options to consider.

- If the velum lay on the tongue with no space between them, then their joint contour was the one that we would have found for the tongue had it been standalone.
- If the contour's bounding rectangular contained the seed point of the space between the velum and the tongue, it could be the joint velum-tongue contour past the contact point, closer to the lips. To qualify, this contour also needed not to stretch too far to the right.
- If the contour's bounding rectangular contained the seed point of the space between the tongue and the upper part of the pharyngeal wall, it was the joint velum-tongue contour before the contact point, closer to the glottis. To qualify, this contour also needed not to have any points close to the left border of the image.

`Velum`

If the velum did not touch any other articulator, any contour that was sufficiently large and whose bounding rectangular contained the velum seed point was to be labeled as the velum.

Then, for every picture, the set of contours with their potential guesses were examined together to make sure they did not contradict each other: every articulator had to have a contour: one and only one for an isolated articulator, at least one for an articulator in contact; additionally, a contour could have several labels only for articulators in contact.

If the labeling was found to be inconsistent, I proceeded to reassigning labels until they became consistent or the number of repeated attempts to do that got larger than the number of articulators in question (three). To find a consistent labeling, I checked articulators one by one, all the while keeping track of the contours whose labeling had already been determined fairly certainly.

If there was no inconsistency found for a given articulator, its contour or contours got preliminarily approved to be labeled as the articulator and, when applicable, the articulator it touched.

If there was a problem, the algorithm tried to fix it in the ways described below.

If an articulator was not associated to any contour, first I found the contour that was the closest to the articulator's seed. This contour could not be small and located in the down left corner of the image because that was the characteristics of black areas falsely recognized as empty spaces in the tongue.

The same approach was attempted if an articulator was touching another one but had only one contour. If it did not yield a new contour to associate to the articulator in question, it should theoretically have happened only for the joint part of the tongue and the velum closer to the glottis. To identify that, I considered the contour closest to the seed point between the tongue and the upper part of the pharyngeal wall, since by design it is very close to the velum.

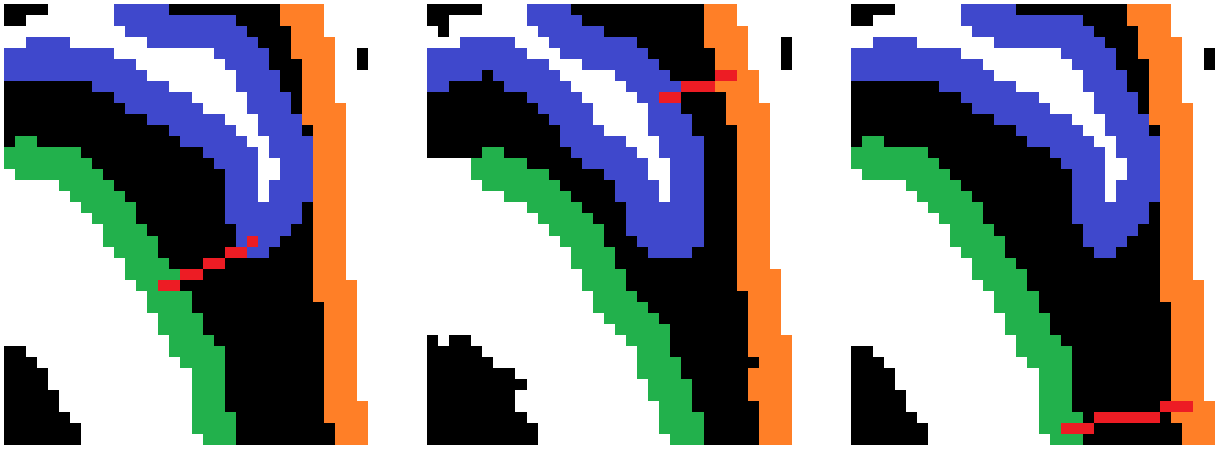
If an articulator was not found to be in contact with any other but got associated to multiple contours, the contour that was the closest to the seed point of the articulator became the one that got to keep the label.

Whenever such problem fixing created a conflict with an already preliminarily approved label or the new labeling turned out to be inconsistent (for example, if the tongue was not touching any other articulator and had its single contour, but then this contour got to be assigned to the velum), it triggered a new round of search for a consistent labeling of the contours, this time applied to the new way of labeling the contours.

If after this process the labeling was still inconsistent, the entire process of contour search and labeling was repeated for the potentially noisy contours extracted from the original smoothed and filtered image (rather than the image where all white areas not containing a single articulator's seed point were filled with black). If the labeling stayed inconsistent even after that, articulatory parameters extracted from the velum window were set to NaNs. Otherwise, having made sure that the shapes present in the frame were all properly labeled and the understanding of the picture seemed to be correct, I was able to calculate the following values as the minimal distances between pixels on borders of the respective contours—see Figure 4.17:

$$v\_t\_dist = \begin{cases} \min(eucl(V,T)), V \text{ for the velum, T the tongue} & \text{if no contact} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$$v\_w\_dist = \begin{cases} \min(eucl(V,W)), V \text{ for the velum}, W \text{ the pharyngeal wall} & \text{if no contact} \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$



(a) v_t_dist=7.62 (b) v_w_dist=5.10 (c) t_w_dist=10.05

Figure 4.17: v_t_dist, v_w_dist and t_w_dist are computed as the minimal distances between the respective articulators.

A complicating issue was the case when there was an articulatory contact: for example, the tongue and the velum touching (see Figure 4.18). Then I would only have access to the joint contour of both articulators, and there will be only one minimal distance between this joint articulator and the pharyngeal wall, instead of two: one for the distance between the velum and the pharyngeal wall, and the other between the tongue and the pharyngeal wall. In this case, I masked the image to have the distance to be computed only in the approximate expected area (in the case of the velum and the tongue in contact, the distance between the tongue and the pharyngeal wall needed to only be in the lower part of the image, and the one between the velum and the pharyngeal wall only in the upper part). Whenever the mask did not capture a single point of the contour, the distance was set to NaN ("not a number").

These two values bear the pieces of information we were searching for: how far the tongue is from the velum (to investigate their behavior when producing such sounds as rhotics) and how open the velopharyngeal port is (to measure nasality). However, there is one more constriction present in the window: the space between the back of the tongue and the pharyngeal wall. In the case of French, this place of articulation is not used consciously, but still is quite informative for the acoustics of the resulting sound, so I extracted it as well (see Figure 4.17).

$$t\_w\_dist = \begin{cases} \min(eucl(T,W)), T \text{ for the tongue}, W \text{ the pharyngeal wall} & \text{if no contact} \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

Additionally, the categorical information, whether there is a contact or not, was stored as

(a) v_w_dist=4.47 (b) t_w_dist=3.00

Figure 4.18: v_w_dist and t_w_dist computed in the case of the velum and the tongue in contact: the minimal distance is computed only in the area where the articulator is expected to be.

the following values:

$$v\_t\_cont = \begin{cases} 0 & \text{if no contact between the velum and the tongue} \\ 1 & \text{otherwise} \end{cases} \tag{4.9}$$

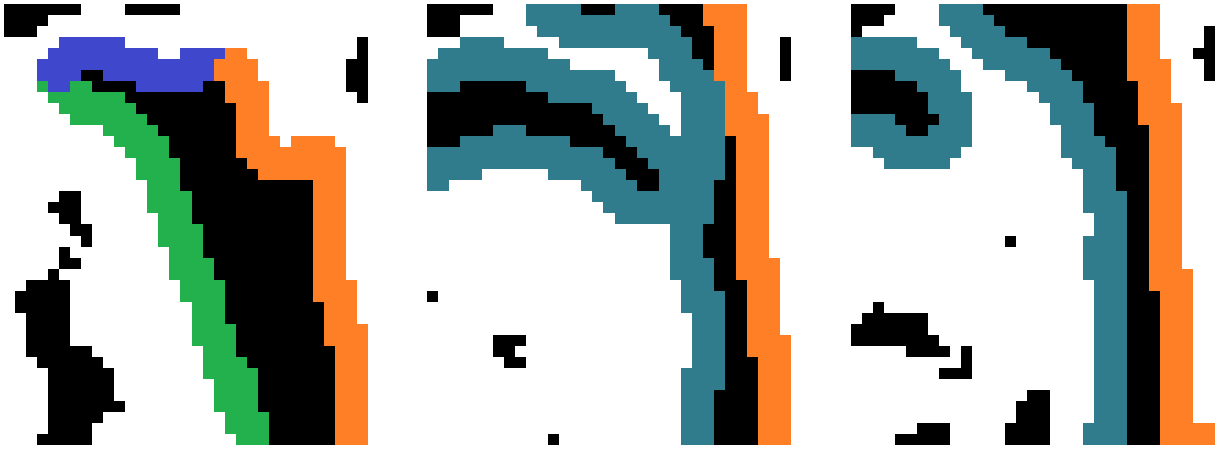
$$v\_w\_cont = \begin{cases} 0 & \text{if no contact between the velum and the pharyngeal wall} \\ 1 & \text{otherwise} \end{cases} \tag{4.10}$$

$$t\_w\_cont = \begin{cases} 0 & \text{if no contact between the tongue and the pharyngeal wall} \\ 1 & \text{otherwise} \end{cases} \tag{4.11}$$

The cont values could be derived without precise contour assignment, so those values could be set even when there was no dist values available.

It should be noted that while the articulatory shape changes gradually and smoothly, the function of minimum between two curves does not have to be smooth, since the points producing the minimal distance do not have stay close across neighboring images. It was less of an issue with the lips, since most of their mid-sagittal behavior can be explained in two dimensions: x for protrusion and y for opening. The lips do not curl and do not have any options for how to come into contact. This is not the case, however, with the more complex organs in the back of the vocal tract. For example, Figure 4.19 showcases three very different cases for how the tongue

and the velum can touch: either because of the tongue going up while the velum is straight and presses up at the pharyngeal wall (as in /k/), or because the velum lowers and at least slightly curls to touch the tongue while opening up (as in /ɔ̃/), or because the velum falls flat on the back of the tongue, and the surface of their contact vibrates in order to produce the rhotic /ʁ/.



(a) /k/: A contact at the expense of the tongue position, with the velum straight and pressing up the pharyngeal wall.

(b) /ɔ̃/: A contact at the expense of the lowered velum.

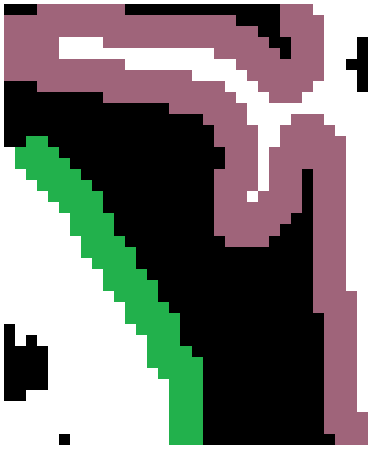
(c) /ʁ/: A turbulent rhotic contact, the velum is shown in an extensive contact with the tongue.

Figure 4.19: The tongue and the velum can come in contact in various ways, but these configurations will lead to the same values $\mathrm{t\_v\_dist} = 0, \mathrm{t\_v\_cont} = 1$. The blue-green color stands for joint contours of the velum and the tongue, green for the tongue alone, blue for the velum, orange for the pharyngeal wall.

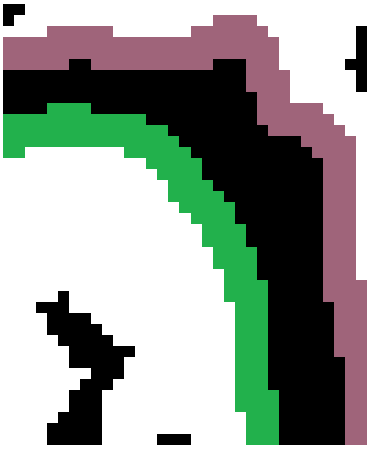
Furthermore, in the case of the lips we were calculating the surface of the contact between the upper and the lower lips, which gave us the possibility to differentiate between a fleeting contact and the lips being firmly pressed together. Since the same computations for the contacts that can occur in the window of the velum would be much more prone to errors, it was decided, as Equations 4.9–4.11 showed, to only store the categorical information about the presence or the absence of the contact (see Figure 4.20 to see how the two types of contact as treated as same); this effectively rules out any way to foresee any change in the $\mathrm{t\_v\_cont}$, $\mathrm{v\_w\_cont}$ and $\mathrm{t\_w\_cont}$ values: in one frame, they can be 1, and in the next one already 0, even though most likely these transitionary frames are not a clear case of either.

Creating articulatory parameter sequences

While having the big advantage of being fully automatic, this approach above has its shortcomings. The major weak point is the robustness of the algorithm. First, whenever there is an error in identifying a point of reference (such as the nose tip or a seed point) used to calculate one parameter or the other, all the subsequent calculations become incorrect. Even when the nose tip is found correctly, it still does not guarantee correct processing of the contours in the image, since contour labeling relies on treating numerous cases, and while the results show that most of the cases were treated, there is a possibility that some of them were not. This can cause dramatic errors down the pipeline: let us consider, for example, the following sequence



(a) A fleeting contact between the velum and the pharyngeal wall, about to disappear.



(b) Firm contact between the velum and the pharyngeal wall, completely blocking off air from entering the nasal cavity.

Figure 4.20: The articulatory parameters extracted from the window of the velum do not allow to track the difference between a well-planted contact between the articulators and a fleeting one: two examples of the velum touching the pharyngeal wall.

of v_w_dist values: $6.00, 6.00, 6.71, 6.09$. They are all quite consistent. If we add the next value that was calculated to be 18.00, it will create a dramatic effect in the estimation of the derivative and will be picked up by the machine learning algorithm.

A possible solution to that is to compare the dynamics of the parameter values in a fixed number of adjacent images and keep out those values that were probably wrong.

Thus, having extracted the values ls_dist (Equation 4.3), ls_cont (Equation 4.4), up_l_pr and lw_l_pr (Equation 4.5) from the window of the lips and the values v_t_dist (Equation 4.6), v_w_dist (Equation 4.7), t_w_dist (Equation 4.8) and v_t_cont (Equation 4.9), v_w_cont (Equation 4.10), t_w_cont (Equation 4.11) from the window of the velum, if the window of a given frame did not produce values consistent with the recent history, its group of parameters were replaced with NaNs.

The exact check for consistency was as follows. The size of the history was chosen through analyzing the intervals of erroneously processed images and concluding that it rarely was more than three images at once and having found no instances of more than four images. Thus, the values had to be consistent with their history of $hist = 4$ most recent images, which corresponds to looking back by approximately 73 ms and accounting for 91 ms in total. If there was not enough recent reliable history in the sequence yet, I checked whether the calculated parameter at least fell into the interval of its reasonable values. If yes, it could be stored, and if not, if it was the first value after a sequence of NaNs, it was replaced by the maximally acceptable value, and otherwise it was discarded.

When there were enough recent values, I could analyze the dynamics of their behavior. For this analysis to be informative, the history needed to be consistent: to qualify, all values could not contain any outliers, which means that they needed to stay close enough to their mean over the course of $hist$ frames. The following two paragraphs are going to quantify this "enough".

It was surmised that the information about the presence or absence of a contact was more reliable than the values calculated as a distance between two potentially mistaken curves. This led me to empirically set two thresholds for the difference between the current value and its mean over the course of the recent history: a weaker threshold, $T_w^l = 2.5$ for the lips and $T_w^v = 3.5$ for the velum, and a stronger threshold, $T_s^l = 3.75$ for the lips and $T_s^v = 6.3$ for the velum. The motivation for the exact values was brought by the range of values for a correct parameter sequence and how much it should be able to change over the fixed number hist of frames.

Thus, a chunk of hist recent parameter values was considered as reliable enough to serve as a ground for keeping or discarding the next value if all its present elements (i.e. all elements but NaNs) did not deviate from their mean value by more than T_w (T_w^l or T_w^v, depending on whether the articulator was studied in the window of the lips or of the velum).

Having access to a reliable average value over the course of the last hist frames, if the change between the average and the new value implied no change in how we considered the articulators in contact in the frame or not, it was compared to T_w, and if it did, to T_s. A change greater than the threshold value drove the new value out of consideration, and we marked it as implausible and discarded it.

Whenever a window had even a single value that was discarded as implausible, either due to too drastic a difference from the past average value or because of going beyond the expected limits, all other parameters extracted from this window were also replaced by NaNs, since it was likely that what was happening in the frame was processed incorrectly (incorrect seed points, incorrect labeling of the contours, etc.)

Since afterwards I was going to apply interpolation, I also needed to set some non-NaN values at the boundaries of the sequence. If the parameter values were increasing from the first non-NaN value in the sequence, I set the sequence to begin with the minimal meaningful value for this parameter; if the values were decreasing, with the maximal one. The case of the end of the sequence went in reverse: if at the end the values were decreasing, I added the minimal value, and if they were increasing, the maximal one. If there was no dynamic near the sequence boundary (stable values, no increase or decrease), the same non-NaN value was copied to the boundary.

After all these steps I finally obtained a sequence of articulatory parameters, a value or a NaN per each frame per each articulatory parameter. Keeping in only the values the algorithm was confident about assured that we would not deduce any erratic transitions from one frame to another. Thus I was able to remove NaNs and re-estimate them through upsampling the signal to match the acoustic rate of 200 Hz (a value once in 5 ms) that was necessary for parametric speech synthesis. Upsampling was done by piecewise 1-d monotonic cubic Hermite interpolation in order to have smooth transitions, the magnitude of each transition section bounded by its corresponding interpolation knots.

Verifying the articulatory parameters

In the subsequent parts of the work, the phonetic labels stayed associated to the extracted articulatory parameters. The labels were oftentimes assigned not precisely in agreement with what would be the judgment of a human annotator. Some of these imprecisions could still maintain the consistency between the label and the articulatory parameters, thus not affecting the interpretability of the results, and some of them could severely disrupt it. Having no ground truth and having to reason only with the calculated articulatory parameters, I could not

verify the first case; for the second, I carried out the following tests on those frames that could potentially create an inconsistency:

- For every phonetic label associated with necessarily open lips (vowels), if the lips never closed during the phoneme production, I labeled the instance as consistent, and if they did, as an error.
- For every phoneme that would be impossible to produce without closing the lips (bilabial stops and nasals /b, p, m/), if the lips closed at least once during the phoneme production, it was marked as consistent labeling, and if not, as an error.
- I extracted all instances marked as:
  - A rounded vowel, that is expected to be articulated with the corners of the lips drawn together and the lips protruded forward;
  - An unrounded vowel, where the lips are not protruded.

  The associated up_l_pr, lw_l_pr values were compared as samples for being different for vowels requiring protrusion and not (Shapiro-Wilk's, D'Agostino's and Anderson's tests for checking that the samples did not follow a normal distribution, Kolmogorov-Smirnov and Mann-Whitney tests to compare the samples). It would be expected that overall, the values for protruded vowels should be larger.
- For any phoneme requiring a contact between the tongue and the velum (/ʁ, k, g, ɲ/), if these two articulators came into contact at least once during the phoneme production, it was marked as consistent labeling, and if not, as an error.
- For any phoneme that would be impossible to produce with a contact between the tongue and the velum (the vowels and more frontal phonemes /j, ʒ, ʃ, ɥ, l, d, t, z, s, v, f, w, m, p/), if there was such contact (t_v_dist values equal to zero, positive t_v_cont values), it was marked as an error.
- For any nasal phoneme (a vowel or a consonant), if it was produced with a closed velopharyngeal port ($v\_w\_dist = 0, v\_w\_cont = 1$ — a contact between the velum and the pharyngeal wall), it was marked as an error.
- For any oral phoneme (a consonant or a vowel), if it was produced nasalized at any moment (without a contact between the velum and the pharyngeal wall: $v\_w\_dist > 0, v\_w\_cont = 0$), it was marked as an error.

This means that when the labels were found to be inconsistent, there were two possible error types: either an articulatory contact that should not have happened for this phoneme (this type is further referred to in figures as "unacceptable", for an unacceptable contact), or an absent contact that should have been there (referred to as "absent"). For every error, it was also important not only to mark its ratio in the entire distribution of consistent and inconsistent phonetic labels, but also to mark how many times it could have occurred, as an error occurring 25 times out of 26 times possible is very different from an error occurring 25 times out of 2500.

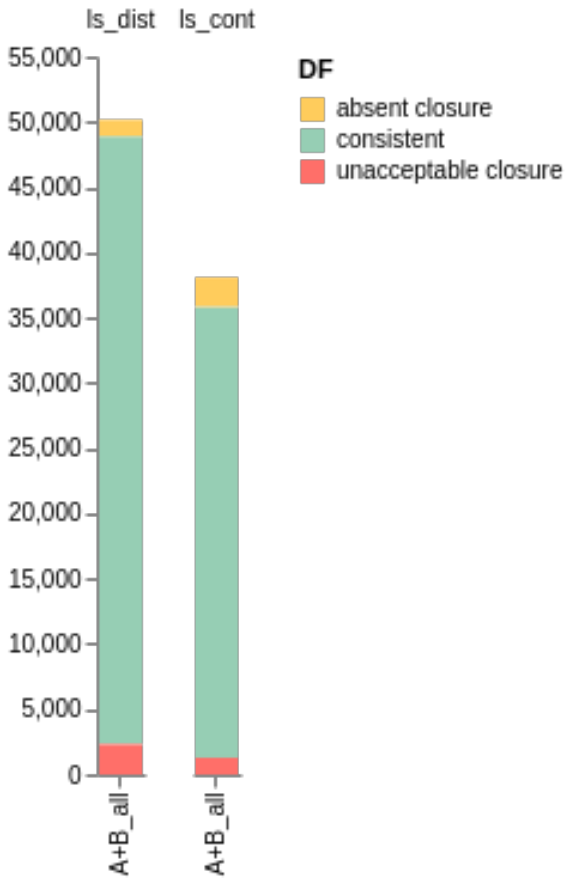
The results of all the checks above were aggregated by speakers (S_A, S_B or both) and by speech styles (spontaneous, not spontaneous or both).

The closure was consistent with the label 88.92% for S_A (when breaking down by spontaneous and non-spontaneous speech, 91.09% and 88.05% respectively), 92.52% for S_B (94.37% and 91.75%, respectively). In total it makes in 90.85% cases (92.87% and 90.02%).

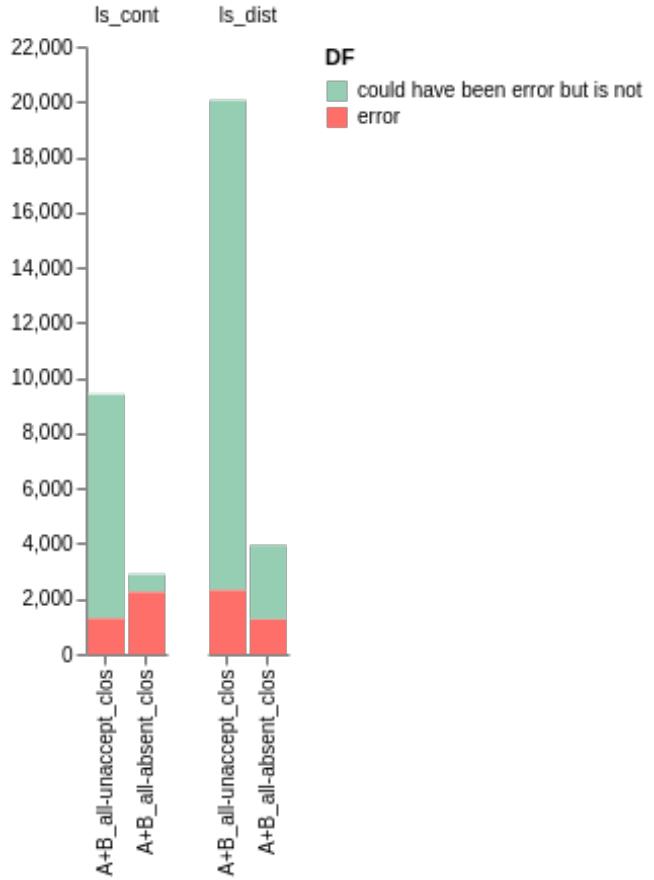
The majority of the errors, 53.21%, were the case of an absent contact between the articulators. That represents 29.31% of all cases where the presence of a contact was critical.

The rest of the errors, 46.79%, were the case of an extra closure during a phoneme that prohibit it. That represents 10.80% of the cases when this error could have occurred.

Figures from 4.23 to s4.24 visualize the possible comparisons.

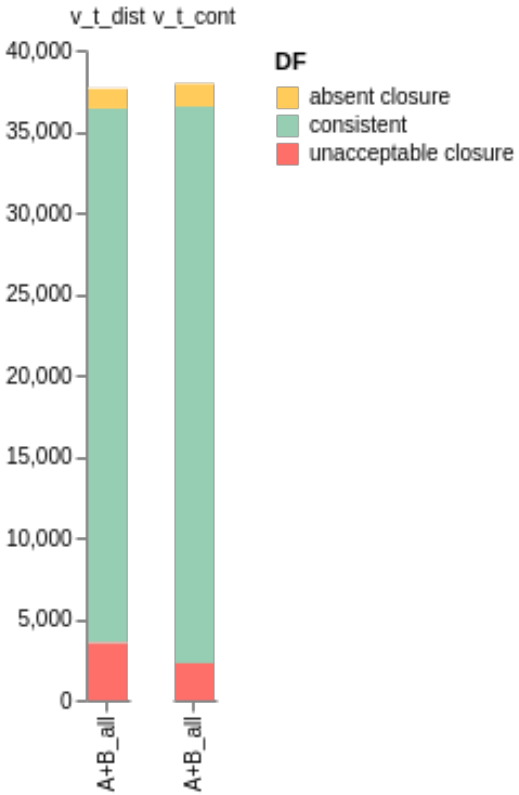


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. The majority of the errors are extra closures for ls_dist and absent closures for ls_cont.

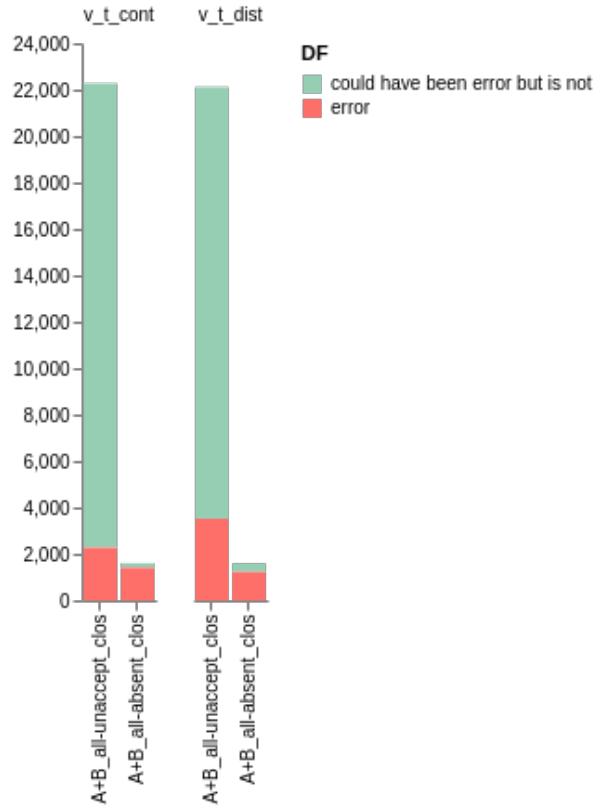


(b) ls_dist, ls_cont: absent and extra closure occurrences. Absent closures (no lip contact when supposedly producing /b, p, m/) occurred much more often with respect to how often they could have in both parameters.

Figure 4.21: Articulatory parameter consistency in the original corpus, overall—cont. on the next page.

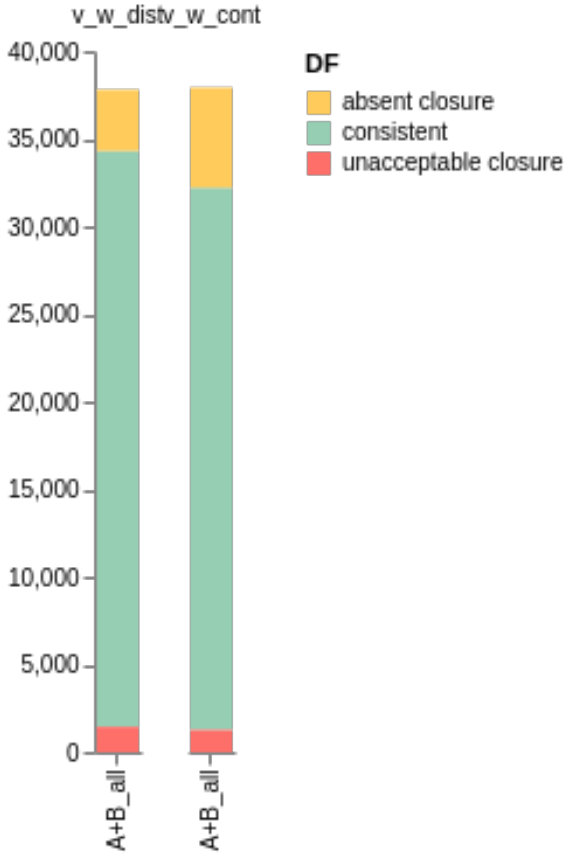


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. The error type of extra closures (e.g. a velum-tongue contact when producing /i/) is prevalent in both parameters.

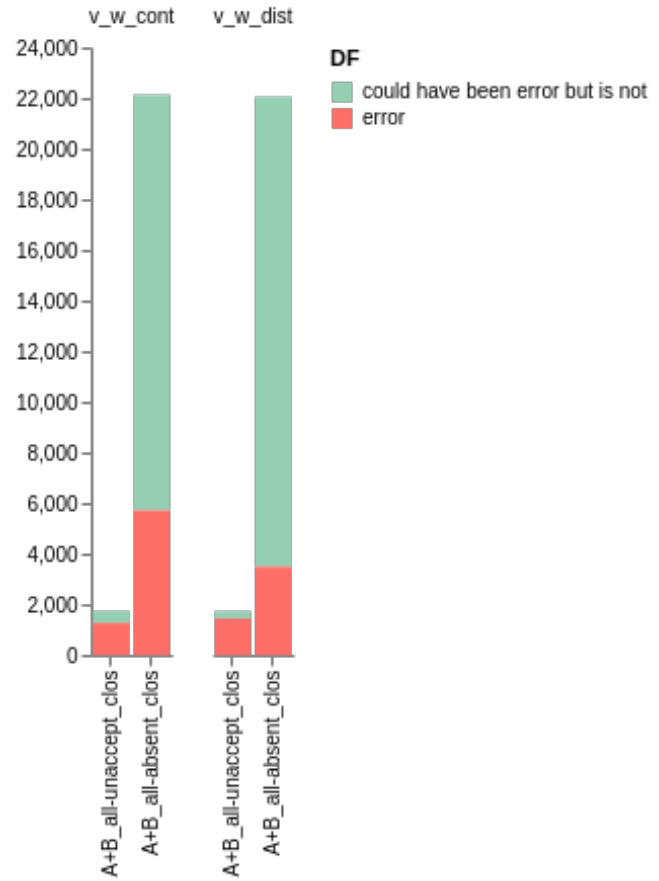


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. Despite a lower total count of absent closure errors, with respect to the number of times they were checked they occurred much more frequently than extra closures.

Figure 4.21: (Cont.) Articulatory parameter consistency in the original corpus, overall—cont. on the next page.

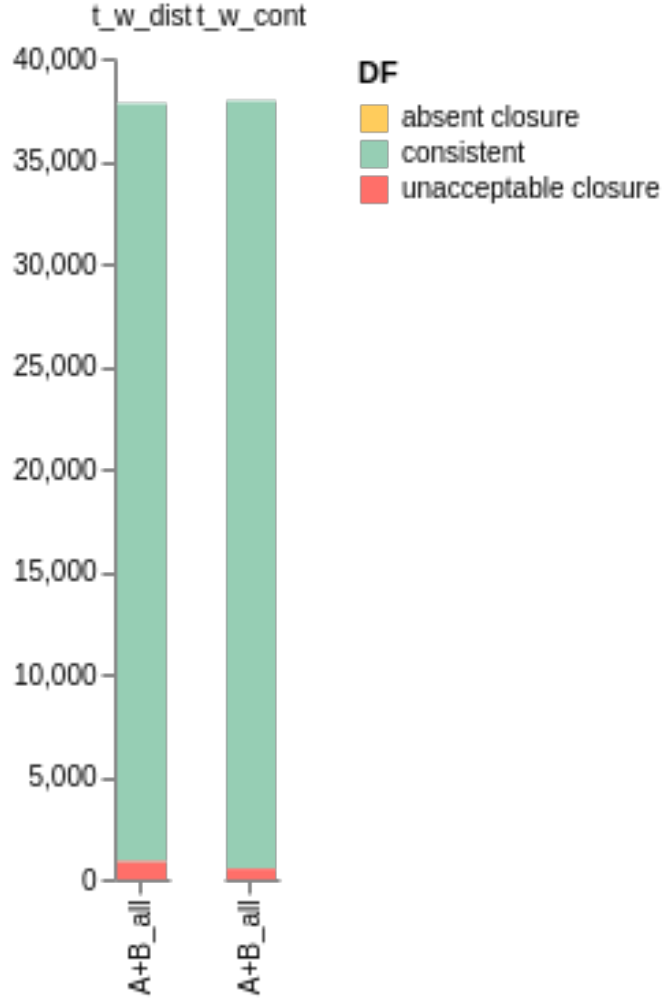


(e) v_w_dist, v_w_cont: consistent labels, absent closure and extra closure. This parameter pair has the lowest precision out of all: as low as 81.59% in the case of v_w_cont. Most errors are absent contacts (meaning that a sound would be produced nasalized rather than oral).

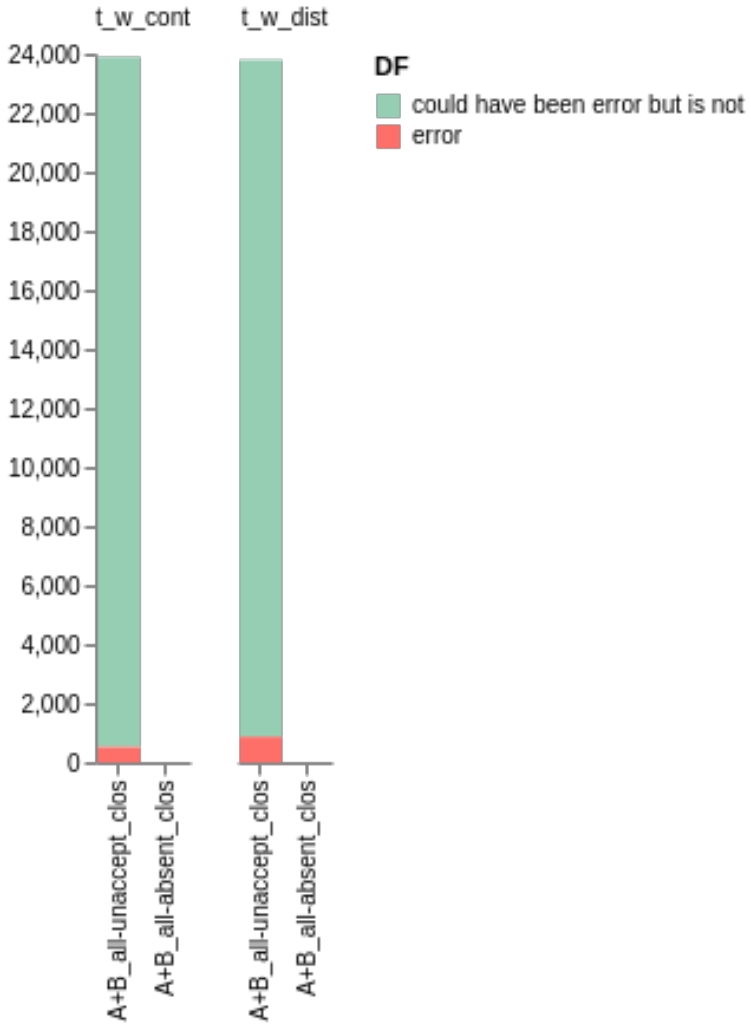


(f) v_w_dist, v_w_cont: absent and extra closure occurrences. Despite a lower total count of extra closures, with respect to the number of times they were checked they occurred much more frequently.

Figure 4.21: (Cont.) Articulatory parameter consistency in the original corpus, overall—cont. on the next page.

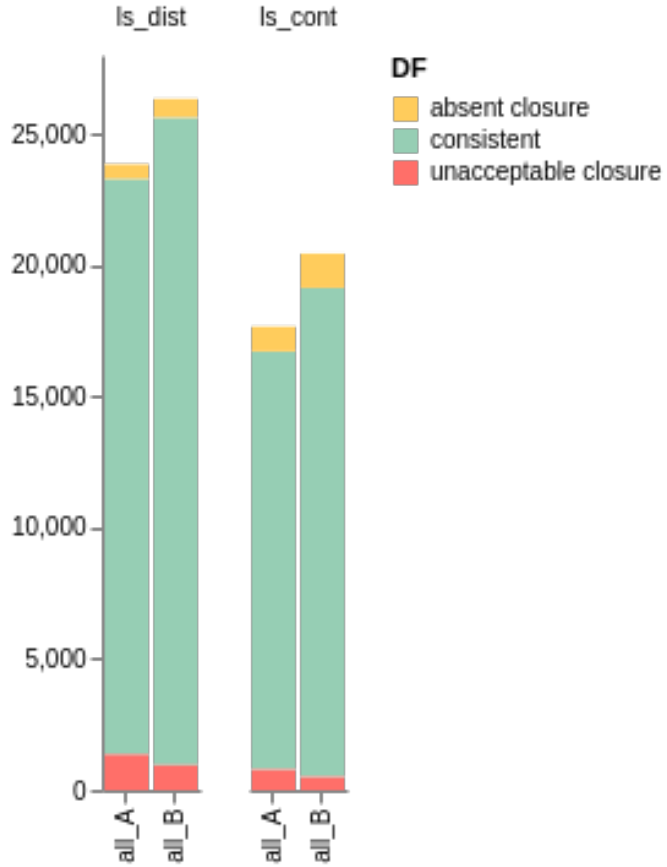


(g) t_w_dist, t_w_cont: consistent labels, absent closure and extra closure. No closure is necessary to produce any sound, so the only error type that was checked was extra closure.

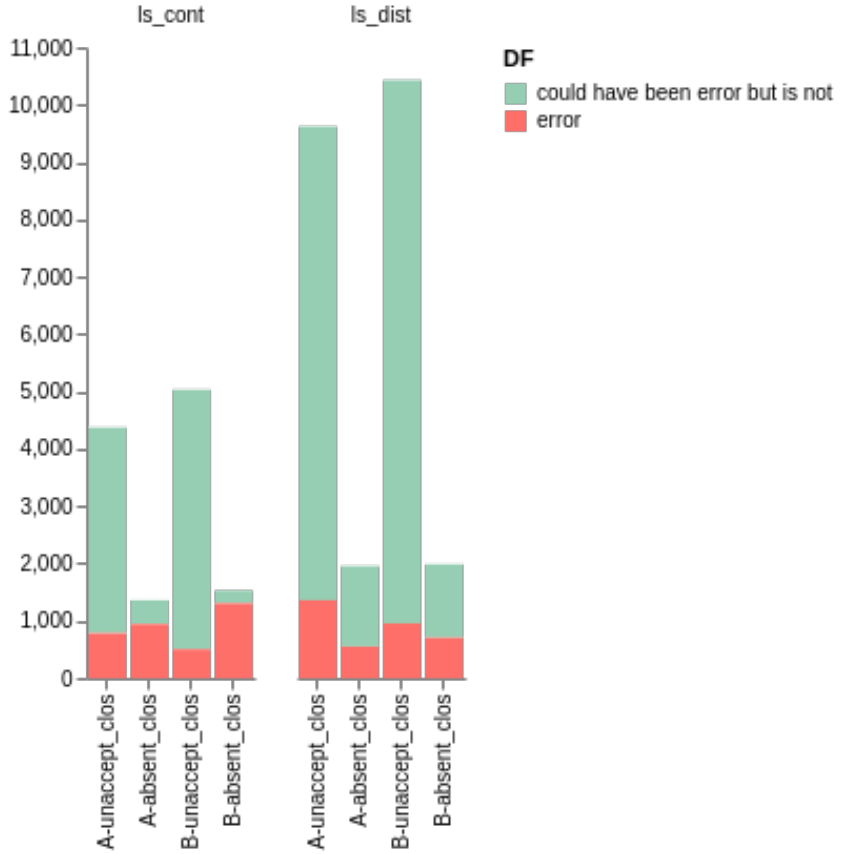


(h) t_w_dist, t_w_cont: absent and extra closure occurrences (no absent closure occurrences since it is not necessary to produce any sound).

Figure 4.21: (Cont.) Articulatory parameter consistency in the original corpus, overall

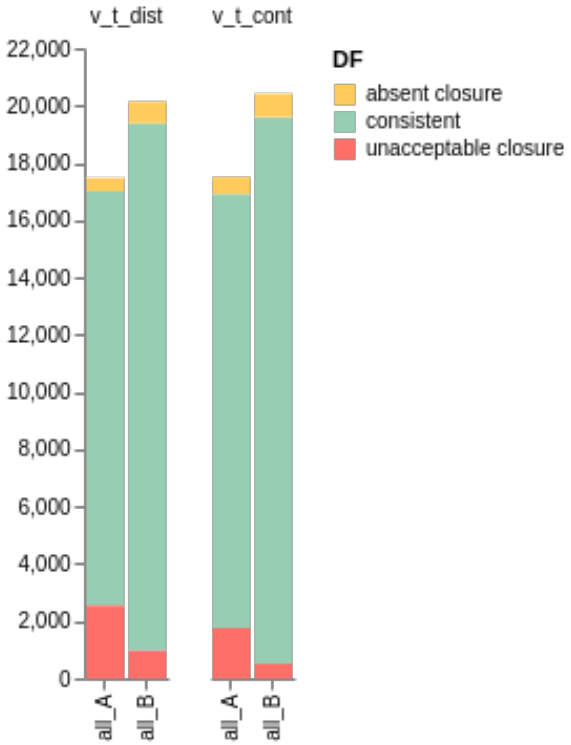


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. Precision is very similar for S_A and S_B. S_A is more likely to have extra closure errors than S_B is.

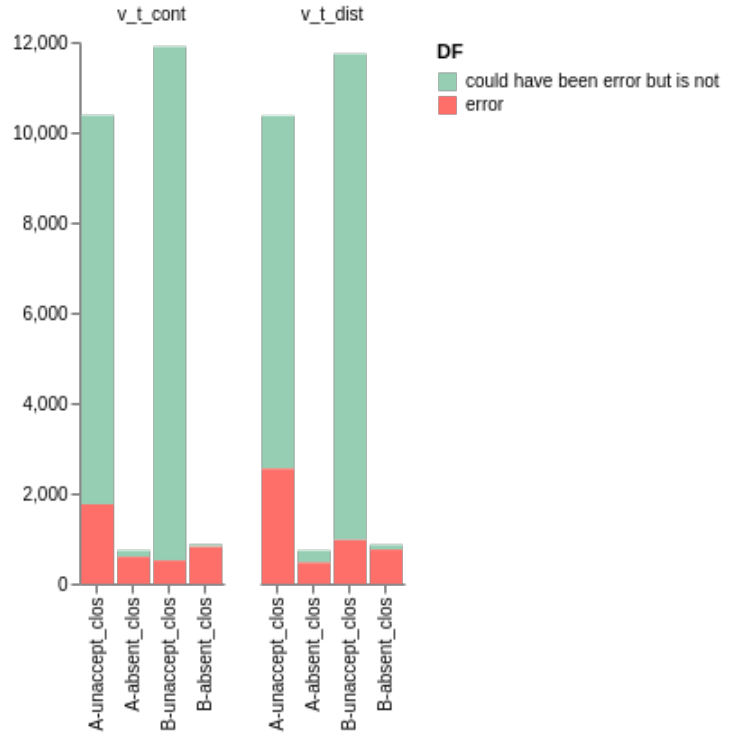


(b) ls_dist, ls_cont: absent and extra closure occurrences. S_A is more likely to have extra closure errors than S_B is.

Figure 4.22: Articulatory parameter consistency in the original corpus, broken down by speakers (cont. on the next page)

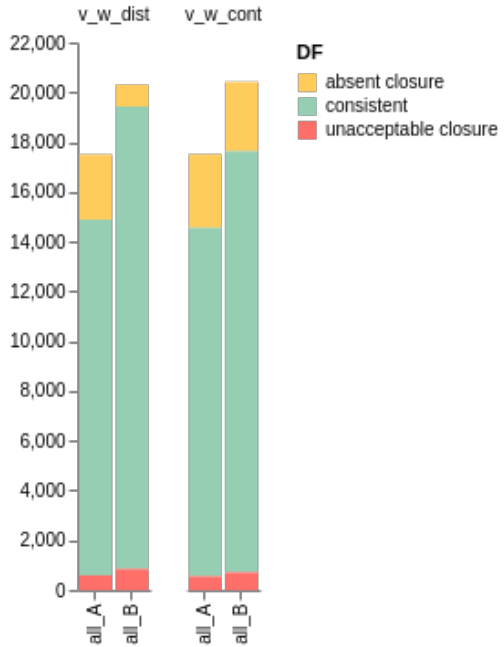


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. S_A is more likely to have extra closure errors than S_B is; extra closures are so prevalent that the overall precision for S_A considerably drops.

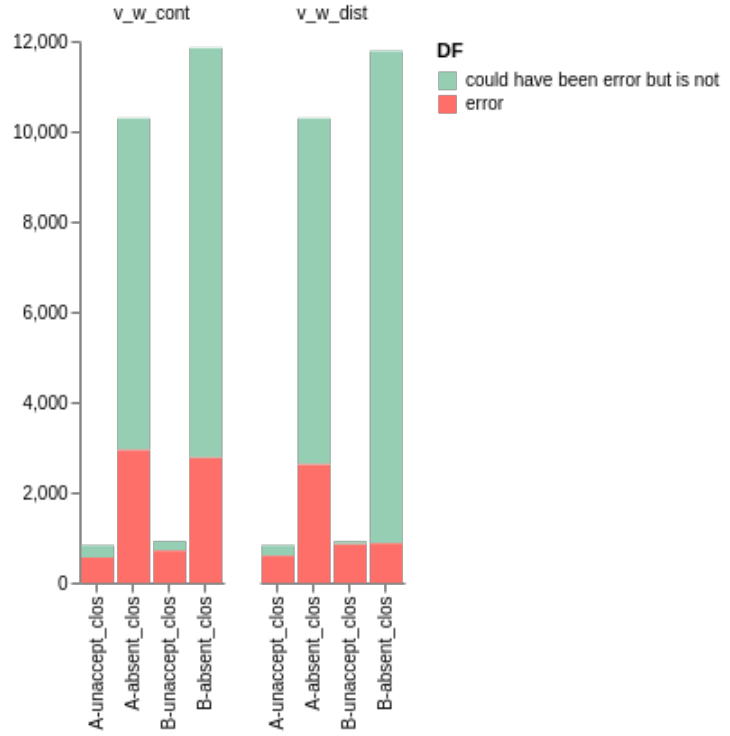


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. S_A is more likely to have extra closure errors than S_B is.

Figure 4.22: (Cont.) Articulatory parameter consistency in the original corpus, broken down by speakers (cont. on the next page)

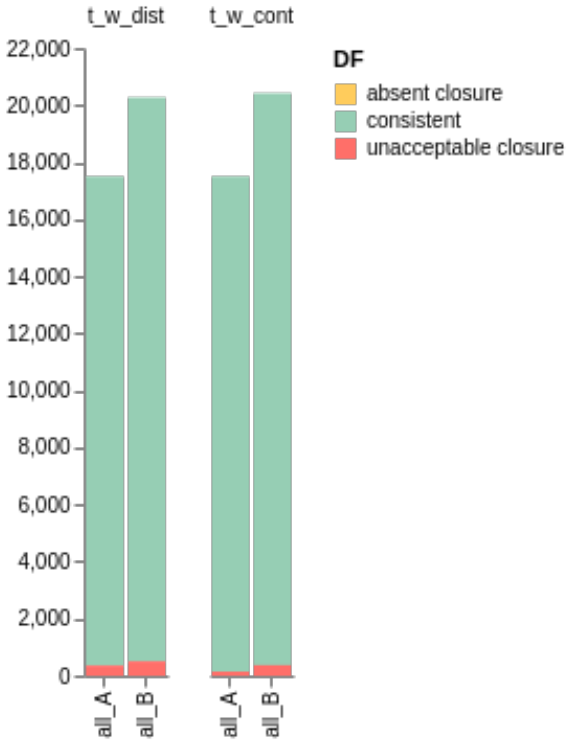


(e) v_w_dist, v_w_cont: consistent labels, absent closure and extra closure. In v_w_dist, S_A is more likely to have absent closure errors (i.e. to produce supposedly oral sounds nasalized) than S_B is. This error brings down S_A's precision. In v_w_cont, both speakers' precision suffers from an increase in absent closure errors.

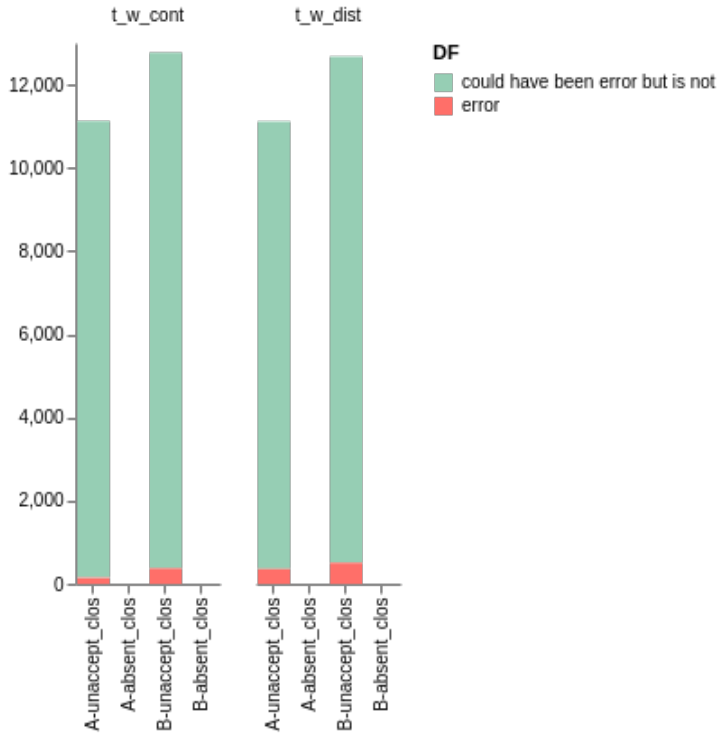


(f) t_v_dist, t_v_cont: absent and extra closure occurrences. Absent closure was a more considerable issue.

Figure 4.22: (Cont.)Articulatory parameter consistency in the original corpus, broken down by speakers (cont. on the next page)

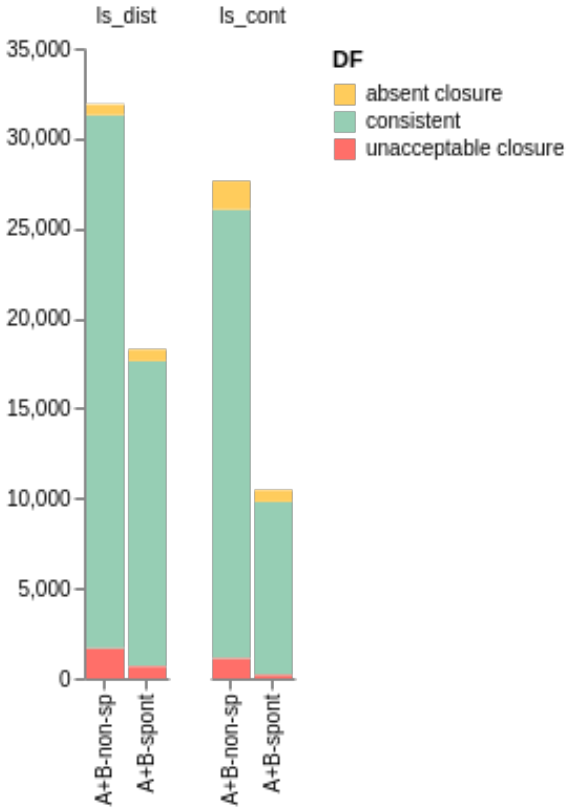


(g) t_w_dist, t_w_cont: consistent labels, absent closure and extra closure. The results are comparable for the two speakers.

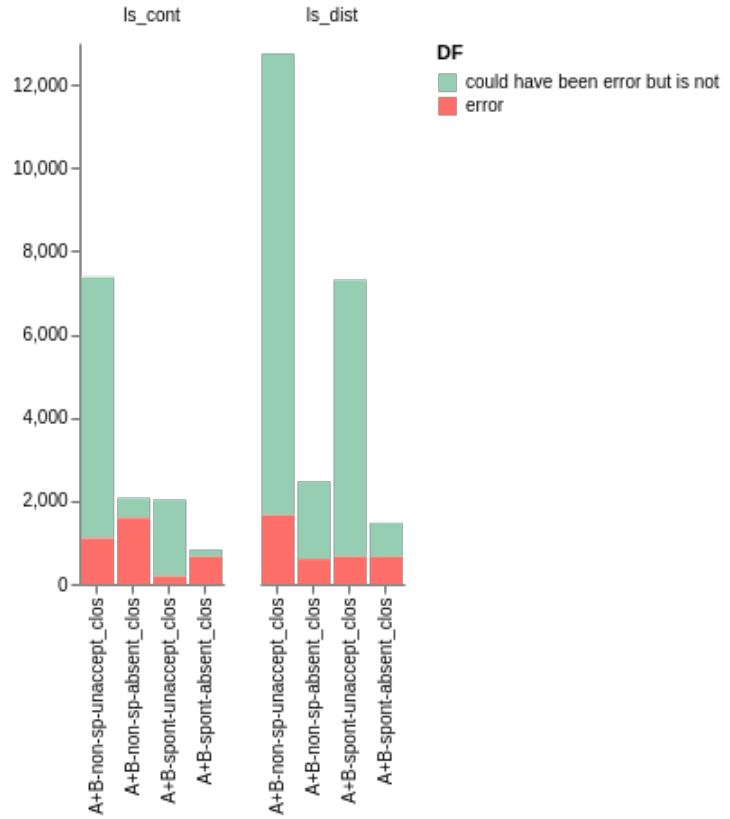


(h) t_w_dist, t_w_cont: extra closure occurrences

Figure 4.22: (Cont.) Articulatory parameter consistency in the original corpus, broken down by speakers.

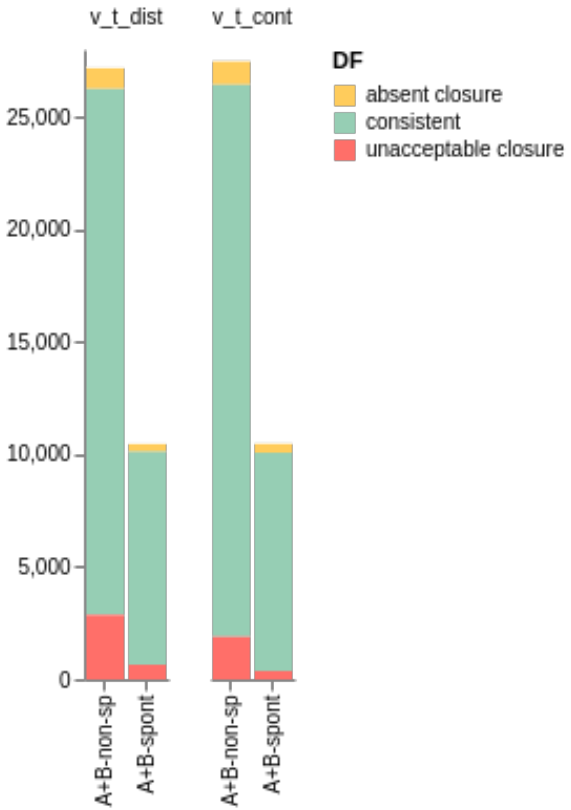


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. Non-spontaneous speech has a lower rate of absent closure errors, which could be related to a more careful articulation.

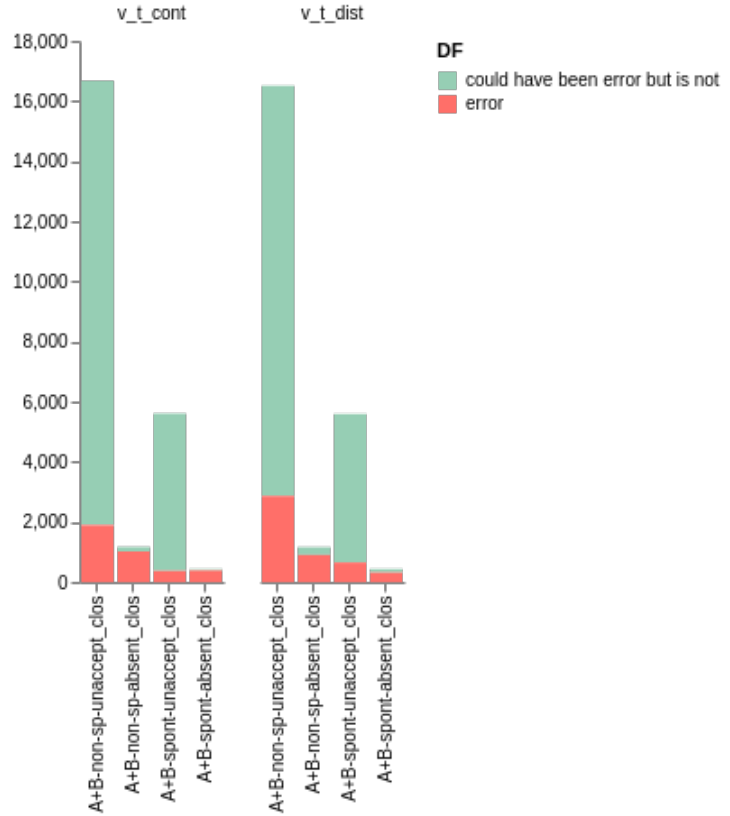


(b) ls_dist, ls_cont: absent and extra closure occurrences. Non-spontaneous speech has a lower rate of absent closure errors.

Figure 4.23: Articulatory parameter consistency in the original corpus, broken down by spontaneity—cont. on the next page.

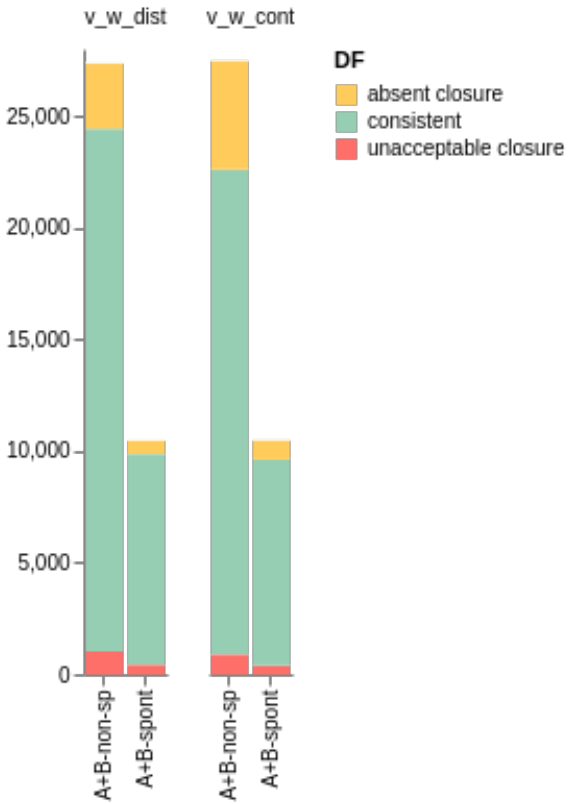


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. The total precision for non-spontaneous speech is lower, mostly because an increase in extra closures, probably due to the fact that in general it is more likely to capture this contact that spontaneous speech is.

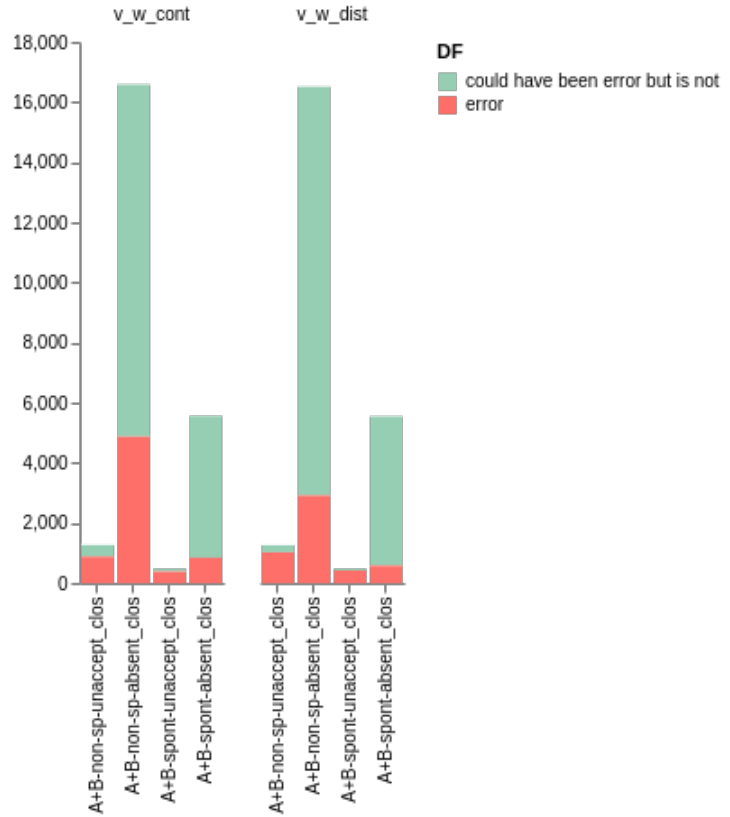


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. Non-spontaneous speech has a lower rate of absent closure errors, and spontaneous of extra closure ones, probably due to the fact that in general it struggles more at capturing this contact.

Figure 4.23: (Cont.) Articulatory parameter consistency in the original corpus, broken down by spontaneity—cont. on the next page.

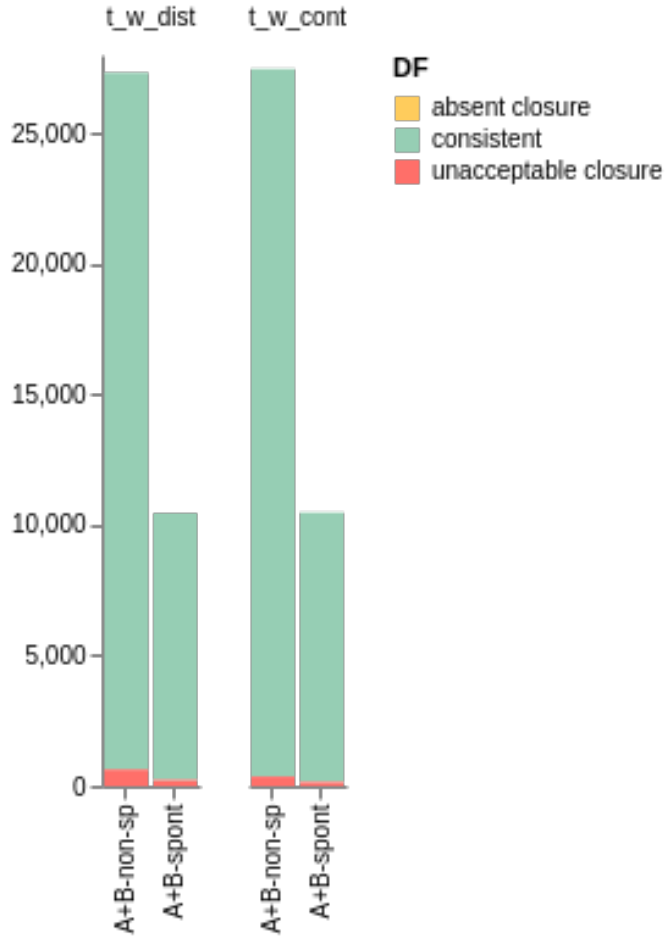


(e) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. Spontaneous speech has a greater precision due to the lower counts of absent closure.

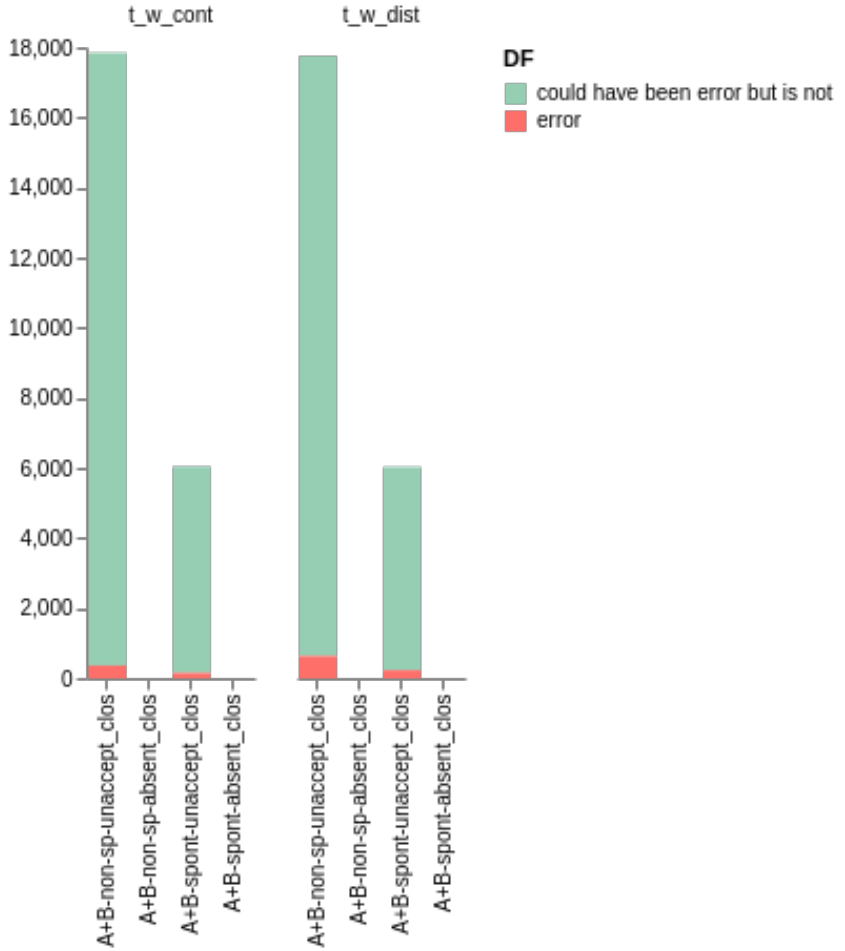


(f) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. Spontaneous speech fares better regarding absent closure.

Figure 4.23: (Cont.) Articulatory parameter consistency in the original corpus, broken down by spontaneity—cont. on the next page.

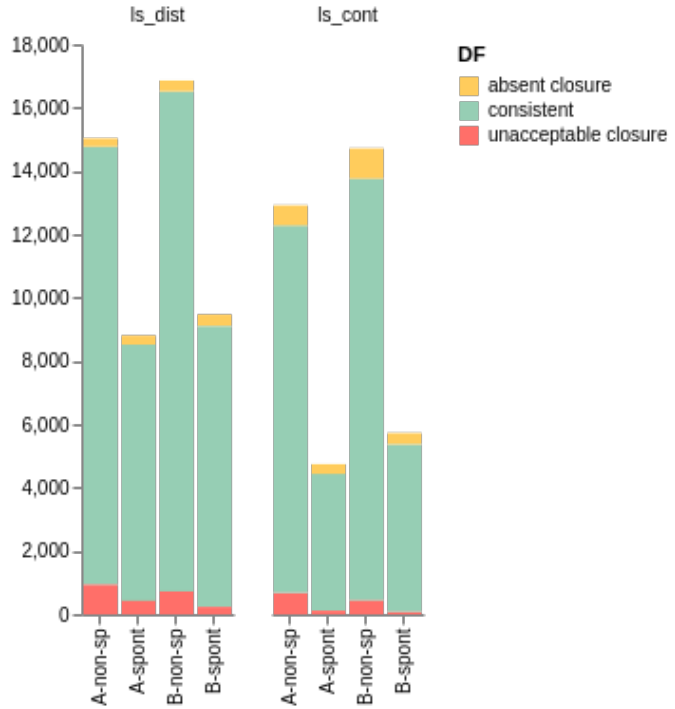


(g) t_w_dist, t_w_cont: consistent labels and extra closure. No significant difference between spontaneous speech and not.

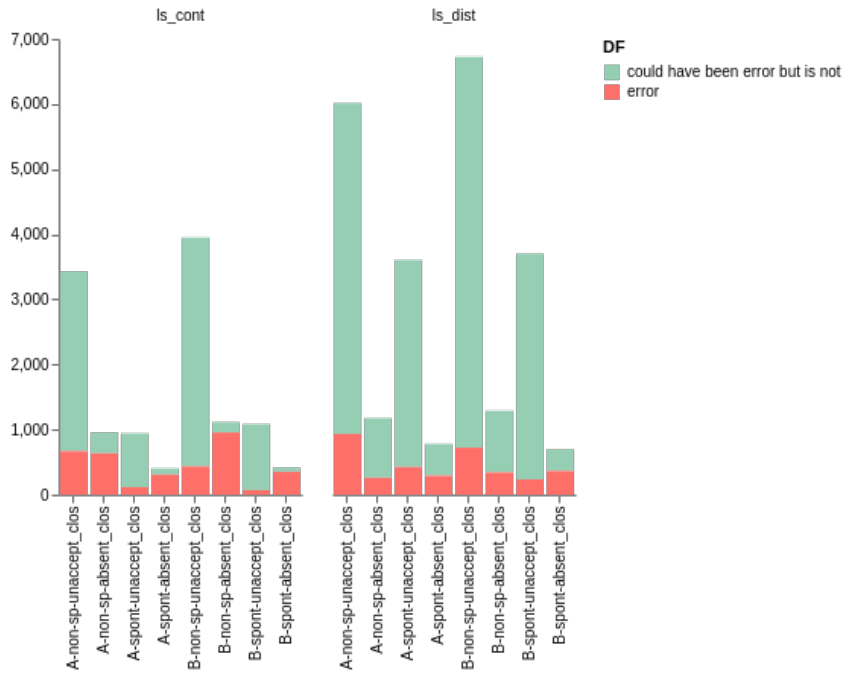


(h) t_w_dist, t_w_cont: extra closure occurrences. No significant difference between spontaneous speech and not.

Figure 4.23: (Cont.) Articulatory parameter consistency in the original corpus, broken down by spontaneity.

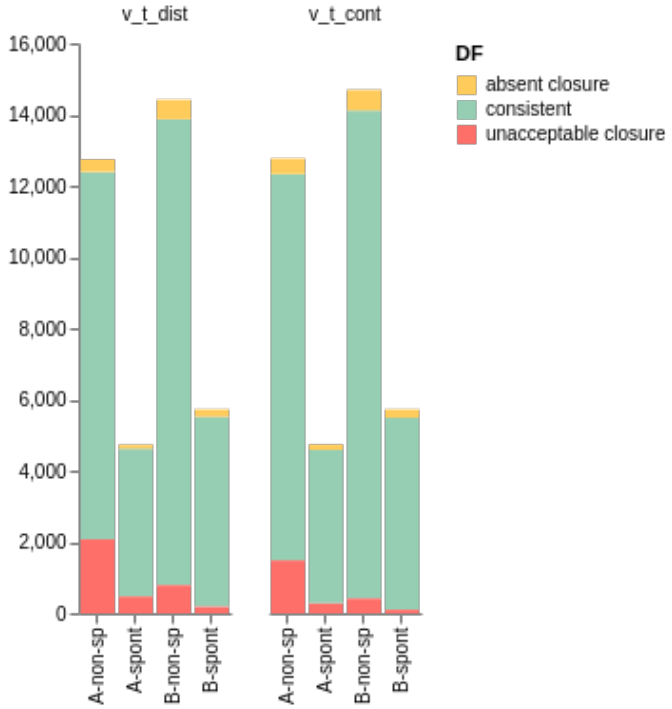


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. For extra closures, there is a greater impact of who the speaker is; for absent closures, whether it is spontaneous or non-spontaneous speech. Considering the total distribution of error types, precision is more defined by the kind of speech rather than by the identity of the speaker.

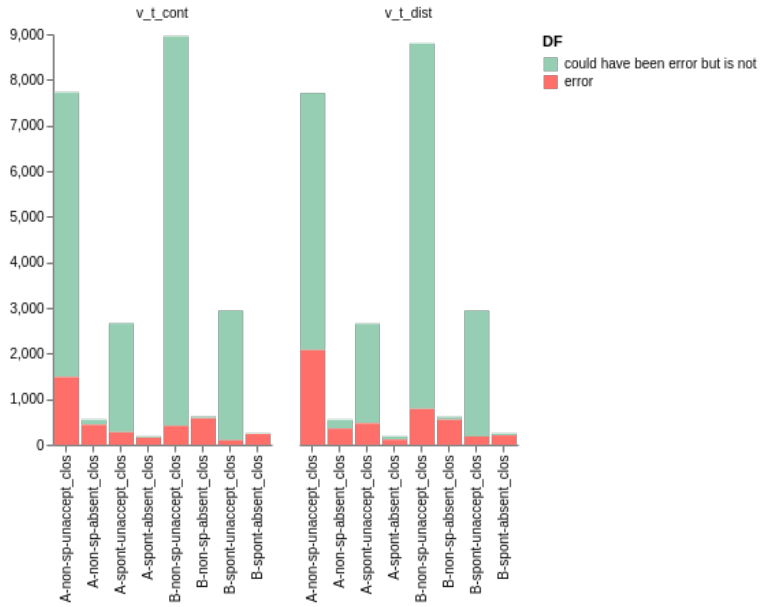


(b) ls_dist, ls_cont: absent and extra closure occurrences. Considering the number of times each error type was checked, for extra closures, there is a greater impact of who the speaker is, and for absent closures, whether it is spontaneous or non-spontaneous speech.

Figure 4.24: Articulatory parameter consistency in the original corpus, broken down by speakers and spontaneity—cont. on the next page.

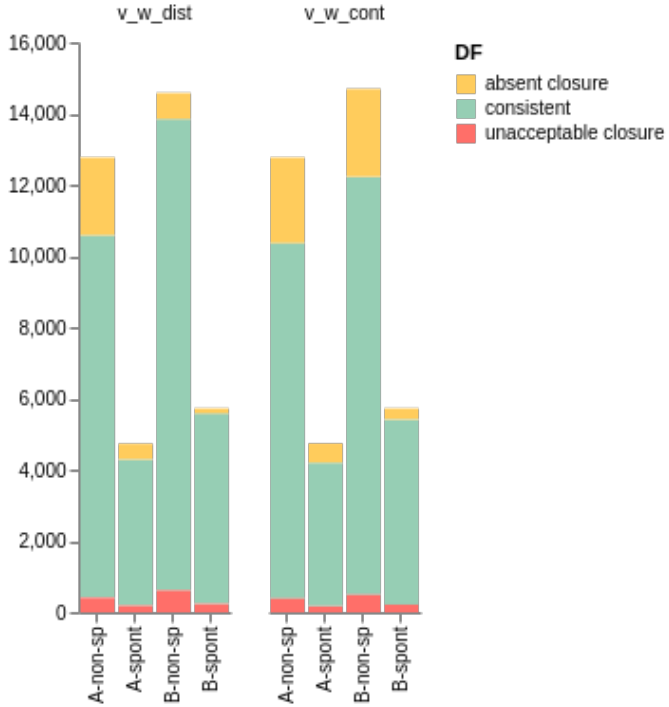


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. For both types of errors, the major defining factor is the identity of the speaker rather than the kind of speech.

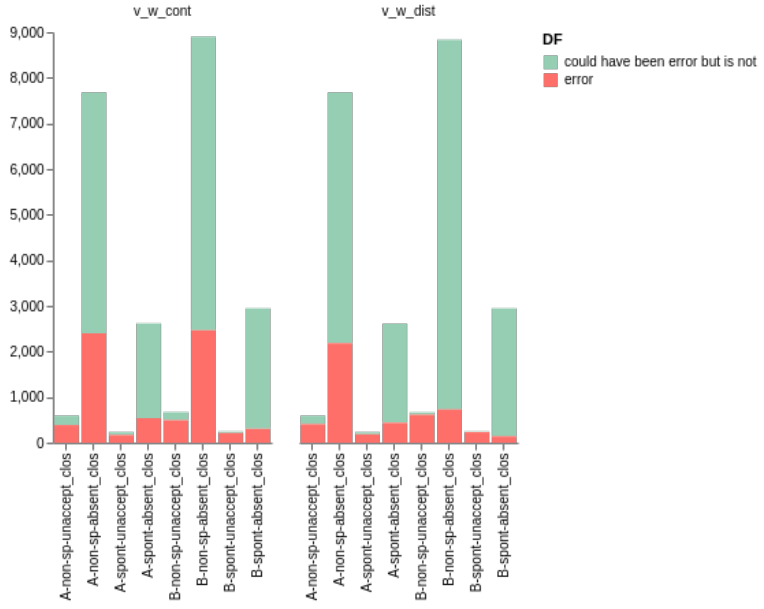


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. For both types of errors, the major defining factor is the identity of the speaker rather than the kind of speech.

Figure 4.24: (Cont.) Articulatory parameter consistency in the original corpus, broken down by speakers and spontaneity—cont. on the next page.

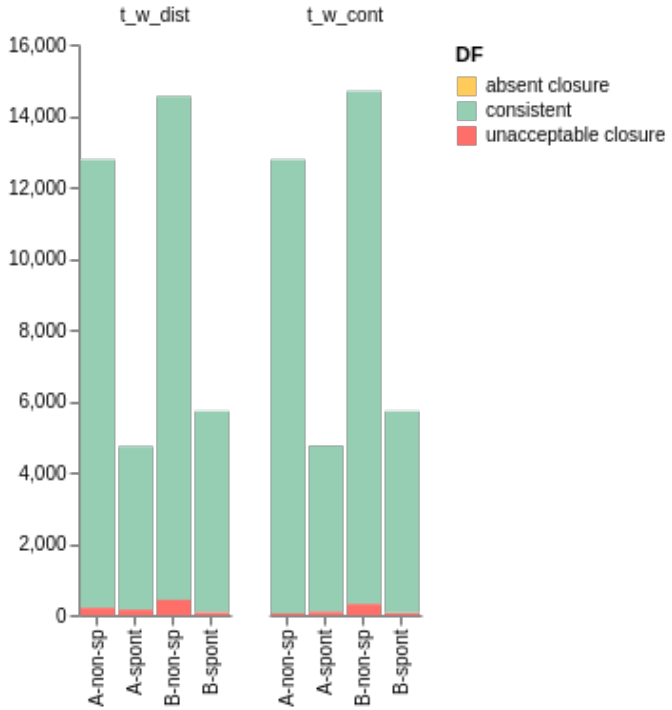


(e) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. In *v_w_dist*, for both types of errors and the resulting precision, the major defining factor is the identity of the speaker rather than the kind of speech. In *v_w_cont*, both kinds of errors are more defined by the kind of speech.

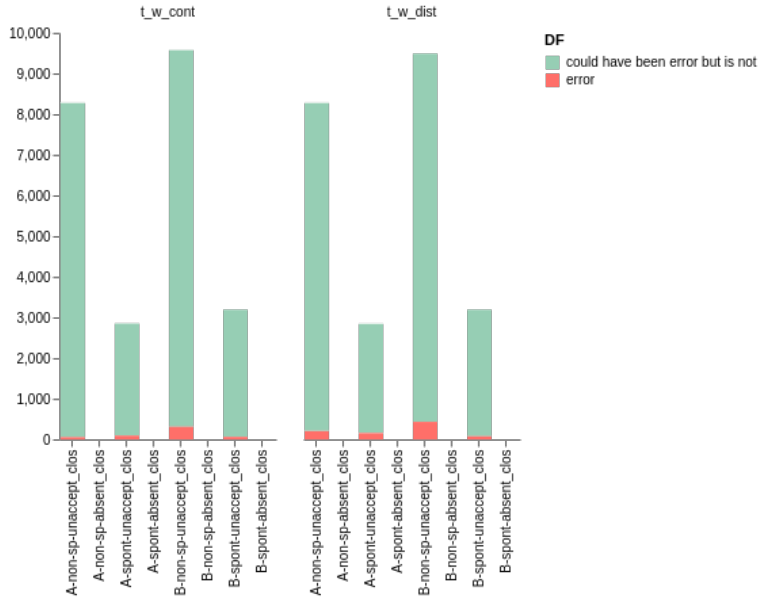


(f) *v_w_dist*, *v_w_cont*: absent and extra closure occurrences. In *v_w_dist*, for both types of errors, the major defining factor is the identity of the speaker rather than the kind of speech. In *v_w_cont*, both kinds of errors are more defined by the kind of speech.

Figure 4.24: (Cont.) Articulatory parameter consistency in the original corpus, broken down by speakers and spontaneity—cont. on the next page.



(g) t_w_dist, t_w_cont: consistent labels, absent closure and extra closure. No substantial differences across the speaker identities and the kind of speech produced.



(h) t_w_dist, t_w_cont: extra closure occurrences. No substantial differences across the speaker identities and the kind of speech produced.

Figure 4.24: (Cont.) Articulatory parameter consistency in the original corpus, broken down by speakers and spontaneity.

As Figures from 4.23 to 4.24 show, each pair of articulators (the lips, the tongue and the velum, the pharyngeal wall and the velum and the tongue and the pharyngeal wall) exhibit their own particularities.

ls_dist and ls_cont were the pair with one of the highest precision rates, 92.87% and 90.71%. While the majority of the errors were extra closures for ls_dist and absent closures for ls_cont, absent closures (no lip contact when supposedly producing /b, p, m/) occurred much more often with respect to how often they could have in both parameters. S_A was more likely to have extra closure errors than S_B was. Non-spontaneous speech had a lower rate of absent closure errors, which could be related to a more careful articulation. The last two points are also reflected in the conclusion when comparing the impact of the speaker identity (S_A or S_B) and the kind of speech (spontaneous or not): for extra closures, the impact was greater from who the speaker was; for absent closures, whether it was spontaneous or non-spontaneous speech. Considering the total distribution of error types, precision was more defined by the kind of speech rather than by the identity of the speaker.

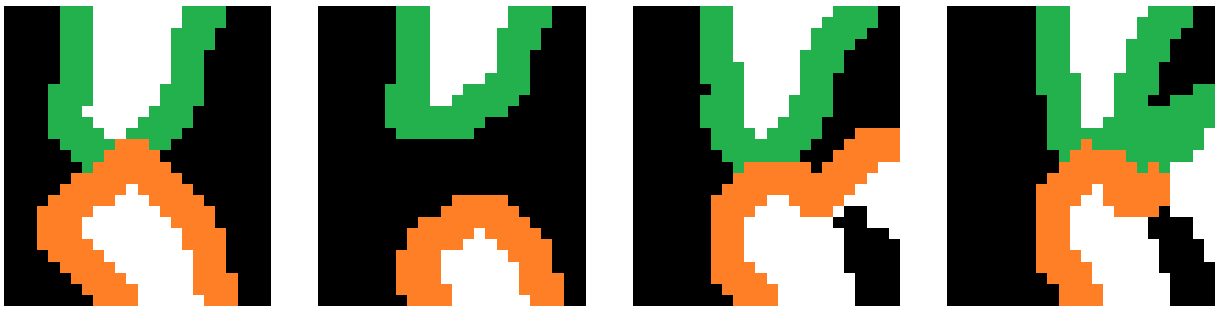
t_v_dist and t_v_cont had a slightly lower precision (87.39% and 90.28%). The rate of extra closures (i.e. having a velum-tongue contact when producing, for example, /i/) was prevalent as an error type in both parameters. However, it was related to the overall number of phonemes prohibiting such contact rather than to the frequency with which this error occurred with respect to how many times it was checked, since absent closure errors occurred much more often (76.61% and 87.80% of the cases when they could have occurred). S_A was more likely to have extra closure errors than S_B was; in fact, they were the reason for the drop in precision for S_A. Those extra closures were also the reason why the total precision for non-spontaneous speech was lower than that of spontaneous speech. The probable explanation is probably that in general spontaneous speech struggled more at capturing the contact between the tongue and the velum; hence, fewer extra closure errors in spontaneous speech and fewer absent closure errors in non-spontaneous speech. Anyway, the major defining factor for the distribution of consistent and wrong labels was the identity of the speaker rather than the kind of speech.

v_w_dist and v_w_cont exhibit the lowest precision out of all pairs (86.93% and 81.59%). Despite the fact that there were much fewer extra closures than absent contacts, with respect to the number of times that extra closures were checked they occurred much more frequently. In v_w_dist, S_A was more likely to have absent closure errors (i.e. to produce supposedly oral sounds nasalized) than S_B was, and this error brought down S_A's precision. In v_w_cont, both speakers' precision suffered from an increase in absent closure errors. Spontaneous speech had a greater precision than non-spontaneous due to the lower counts and a lower relative frequency of absent closure. Again, the probable explanation is probably that in general spontaneous speech struggled more at capturing the contact between these two articulators. As for the defining factor for the distribution of consistent labels and error types, there was a difference for the two parameters: in v_w_dist, for both types of errors and the resulting precision, the major defining factor is the identity of the speaker, and in v_w_cont, by the kind of speech.

t_w_dist and t_w_cont are a special case since no closure is necessary to produce any sound, so the only error type that was checked was extra closure, and precision went as high as 97.70% and 98.63%. There was no considerable difference in label consistency across speakers and spontaneity.

The analysis of error patterns showed the following problems:

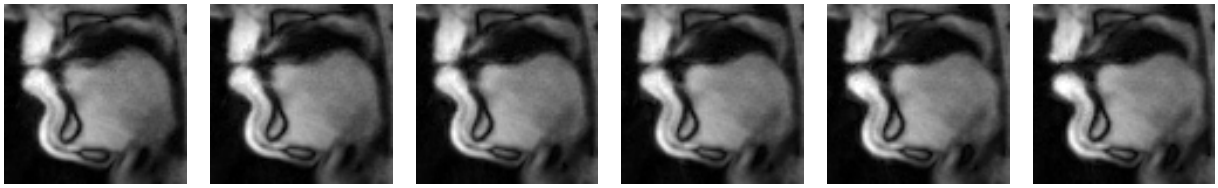
- Time shifts and other errors in labeling (missing, extra or wrong phonemes), commonly causing multiple errors in a single file. For example, if a $+$ means closure and $-$ means none, with a time shift a frame sequence labeled as /b, b, b, b, i, i, i, i, i, i, i, i, i/ could produce $----+++---------$, revealing that /b/ actually started only when the phonetic label already changed to /i/ (or this was no /bi/ in the first place);
- Incorrect identification of the lips:
  - Wrong nose tip identification due to noise being too strong in the image or due to accidental filtering out of the nose area, leading the algorithm to misplace the leftmost point recognized as the speaker's face. I improved this by correcting the nose tip values that get too far from the other ones that are observed in the sequence (for example, if all previous frames had the nose tip at around $(9, 20)$ and the next frame came with a nose tip at $(20, 19)$, I corrected it);
  - Choosing a wrong seed (a white pixel approximately where the lips, the velum, the tongue or the pharyngeal wall are expected to be) brought by a slightly misplaced window due to an unexpected speaker's movement or noise;
  - Other shapes getting involved: the tongue touching the lips (compromising the estimation whether the lips touch or not—see Figure 4.25 to get an idea for how the algorithm interpreted the upper and lower lip then), or in noise canceling, the lips getting accidentally merged with the area outside the vocal tract; mishandling the complicated geometry of the contours in the back of the vocal tract.
- When the contact was fleeting, the pixels where the articulators touched could stay too dark for the algorithm to pick up on their contact (see Figure 4.26 for an example of the lips).



The lips closed, /p/. | The lips open, preparing to speak. | The tongue touching the lips (1), silence | The tongue touching the lips (2), silence

Figure 4.25: Different cases for the annotation of lip contours, green for the upper lip and orange for the low lip.

As for the lip protrusion, both the upper and the lower lip were shown to have, in general, larger protrusion values up_l_protr and lw_l_protr on protruded vowels rather than on the non-protruded ones:



/s/, ls_dist = 1.0 | /m/, ls_dist = 1.0 | /m/, ls_dist = 1.0 | /m/, ls_dist = 3.16 | /ã/, $ls_dist = 4$ | /ã/, $ls_dist = 5$

Figure 4.26: The vocal-tract window of the sequence /smã/ (with only the ending of /s/ and the beginning of /ã/) and the respective extracted distance between the lips: a case when the contact between the lips is too fleeting to be recognized.

Sp.	Par.	$M_{protr}[par] \pm \sigma_{protr}[par]$	$M_{not\ protr}[par] \pm \sigma_{not\ protr}[par]$
S_A, non-sp	up_l_protr	13.55 ± 2.39	13.03 ± 2.22
S_A, spont	up_l_protr	13.69 ± 1.86	12.96 ± 1.62
S_B, non-sp	up_l_protr	11.50 ± 1.48	11.15 ± 1.42
S_B, spont	up_l_protr	12.64 ± 0.99	12.06 ± 1.00
S_A, non-sp	lw_l_protr	12.18 ± 2.22	11.60 ± 2.10
S_A, spont	lw_l_protr	12.62 ± 1.82	11.49 ± 1.46
S_B, non-sp	lw_l_protr	13.17 ± 1.88	12.71 ± 1.78
S_B, spont	lw_l_protr	13.61 ± 1.37	12.14 ± 1.38
S_A, all	up_l_protr	13.57 ± 2.32	13.01 ± 2.08
S_B, all	up_l_protr	11.62 ± 1.47	11.35 ± 1.39
S_A, all	lw_l_protr	12.24 ± 2.18	11.57 ± 1.95
S_B, all	lw_l_protr	13.22 ± 1.83	12.59 ± 1.72
S_{A+B}, non-sp	up_l_protr	12.41 ± 2.19	11.90 ± 2.00
S_{A+B}, spont	up_l_protr	13.15 ± 1.57	12.48 ± 1.40
S_{A+B}, non-s	lw_l_protr	12.73 ± 2.10	12.27 ± 1.99
S_{A+B}, spont	lw_l_protr	13.12 ± 1.68	11.84 ± 1.46
S_{A+B}, all	up_l_protr	12.48 ± 2.09	11.96 ± 1.95
S_{A+B}, all	lw_l_protr	12.78 ± 2.06	12.17 ± 1.88

Table 4.2: Lip protrusion is shown to be different on the vowels to require protrusion (larger up_l_protr, lw_l_protr values) and those that do not (smaller up_l_protr, lw_l_protr values).

4.2.2 Implementation

The speech synthesizer was built using Merlin as frontend [WWK16] based off their standard "build your own voice" recipe and using WORLD vocoder, for each speaker separately. After excluding the phrases where there was no phonetic labeling (due to bugs in elite HTS) or articulatory parameters (when, for example, the articulatory sequence was to be estimated based off too few samples that were labeled as a phoneme rather than silence or pause) available, I

obtained 618 utterances, 52.65 minutes of speech for speaker S_A and 917 utterances, 59.53 minutes of speech for speaker S_B.

The proportions of random division of the data for model training, validation and testing were 90%, 9% and 1% respectively.

All input parameters were normalized before being processed by the network.

Linguistic specification

To extract linguistic features from the labels, I identified potentially significant phonetic phenomena that were marked in the HTS labels:

- Specific phoneme;
- Groups of phonemes exhibiting allophonic variation, such as /e/ and /ɛ/ or /o/ and /ɔ/;
- Phonetic class—a feature per each position in the quinphone: vowel and its type (open, closed, etc.; front, back, etc.; rounded or unrounded); consonant and its type (voiced or unvoiced; stop, fricative, nasal, etc.; articulated at the front, center or the back of the vocal tract);
- Position of the phoneme in the syllable, of the syllable in the word, of the word in the phrase: an exact number as well as within a range;
- Number of phonemes in the syllable, of syllables in the word, of words in the phrase: an exact number as well as within a range;
- Stress and accent in the current syllable as well as in the neighbors;
- Distance to the stressed / accented syllables forwards and backwards;
- Part-of-speech tags: backwards, forwards and now.

Duration model

The duration model was a feed-forward 6-layer network, each layer with a hyperbolic tangent activation ($\tanh$). The choice of the activation function was done to allow the model parameters to update regularly and avoid the model parameters getting "stuck".

The batch size was 64, the learning rate was 0.002 with an exponential decay, 25 training epochs.

The output of the model is the value dur for duration.

Articulatory-acoustic and acoustic models

For comparison purposes, I made two setups: `full art` and `no art`: with articulation and without, respectively.

Acoustically a signal is characterized by a sequence of f0, the smooth spectral envelope and aperiodic energy. Then they produce the following parameters—both for `full art` and `no art`:

- The mgc parameters: Mel-Generalized Cepstral coefficients, extracted from the signal's spectral envelope, its dimensionality of 60, and the corresponding dmgc parameters estimating the derivative of mgc, the dimensionality of $3 \times 60 = 180$.

- The lf0 parameters: log of f0 values in cases where there is voicing identified, a fixed large negative value otherwise—its dimensionality of 1, and the corresponding dlf0 for the derivative, $3 \times 1 = 3$.

- The bap parameters: band aperiodicities, the aperiodic energy of the signal—its dimensionality of 1, and the corresponding dbap of 3.

- The vuv parameters: voiced or unvoiced, a boolean value (no derivative).

Acoustic parameters were treated by the WORLD vocoder.

Articulatorily (`full art` only) I enhanced the signal with the upsampled sequences of the articulatory parameters calculated above:

- ls_dist $\in [0, h_{window}]$ (before normalization),
- ls_cont $\in [0, w_{window}]$,
- up_l_protr $\in [0, w_{window}]$,
- lw_l_protr $\in [0, w_{window}]$,
- t_v_dist $\in [0, \sqrt{h^2_{window} + w^2_{window}}]$,
- t_v_cont $\in [0, 1]$,
- v_w_dist $\in [0, \sqrt{h^2_{window} + w^2_{window}}]$,
- v_w_cont $\in [0, 1]$,
- t_w_dist $\in [0, \sqrt{h^2_{window} + w^2_{window}}]$,
- t_w_cont $\in [0, 1]$.

Same as the duration model, the network setup started off the standard Merlin recipe: a feed-forward network with 6 layers, each layer with a hyperbolic tangent activation (tanh).

The batch size was larger than in the duration model, 256, since the acoustic search space constituted more parameters and therefore needed to benefit from a more accurate gradient computation. The learning rate was 0.002 with an exponential decay; the number of training epochs was set to 25.

The output of the network was a set of parameters: the acoustic parameters used to generate audio (mgc, lf0, bap) and all the articulatory parameters.

4.3 Evaluation

4.3.1 Evaluation components and criteria

The system is a joint articulatory and acoustic synthesizer; hence, it should be evaluated as such.

First, I would like to know how the adding of the articulation influenced the resulting quality of speech: intelligibility and naturalness. Without going into details there could be three scenarios: (a) the quality improves because articulation is helpful; (b) the quality reduces because articulatory information does not behave consistently with the acoustic and linguistic training data; (c) the quality stays the same. At a closer scrutiny, the results could be mixed: an improvement in some aspects while a deterioration in others.

Second, I needed to go deeper into the development/test evaluation and evaluate specifically the behavior of the articulatory parameters. The synthesized sequences should stay interpretable, which is not guaranteed considering the network does not have the knowledge about the meaning of each articulator's values, and they should not deviate far from what we observe in the corpus, which is not guaranteed either as the network could try to keep the cost function low by working primarily on the acoustic parameters and largely disregarding the articulatory ones.

4.3.2 Evaluation data and methods

Large-scale articulatory-acoustic evaluation

The first dataset was the rest of the [MCTO11] corpus: 281 sentences that were not included in the original corpus. They were synthesized both within `no art` and `full art` setups. Their duration models were naturally the same. What was different was their acoustic models, which in the case of `full art` was an articulatory-acoustic model.

First, the results were compared objectively: with MCD (mean mel-cepstrum distortion between the generated parameter sequence and the target one), BAP (band aperiodicity prediction error), and three measures for F0: RMSE (root mean square error), CORR (correlation coefficient) and V/UV (frame-level voiced/unvoiced error).

Then they were evaluated perceptually.

As for articulation, since `no art` did not treat it, it was not possible to compare it for the two setups. However, one of my evaluation criteria was to keep the generated articulatory parameter sequences interpretable, and the number of sentences was high enough to provide a meaningful comparison of parameter interpretability on the generated set to the one that was in the original corpus.

One-sentence-out articulatory evaluation

To evaluate specifically the behavior of the generated articulatory parameter sequences, I trained ten instances of the model with the same input data, but only one sentence taken out (i.e. all samples that contained this sentence). The sentences were chosen randomly according to the following criteria:

- It needed to have no fewer than 3 instances in the database of each speaker and no more than 10. Those instances were found as sentences with Jaccard similarity [Jac01] greater than 0.6;
- It could not have a disbalanced representation between the speakers: the numbers of its occurrences in the sets of S_A and S_B could not differ by more than 40%;
- It could not be shorter than the shortest sentence that made sense to analyze: "Il a pas mal" (thus ruling out, for example, /aβa/ and /ara/);
- It could not be only a part of an actual sentence, starting in the middle of it due to the imprecise corpus split into sentences or ending midway or with the speaker's hesitation.

Then the generated sequences were compared to the original ones with dynamic time warping, or DTW. Essentially it considers the source signal as one that needs to be morphed into the target signal at the lowest expense. Given a graph of pattern values over time, it uses dynamic programming to find the shortest path in this graph. The final cost of this path can serve as a measure of similarity between two temporal patterns.

Finally, I compared the rates of interpretability with respect to the phonetic label of the original and the generated sequences.

4.3.3 Evaluation results

Synthesis example

Let us consider the synthesis of "Bonjour" (*Fr.* "Hello"). Its phonetized form is /bɔ̃ʒuʁ/. Figure 4.27 shows the spectrogram of the generated signal with the voice of S_A.

This sound was generated accompanied by the sequences of each of the articulatory parameters.

Lip opening and contact parameters, ls_dist and ls_cont (Figure 4.28), show that the lips correctly closed to produce a /b/ (closure is marked with red) and correctly remained open throughout the rest of the utterance.

Lip protrusion parameters up_l_protr and lw_l_protr (Figure 4.29) were generated with the lips being more protruded on the vowels /ɔ̃/ and /u/ and their neighborhood, which is correct.

Tongue-velum distance and contact parameters t_v_dist and t_v_cont (Figure 4.30) correctly shows, as the velum lowers to produce the nasal /ɔ̃/, the decrease in the tongue-velum distance; there is no contact between the velum and the tongue to produce the rhotic /ʁ/, but this is something we observe in the original data as well (Figure 4.31), since the algorithm recognizes the space created by the velar and uvular vibration as an absence of contact. Furthermore, the increase of the t_v_cont values right before the end of the utterance could indicate that the problem was us getting too open a distance at /u/ and having no time to close it for /ʁ/ without a dramatic effect on the derivative.

Velum-pharyngeal wall distance and contact parameters v_w_dist and v_w_cont (Figure 4.32) generally exhibit a correct division between oral and nasal sounds over the course of all phonemes but /b/ in the case of v_w_dist: in this configuration, /b/ would be produced as

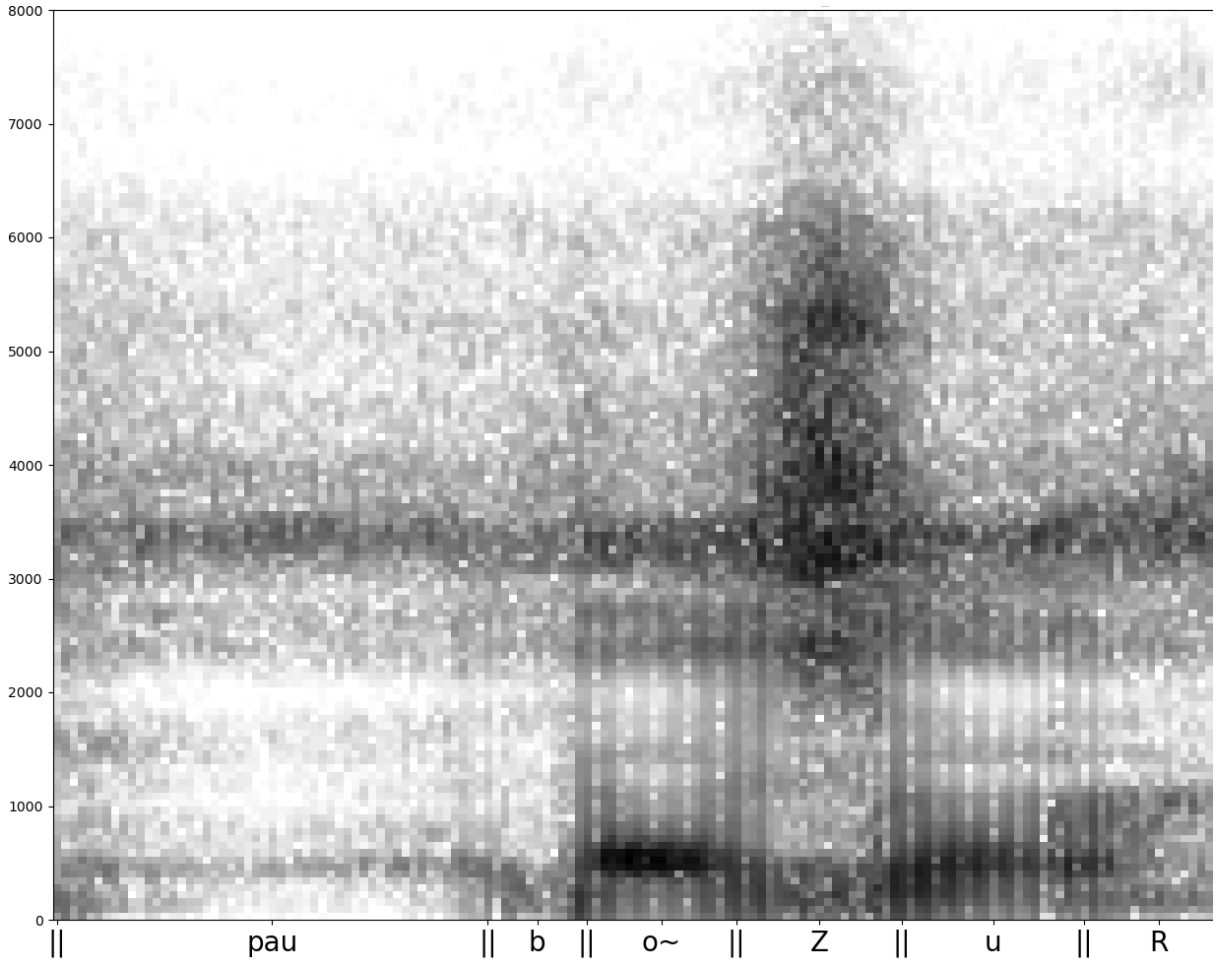


Figure 4.27: The `full art` synthesis of "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A. The transition of formants corresponds to the change of phonemes in production.

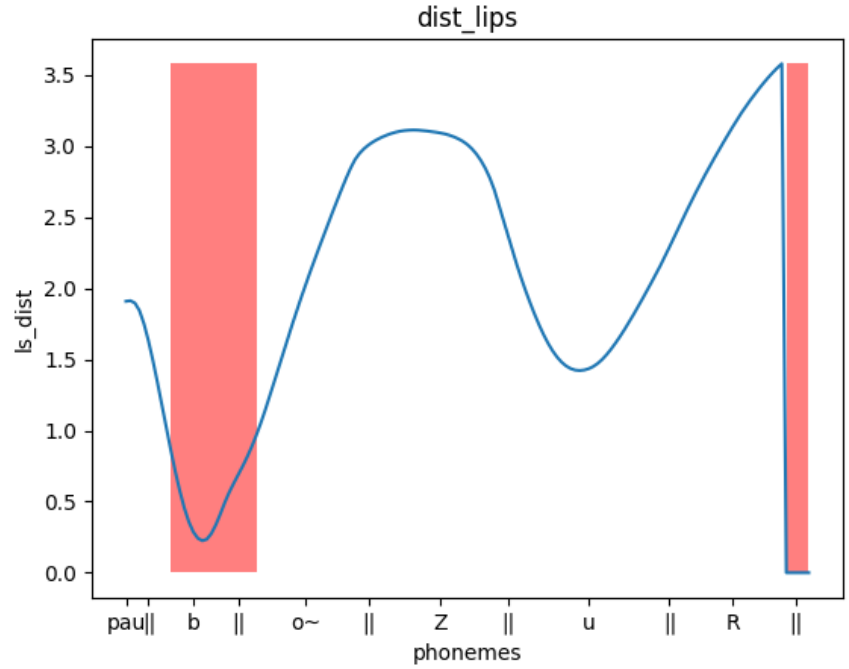
/m/. This should be caused by the same effect as in t_v_dist and t_v_cont: the silent position for the velum is to be lowered, and the velum was generated to have such a high v_w_dist that the value did not have the time to get as low as necessary for /b/ without affecting the derivative (while it should be noted that the decrease from *pau* to /b/ is quite steep), and then it was already time to prepare for the nasal /ɔ̃/ coming next. Otherwise, both /b/ and the final phonemes correctly register as oral sounds.

Tongue-pharyngeal wall distance and contact parameters t_w_dist and t_w_cont (Figure 4.33) show a correct absence of the contact throughout the entire sequence.

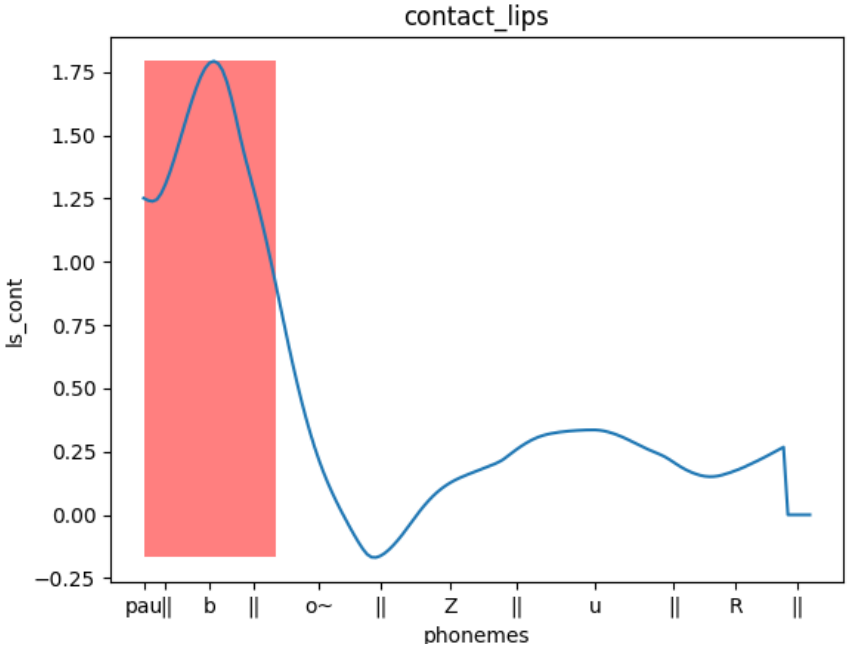
Large-scale articulatory-acoustic evaluation

As mentioned above, two setups, `no art` and `full art`, were trained. They were objectively evaluated on their common part, acoustics—see Table 4.3.

For the reference, the duration model's evaluation (in both setups) was as follows:



(a) ls_dist



(b) ls_cont

Figure 4.28: The synthesized sequences of ls_dist and ls_cont for "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A. The lip closure (in red) is consistent with the production of the labial stop /b/ and the narrowed labial opening for /u/ and with the absence of labial contact throughout the rest of the utterance.

- S_A:

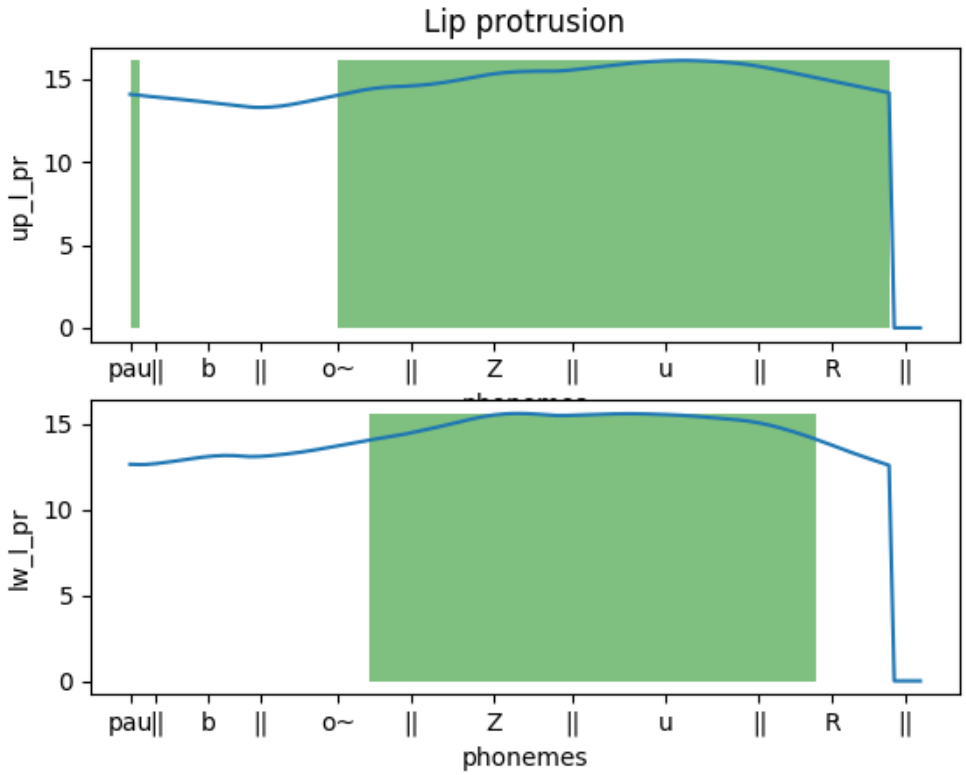
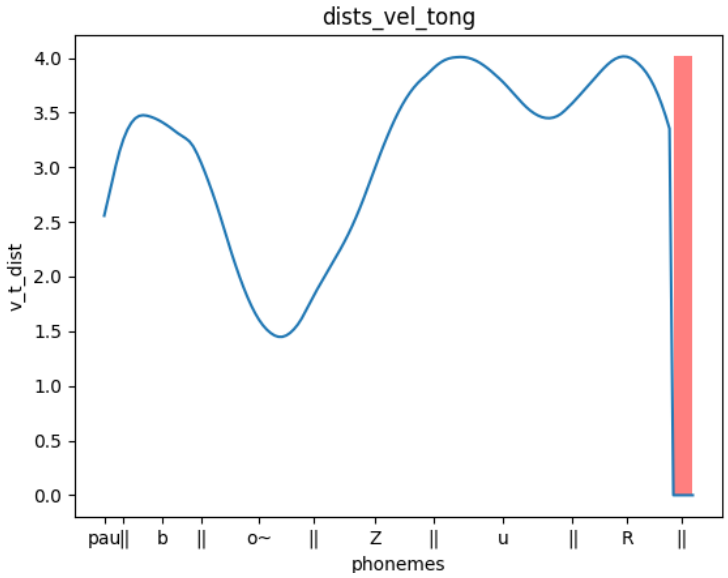


Figure 4.29: The synthesized sequences of ls_dist and ls_cont for "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A. Values within the range associated with protrusion are marked green. Indeed, both /ɔ̃/ and /u/ require lip protrusion, and it should be anticipated.

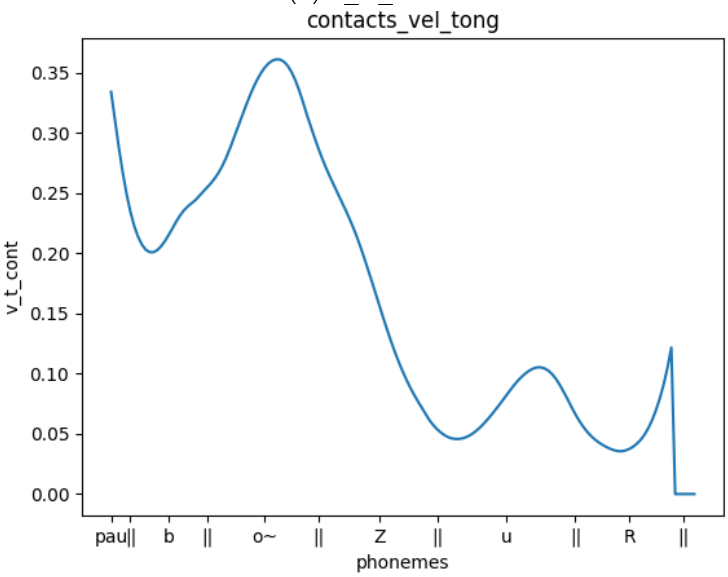
  - Development: RMSE: 24.110 frames/phoneme, CORR: 0.738;
  - Test: RMSE: 26.476 frames/phoneme, CORR: 0.676;

- S_B:
  - Development: RMSE: 34.212 frames/phoneme, CORR: 0.585;
  - Test: RMSE: 19.195 frames/phoneme, CORR: 0.580.

As for the interpretability of the generated labels, it generally follows that of the corpus. The major issue for all of the articulators is attaining a contact, thus raising the absent closure error counts and frequency. Figures 4.34 and 4.35 show the distributions.

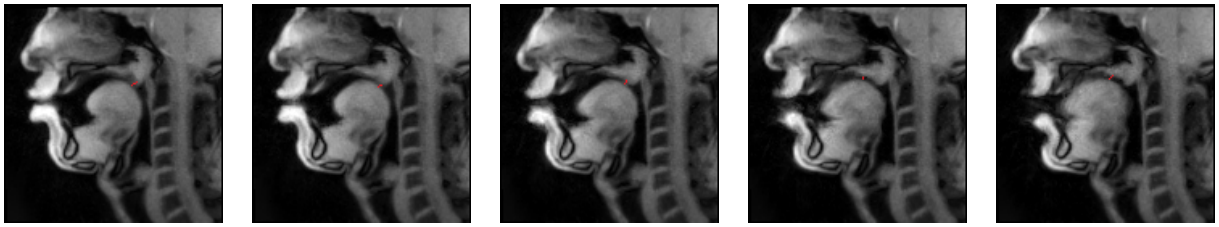


(a) t_v_dist



(b) t_v_cont

Figure 4.30: The synthesized sequences of t_v_dist and t_v_cont for "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A. As the velum lowers to produce the nasal /ɔ̃/, the tongue-velum distance correctly decreases; the behavior of the velum and the tongue for the rhotic /ʁ/ follows that of the original data, and, furthermore, the increase of the t_v_cont values right before the end of the utterance could indicate that the problem was getting too open a distance at /u/ and having no time to close it when /ʁ/ without a dramatic effect on the derivative.

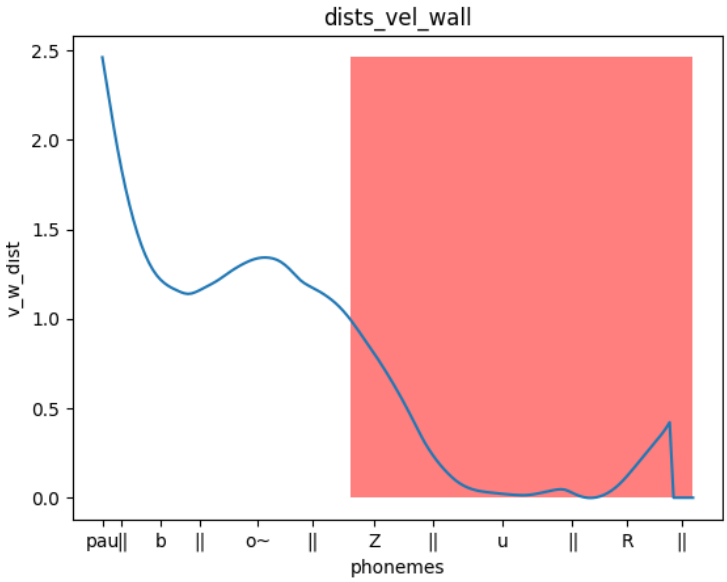


/u/, t_v_dist = 4.0 | /u/, t_v_dist = 2.8 | /ʁ/, t_v_dist = 3.0 | /i/, t_v_dist = 2.2 | /i/, t_v_dist = 2.8

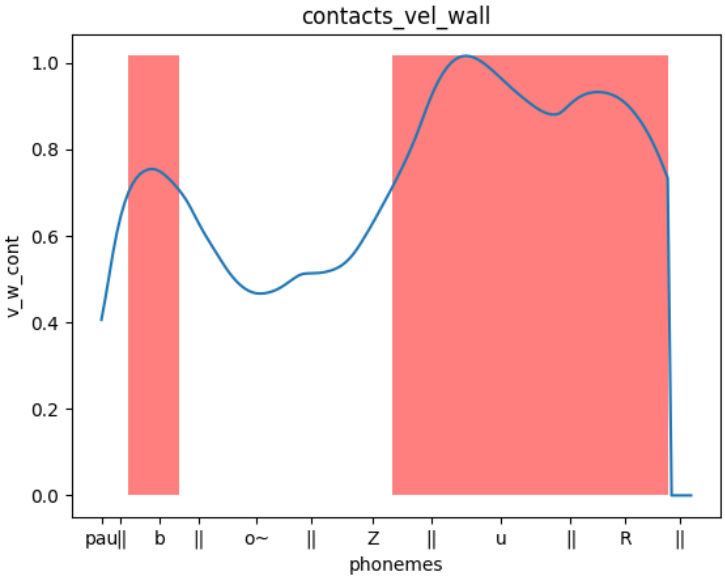
Figure 4.31: Every other frame when producing /uʁi/ and the associated t_v_dist values. The points producing the values are marked in red. Despite the interaction between the tongue and the velum to produce the rhotic /ʁ/, the extracted distances between the tongue and the velum are positive, at most recognizing frication but not contact.

Setup	Speaker	Dev/Test	MCD	BAP	F0 RMSE	F0 CORR	VUV
`no art`	S_A	Dev	4.871 dB	0.116 dB	16.176 Hz	0.740	8.859%
`no art`	S_A	Test	4.654 dB	0.131 dB	13.669 Hz	0.840	6.253%
`full art`	S_A	Dev	4.878 dB	0.119 dB	16.131 Hz	0.743	8.693%
`full art`	S_A	Test	4.709 dB	0.135 dB	15.785 Hz	0.781	6.278%
`no art`	S_B	Dev	5.352 dB	0.133 dB	18.645 Hz	0.516	7.631%
`no art`	S_B	Test	5.399 dB	0.157 dB	15.304 Hz	0.549	10.725%
`full art`	S_B	Dev	5.390 dB	0.132 dB	18.530 Hz	0.504	8.200%
`full art`	S_B	Test	5.498 dB	0.161 dB	14.777 Hz	0.563	11.567%

Table 4.3: Objective evaluation of the `no art` and `full art` setups of the acoustic-articulatory models. MCD: mean mel-cepstrum distortion (the distortion between the generated sequence and the target one). BAP: bap (band aperiodicity) prediction error. F0 RMSE: root mean square of F0. CORR: F0 correlation coefficient. VUV: frame-level voiced/unvoiced error. The `full art` setup takes slightly worse values that could be explained by the fact that the network size of `full art` stayed the same while managing more parameters—maybe it needed one more epoch; however, S_B's F0 RMSE and F0 CORR were slightly improved in `full art`. In general, S_A's results are better than S_B's.

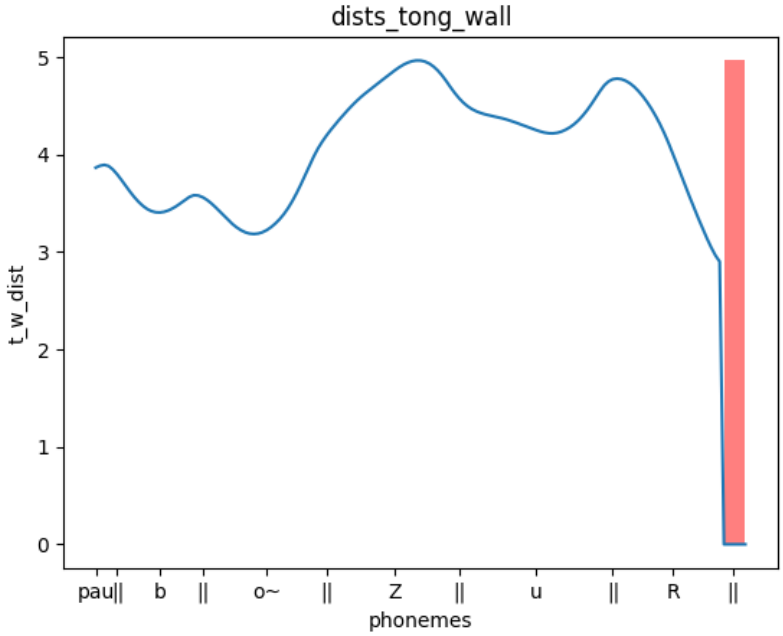


(a) v_w_dist

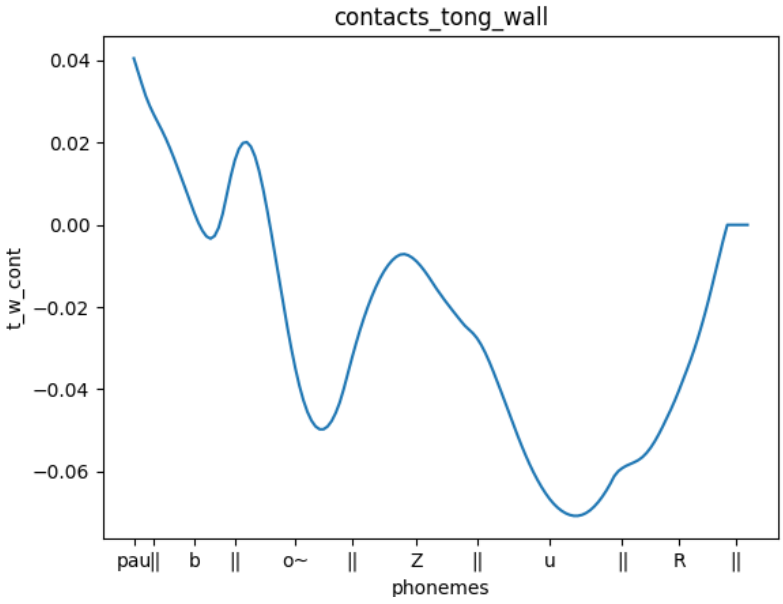


(b) v_w_cont

Figure 4.32: The synthesized sequences of v_w_dist and v_w_cont for "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A (oral samples marked in red). The division between oral and nasal sounds on all phonemes is correct with the exception of v_w_dist values during the production of /b/, which would be produced as /m/ in this configuration. This should be caused by the same effect as in t_v_dist and t_v_cont: the silent position for the velum is to be lowered, and the velum was generated to have such a high v_w_dist that the value did not have the time to get as low as necessary for /b/ without affecting the derivative (while it should be noted that the decrease from *pau* to /b/ is quite steep), and then it was already time to prepare for the nasal /ɔ̃/ coming next.

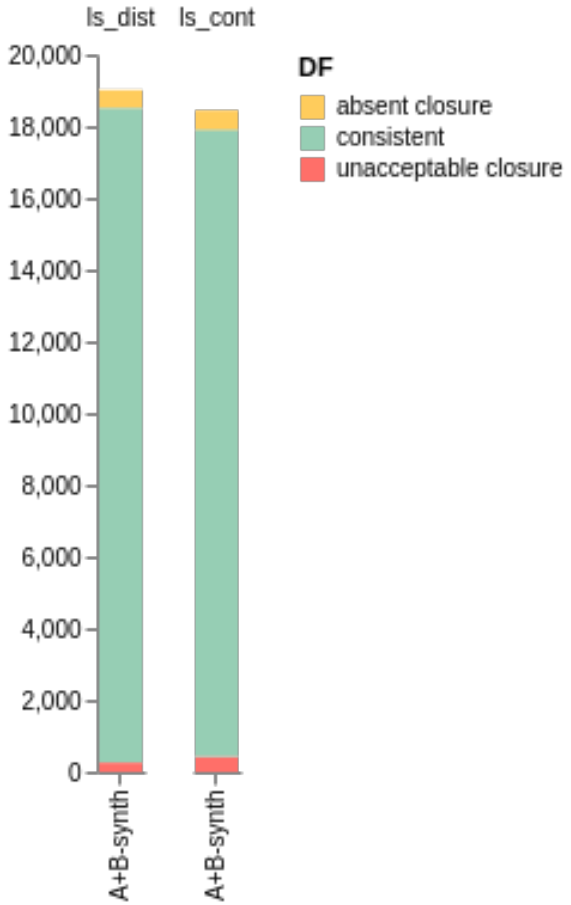


(a) t_w_dist

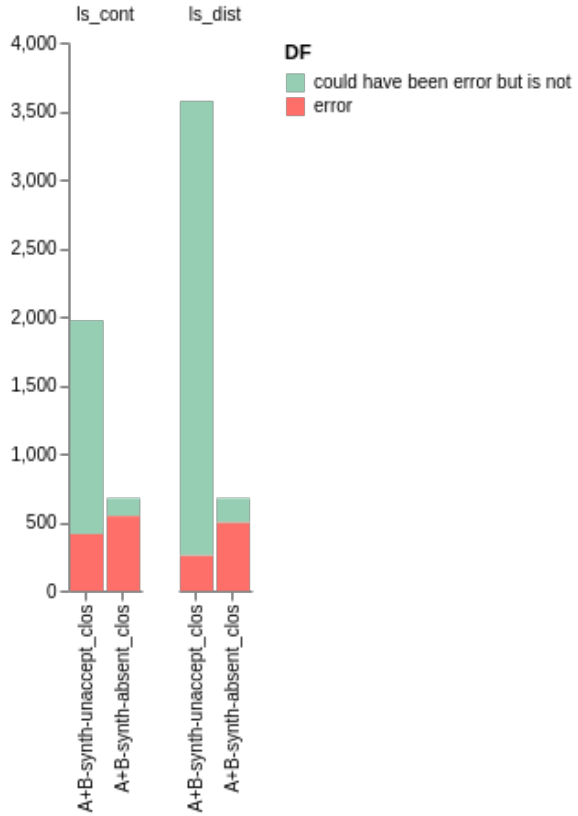


(b) t_w_cont

Figure 4.33: The synthesized sequences of t_w_dist and t_w_cont for "bonjour" /bɔ̃ʒuʁ/ with the voice of S_A. As it should be, there is no contact between the tongue and the pharyngeal wall throughout the entire sequence.

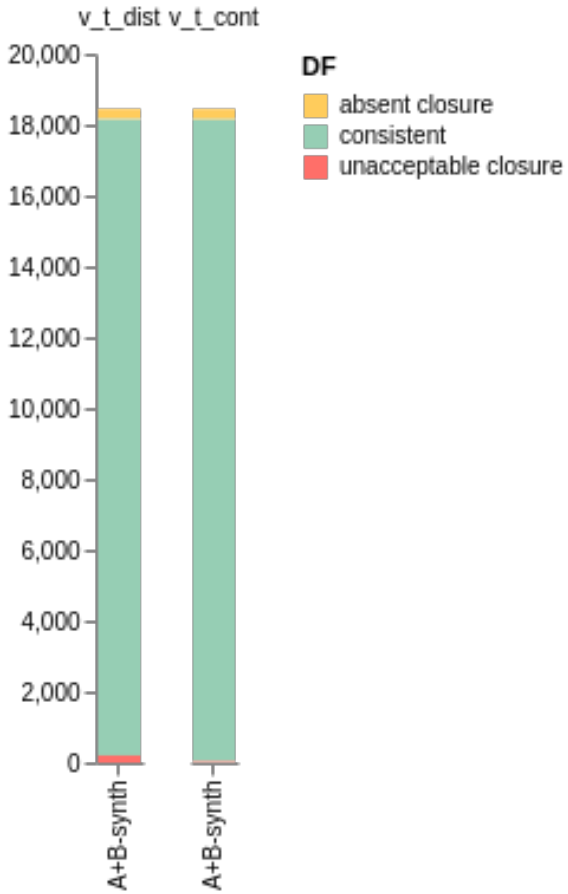


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. The number of consistent labels is high. Most errors are due to the lips not being closed.

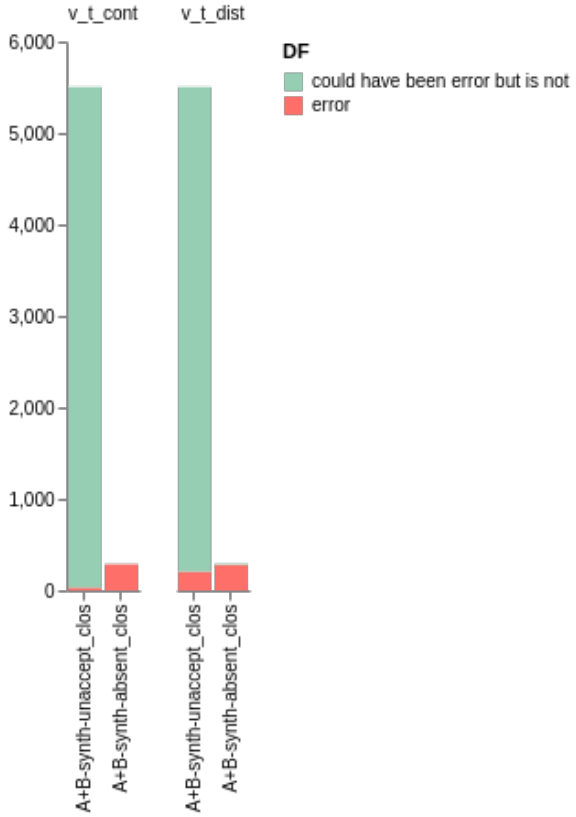


(b) ls_dist, ls_cont: absent and extra closure occurrences. The frequency of absent lip closure is higher than that of the lips being closed.

Figure 4.34: Articulatory parameter consistency in the synthesized 261 sentences—cont. on the next page.

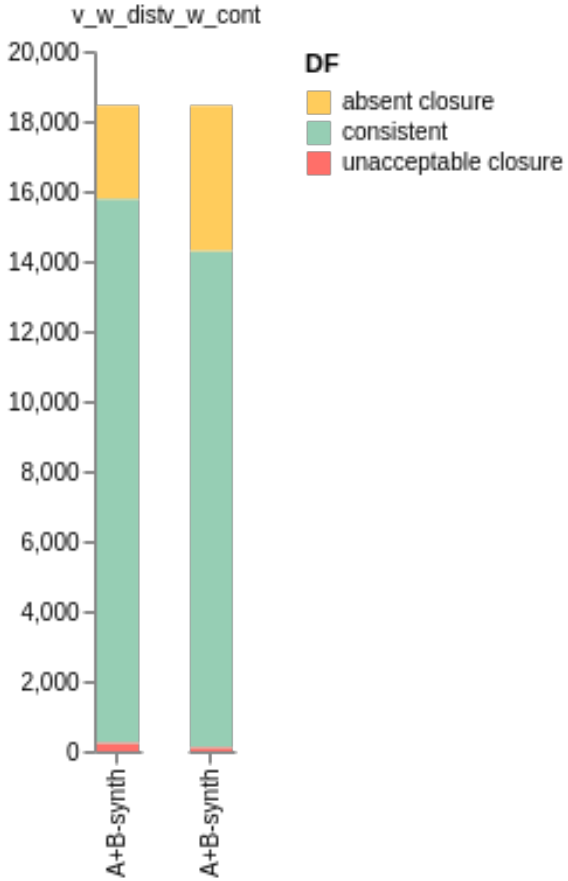


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. Almost all labels are consistent. Most errors are due to the tongue and the velum not reaching contact.

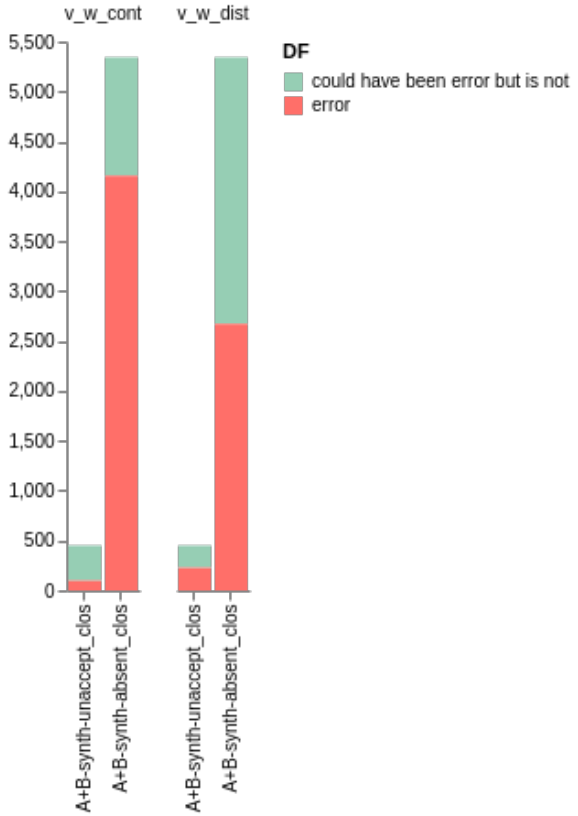


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. The frequency of an absent contact between the tongue and the velum is higher than that of them touching when they should not.

Figure 4.34: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences—cont. on the next page.

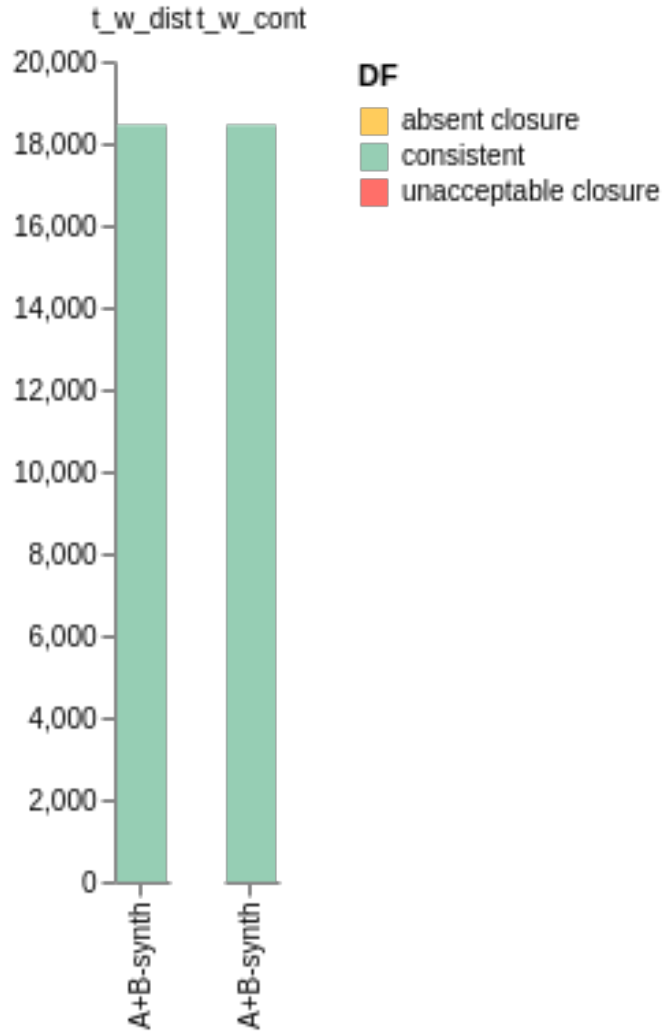


(e) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. The number of consistent labels is lower than in the case of other articulator pairs due to a higher absent closure error, which means that the sounds would be produced nasalized.

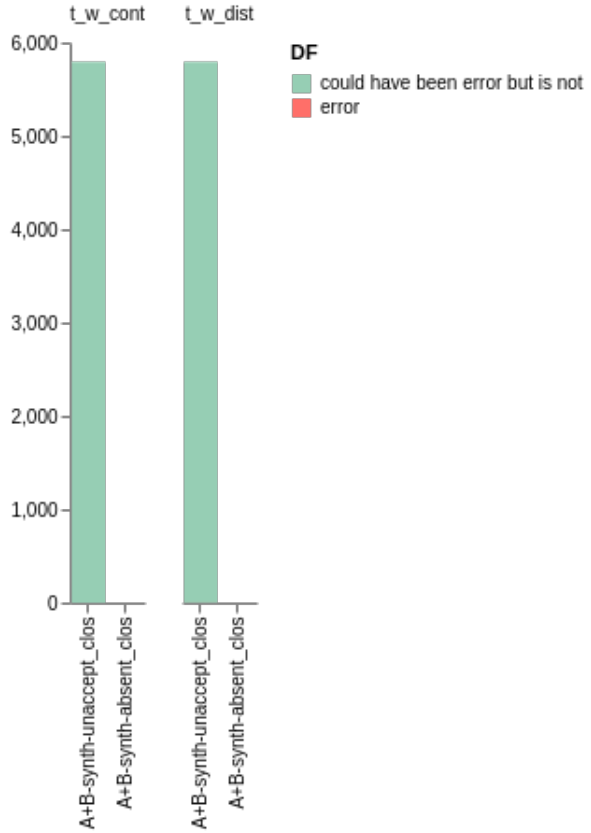


(f) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. The frequency of absent closures is higher than that of extra closures.

Figure 4.34: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences—cont. on the next page.



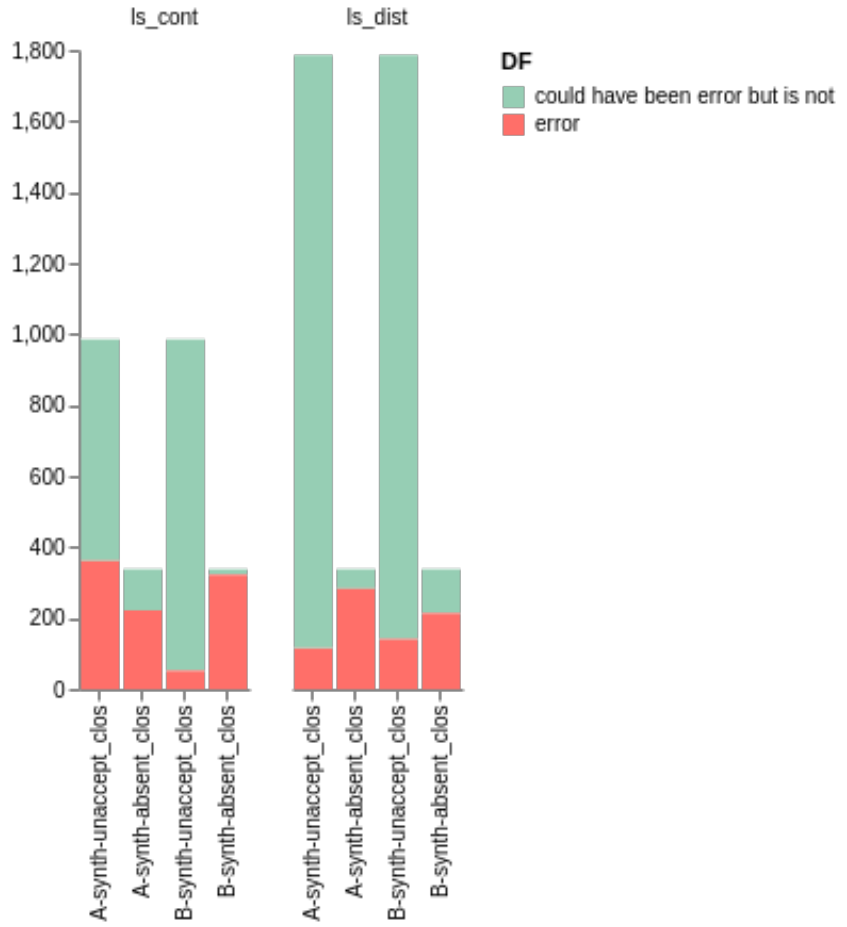
(g) t_w_dist, t_w_cont: consistent labels and extra closure. No extra closure occurrences.



(h) t_w_dist, t_w_cont: extra closure occurrences. No extra closure occurrences.

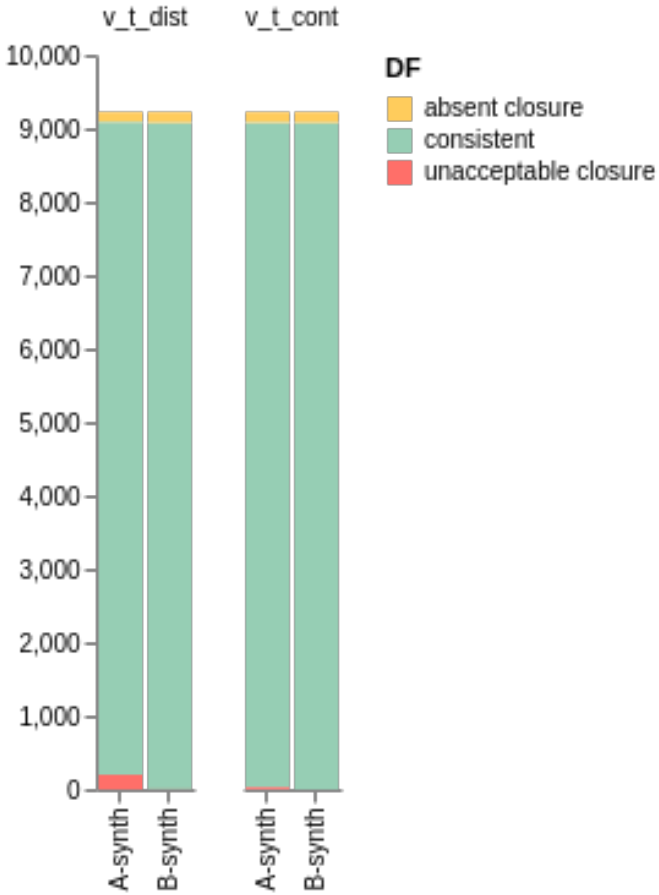
Figure 4.34: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences.

(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure. The results of both speakers are quite comparable except for an increase in lip closure when producing a sound requiring the lips to be open in S_A in ls_cont.

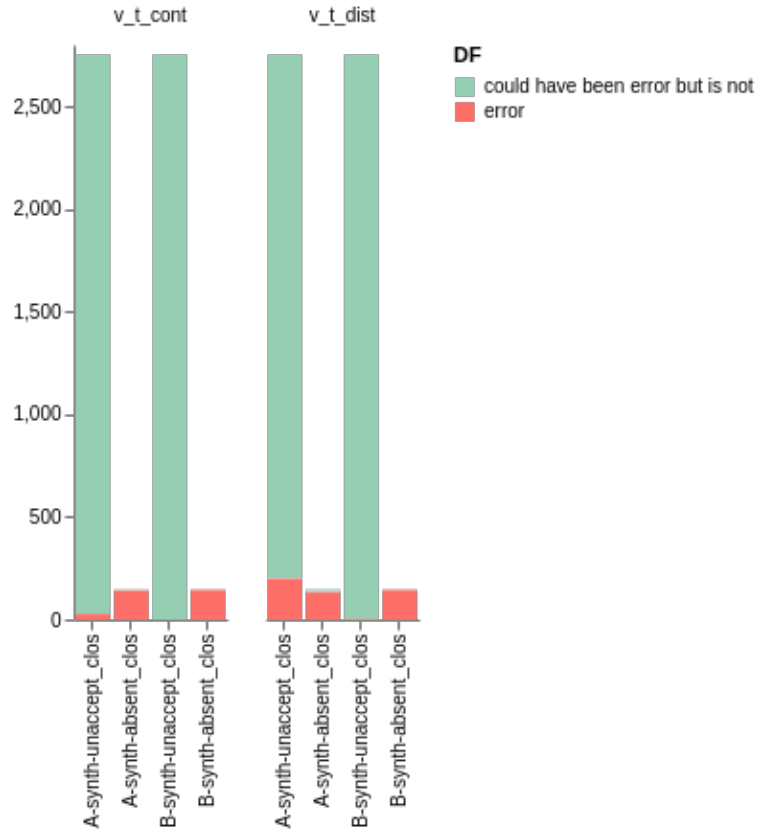


(b) ls_dist, ls_cont: absent and extra closure occurrences. The results are inconclusive.

Figure 4.35: Articulatory parameter consistency in the synthesized 261 sentences, broken down by speakers—cont. on the next page.

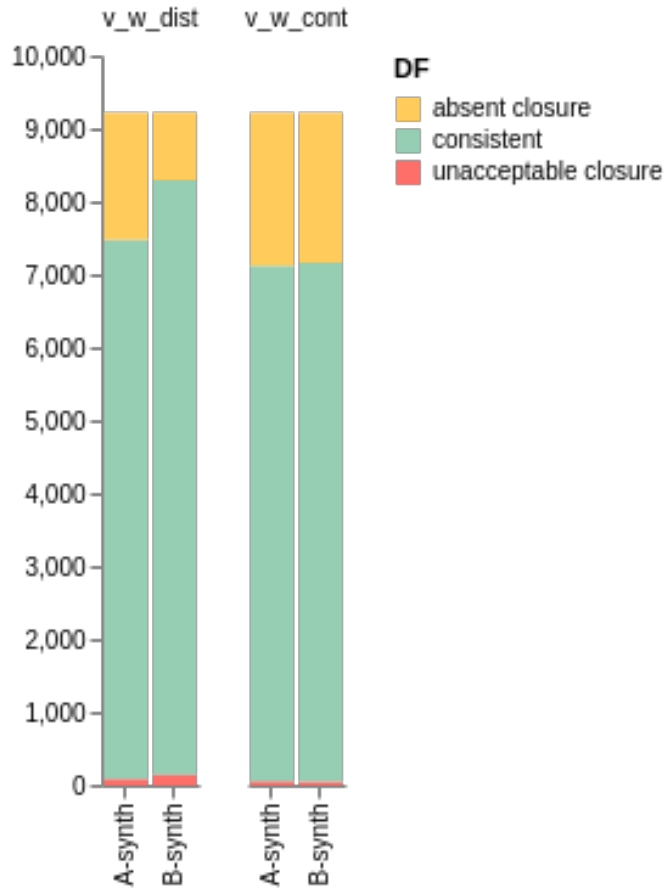


(c) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure. S_A has an increase the velum and the tongue touching during phonemes that are impossible to produce like that.

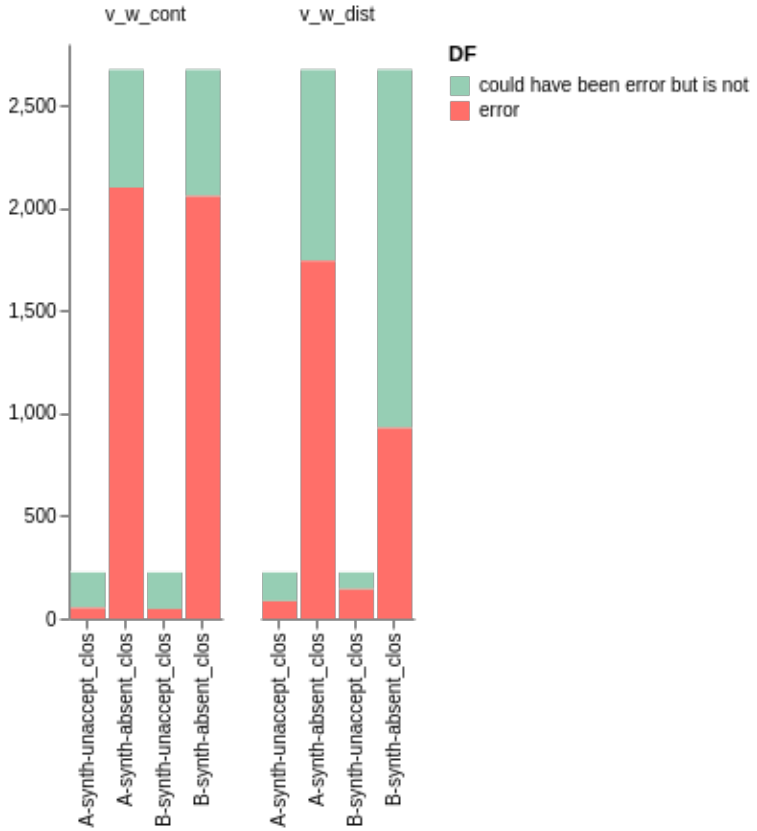


(d) t_v_dist, t_v_cont: absent and extra closure occurrences. The frequency of the tongue and the velum missing a contact is as high as 92.41% for S_A and 100.0% for S_B.

Figure 4.35: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences, broken down by speakers—cont. on the next page.

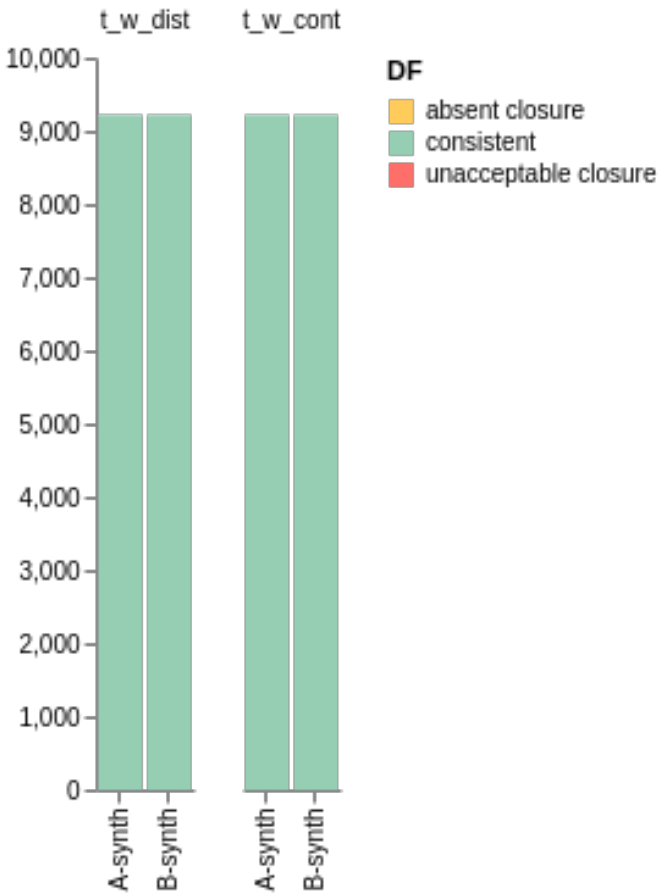


(e) *v_w_dist*, *v_w_cont*: consistent labels, absent closure and extra closure. The number of consistent labels is higher for S_A than for S_B.

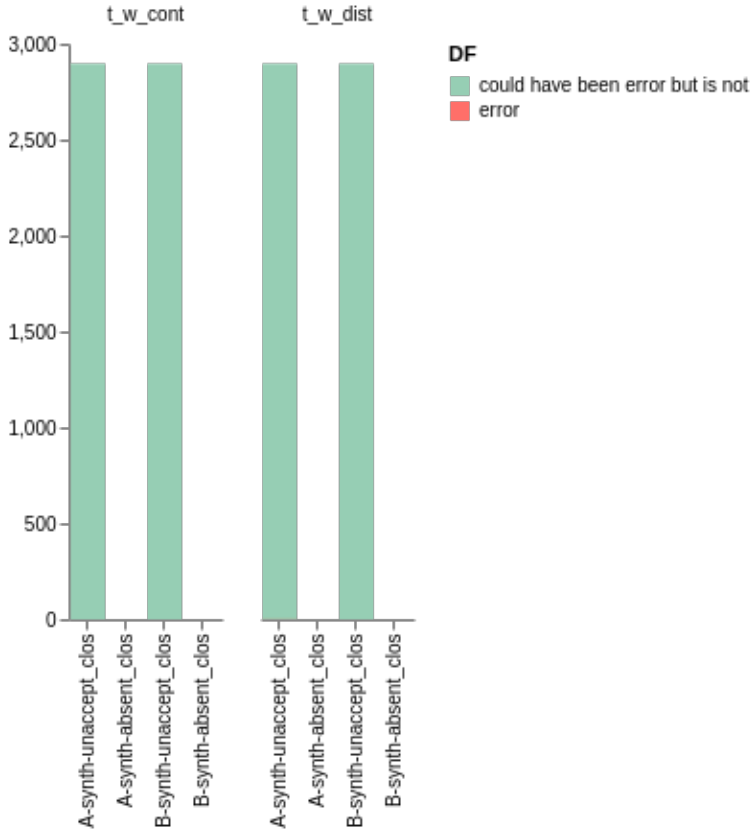


(f) *v_w_dist*, *v_w_cont*: absent and extra closure occurrences. The error frequencies are lower for S_A than for S_B.

Figure 4.35: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences, broken down by speakers—cont. on the next page.



(g) t_w_dist, t_w_cont: no extra closure.



(h) t_w_dist, t_w_cont: no extra closure occurrences.

Figure 4.35: (Cont.) Articulatory parameter consistency in the synthesized 261 sentences, broken down by speakers.

Lip protrusion, too, differed between protruded vowels and not—S_A and S_B together:

- up_l_protr: 11.84 ± 1.91 protruded, 11.49 ± 1.69 not;
- lw_l_protr: 12.69 ± 1.79 protruded, 12.45 ± 1.61 not.

It was decided against carrying out a multi-user perceptual test because both these acoustic evaluation values and an informal preliminary test suggested that while not being identical, the generated audio samples were indistinguishable to the human ear.

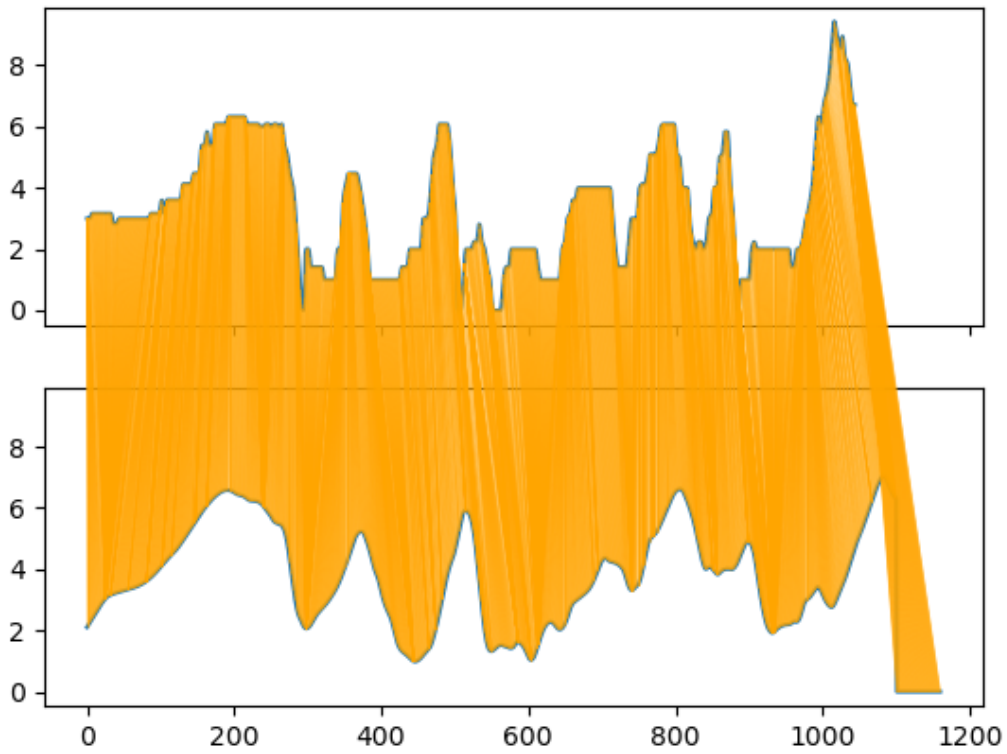
One-sentence-out articulatory evaluation

The main instrument to compare how well the synthesized articulatory sequences modeled the original ones was the distance induced by dynamic time warping.

Table 4.4 compares the DTW distance between the generated articulatory sequences (10 per each parameter per speaker) and their corresponding sequences from each speaker's part of the corpus that were left out in the model's database. While in general the distances between the synthesized sequences and the original ones are higher than those between the original ones themselves, since the variance is high, they still fall within the range of the acceptable. Figures 4.36 show examples of aligned articulatory parameter sequences.

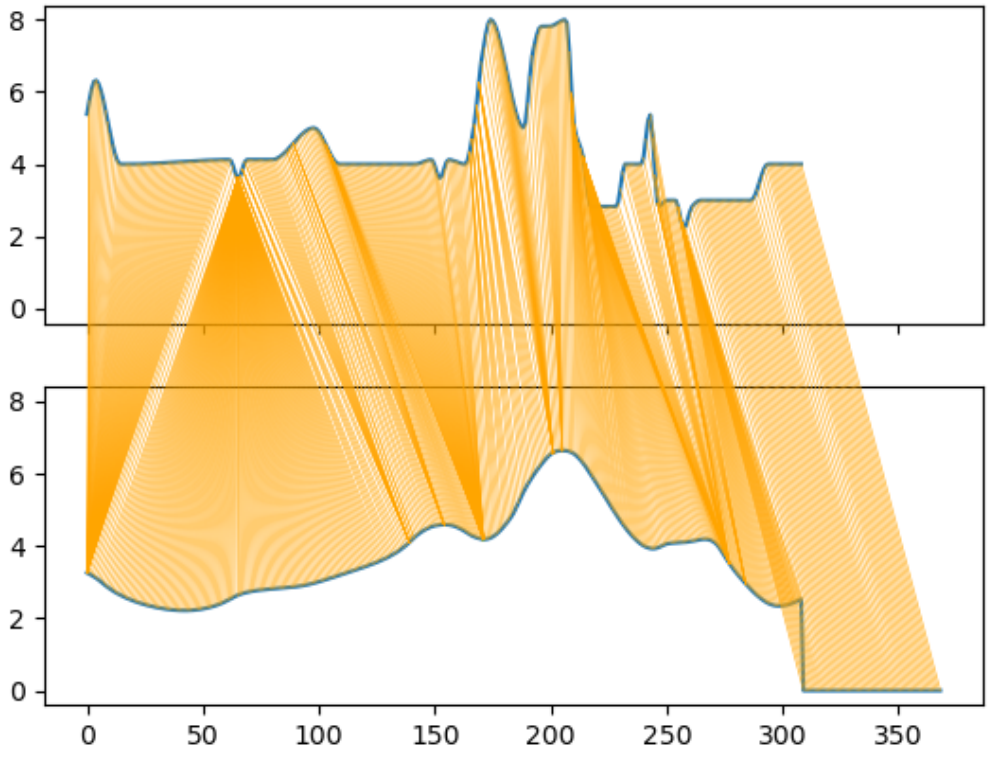
Parameter	Speaker	$M_{\text{synth to corp}}[dtw] \pm \sigma_{\text{synth to corp}}[dtw]$	$M_{\text{intracorp}}[dtw] \pm \sigma_{\text{intracorp}}[dtw]$
ls_dist	S_A	49.67 ± 16.86	39.28 ± 18.53
ls_dist	S_B	47.55 ± 13.88	32.83 ± 19.37
ls_cont	S_A	20.96 ± 13.10	15.51 ± 10.23
ls_cont	S_B	12.85 ± 9.75	11.08 ± 12.81
t_v_dist	S_A	47.10 ± 15.91	35.47 ± 18.28
t_v_dist	S_B	58.64 ± 19.47	46.09 ± 27.90
t_v_cont	S_A	7.92 ± 5.34	4.80 ± 5.68
t_v_cont	S_B	3.41 ± 4.64	2.18 ± 3.77
v_w_dist	S_A	32.63 ± 23.49	29.78 ± 17.09
v_w_dist	S_B	36.31 ± 19.64	31.69 ± 17.82
v_w_cont	S_A	7.30 ± 2.16	4.94 ± 3.15
v_w_cont	S_B	10.50 ± 2.61	6.90 ± 4.64
t_w_dist	S_A	63.82 ± 15.93	26.48 ± 12.82
t_w_dist	S_B	59.72 ± 10.93	37.73 ± 15.67
t_w_cont	S_A	1.02 ± 0.44	0.00 ± 0.00
t_w_cont	S_B	2.76 ± 3.07	3.24 ± 4.02

Table 4.4: The mean dynamic-time-warping distance between the generated articulatory parameter sequence and the corresponding sentences in the corpus (synth to corp), compared to the same distances between the original sentences (intracorp). While synth to corp distances are greater than intracorp, variance is very high, so they still fall within the acceptable range. dist parameters get synthesized closer to the corpus values for S_A; cont, for S_B.



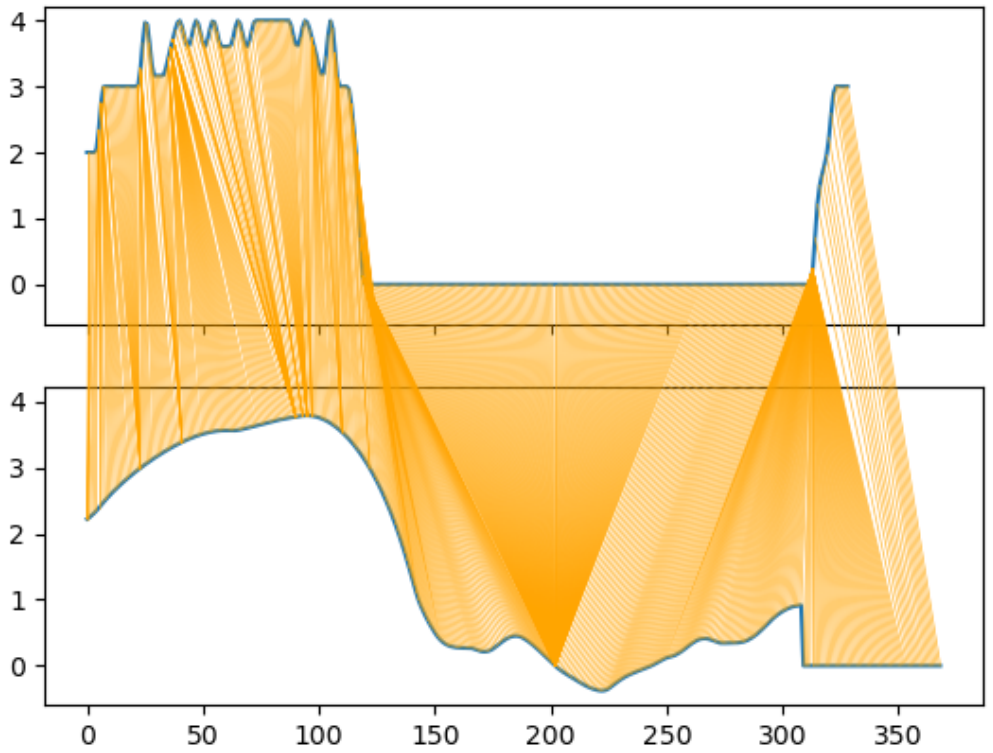
(a) ls_dist, S_B,"Il éblouit le veau et les pioupious qui sautaient à une encablure du Cher", DTW = 41.14, the timing and values are quite well-aligned.

Figure 4.36: DTW alignment between the original (above) and generated (below) articulatory parameter sequences—cont. on the next page.



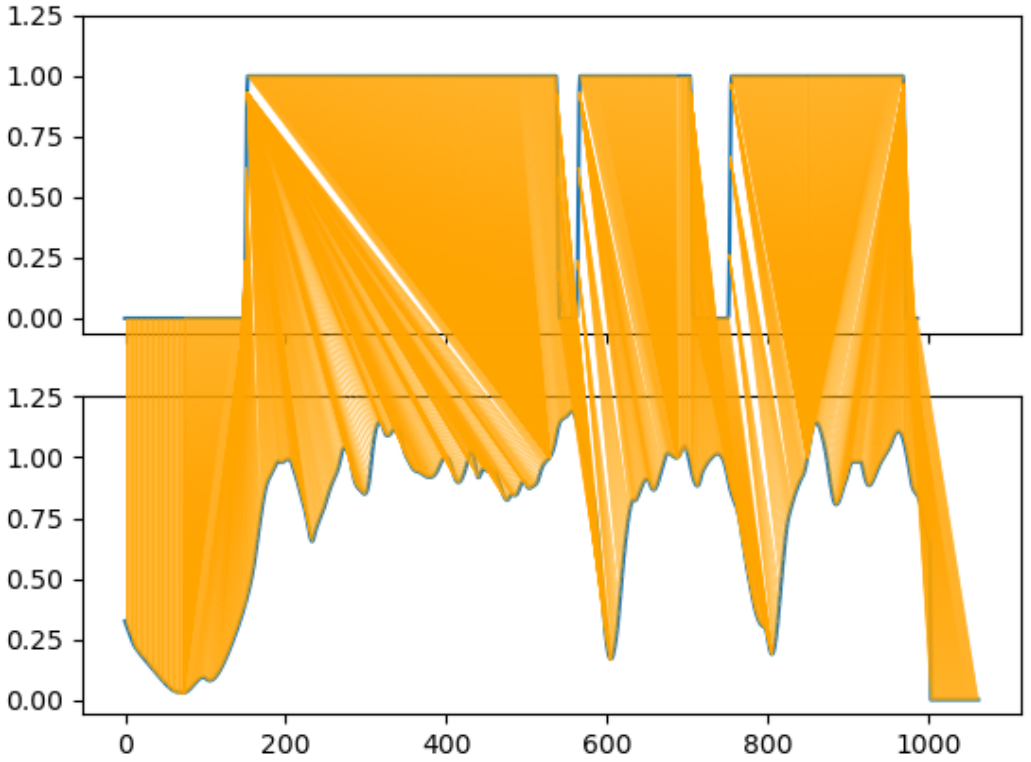
(b) t_v_dist, S_A,"Il a pourri", DTW = 22.55. The peaks are more moderate.

Figure 4.36: (Cont.) DTW alignment between the closest matching original (above) and generated (below) articulatory parameter sequences—cont. on the next page.



(c) *v_w_dist*, S_A,"Il a pourri", DTW = 9.07. The generated sequence is not as stable (fluctuations around 0 while the original sequence was just constant 0), but its approximate values and the timing are correct.

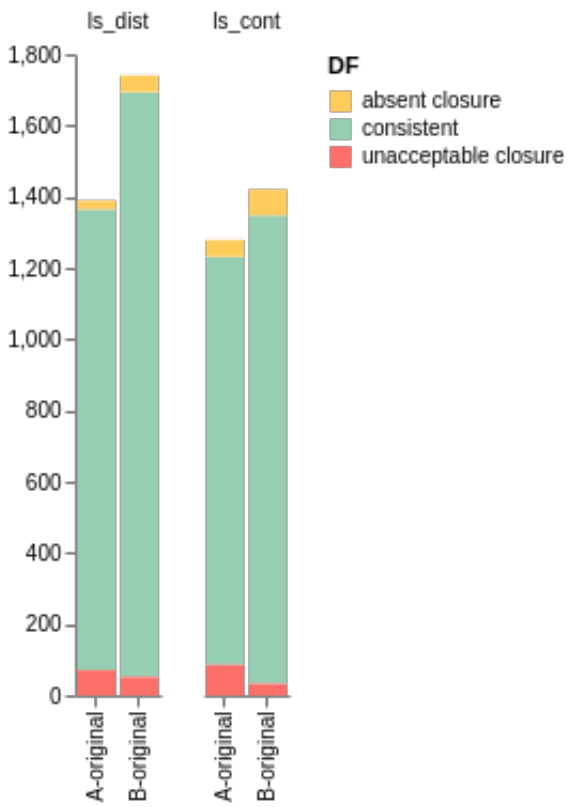
Figure 4.36: (Cont.) DTW alignment between the closest matching original (above) and generated (below) articulatory parameter sequences—cont. on the next page.



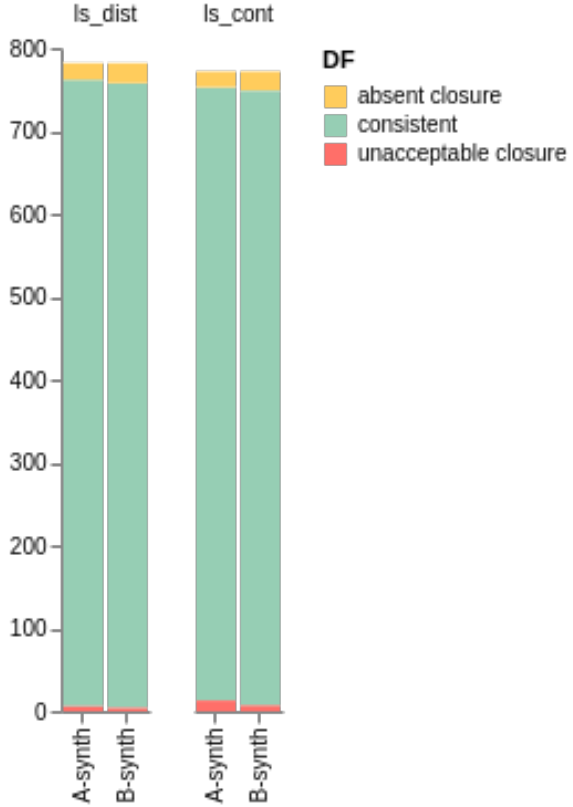
(d) *v_w_cont*, S_A,"Il éblouit le veau et les pioupious qui sautaient à une encablure du Cher", DTW = 3.43. The generated sequence is not as stable, but its approximate value and the timing are correct.

Figure 4.36: (Cont.) DTW alignment between the closest matching original (above) and generated (below) articulatory parameter sequences.

Additionally, I compared phonetic label consistency with the parameter labels for the generated sequences and their original counterparts taken out from the corpus. The conclusion is the same as in the large-scale evaluation: that it is more complicated to generate a contact, therefore the number of unacceptable closure drops in synthesis, and the rate of absent closures increases. As for the frequencies of errors, they come out to be related to what was in the original data not used to train the model.

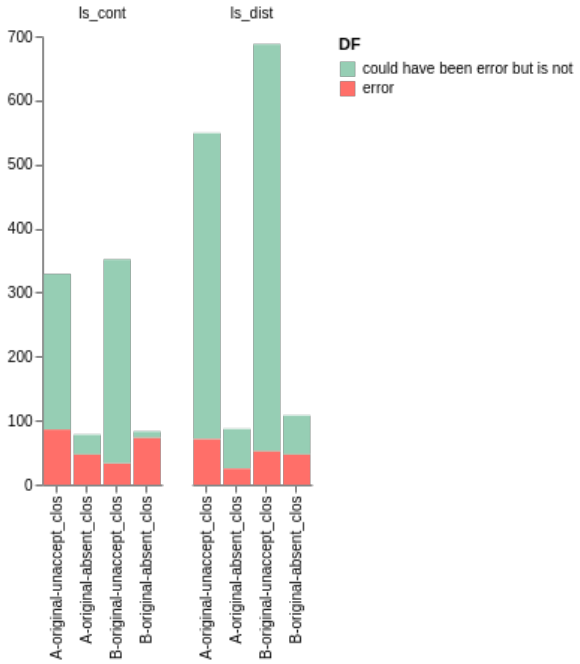


(a) ls_dist, ls_cont: consistent labels, absent closure and extra closure—the original counts.

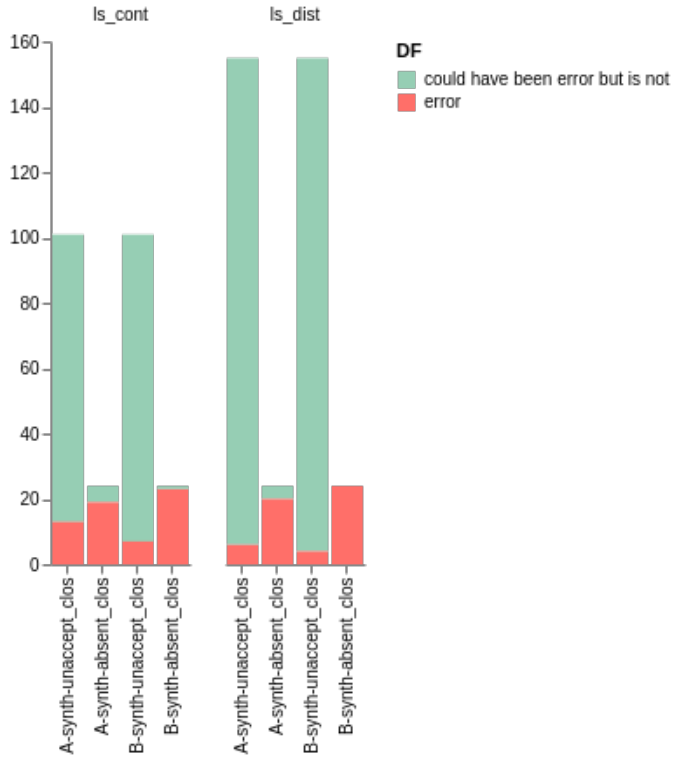


(b) ls_dist, ls_cont: consistent labels, absent closure and extra closure—the synthesis counts.

Figure 4.37: Articulatory parameter consistency compared on the original sentences that were taken out and the generated ones, broken down by speakers—cont. on the next page.

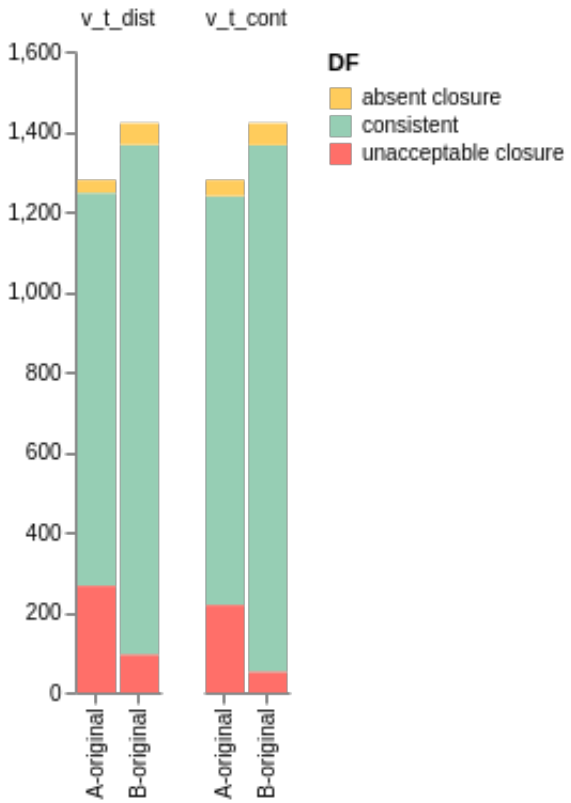


(c) ls_dist, ls_cont: absent and extra closure occurrences—the original counts.

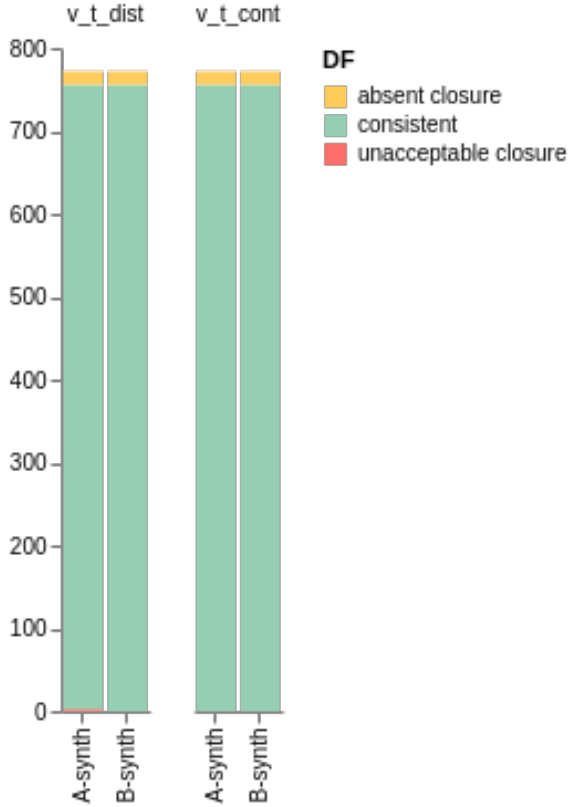


(d) ls_dist, ls_cont: absent and extra closure occurrences—the synthesis counts.

Figure 4.37: (Cont.) Articulatory parameter consistency compared on the original sentences that were taken out and the generated ones, broken down by speakers—cont. on the next page.

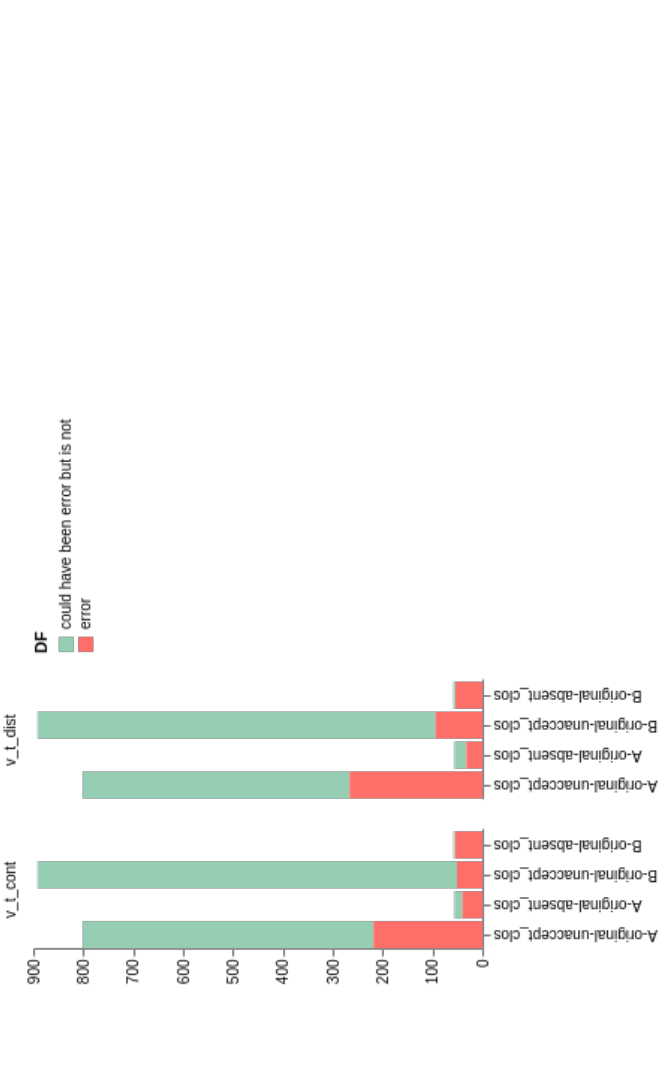


(e) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure—the original counts.

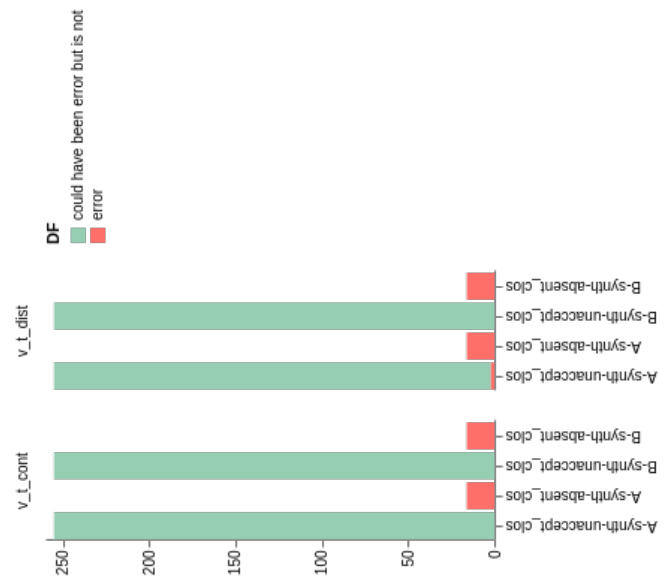


(f) t_v_dist, t_v_cont: consistent labels, absent closure and extra closure—the synthesis counts.

Figure 4.37: (Cont.) Articulatory parameter consistency compared on the original sentences that were taken out and the generated ones, broken down by speakers—cont. on the next page.



(g) `t_v_dist`, `t_v_cont`: absent and extra closure occurrences—the original counts.



(h) `t_v_dist`, `t_v_cont`: absent and extra closure occurrences—the synthesis counts.

Figure 4.37: (Cont.) Articulatory parameter consistency compared on the original sentences that were taken out and the generated ones, broken down by speakers.

Finally, I looked into lip protrusion in the generated sequences and in the original data. Again, the upper lip parameter up_l_protr was better at differentiating between protruded and non-protruded vowels; when verifying whether the values corresponding to protruded vowels were on average larger than those to non-protruded ones, this condition was true 61.40% of the times in synthesis and 59.65% in the corresponding original data.

4.4 Conclusion

4.4.1 Overview of the results

This chapter presented a DNN-based articulatory speech synthesizer, where articulation was extracted fully automatically at a comprehensive rate and modeled jointly with the acoustics of the aligned signal. Substantial work was conducted to represent articulatory information in a consistent way suitable for the network, i.e. as a set of eight parameters corresponding to the lips opening and contact surface, the distance between the velum and the tongue and the categorical parameter stating the presence or absence of a contact between them, the distance between the velum and the pharyngeal wall and its own categorical parameter, and the distance between the tongue and the pharyngeal wall. These parameters should therefore be able to represent numerous articulatory effects in speech production:

- Full behavior of the mid-sagittal slice of the lips: lips protrusion and opening;
- Nasality, represented through the distance and contact between the velum and the pharyngeal wall;
- Partial information of the velar region: no notion of the velum and tongue position, but a set of values that depend on them:
  - Circumstantial evidence of the tongue height through the distance between the tongue and the velum;
  - Circumstantial evidence of the tongue backness through the distance between the tongue and the pharyngeal wall.

The results suggest that the original acoustic model can handle the added information, meaning that the extracted parameters were generally coherent with the acoustic data: objective evaluation for the `full art` setup comes up to almost identical `no art` MCD, BAP, F0 RMSE, F0 CORR and VUV values, and perceptually the generated samples in the two modes are indistinguishable.

The generated articulatory parameter sequences match the original ones acceptably closely. They struggle more at attaining a contact between the articulators, which results in a reduced rate of prohibited closure errors and an increased one in absent contacts.

Thus, within the defined objectives, obtaining a full-fledged articulatory speech synthesizer can be considered complete.

4.4.2 Future work

There are a few experiments and improvements that could still help us gain some insight regarding this line of research.

For instance, it would be useful to study the exact behavior of the distortion and error functions, especially of the articulatory parameters error, over the time of acoustic model training. Since the `full art` setup introduced new parameters, it could be the case that it would be justified to modify the number of training epochs or to adapt the layers in the network.

Then, the study would be much more complete with experimenting with other types of neural networks, especially LSTMs and BLSTMs known to excel in speech synthesis thanks to their improved management of temporal relations, and adjusting their parameters. Also, it would make sense to process the generated articulatory parameters at the final layer so that they stay within the interpretable ranges, for example, without reaching negative values.

From the purely speech production perspective, it is also interesting to study the contribution of each of the parameters. For example, the pair of the tongue and the pharyngeal wall is not part of the place of articulation of any phonemes of French, and throughout the analysis it oftentimes appeared secondary despite its alluded contribution to estimating the backness of the tongue. This brings out the question of whether this piece of information was useful to the network or not. It could turn out, for example, that the lips are much more meaningful in the resulting sound generation than the distance between the tongue and the velum; this is an open research question that merits being studied.

Continuing the line of thought of separating different contributions, we saw some differences between spontaneous and non-spontaneous speech. While spontaneous speech is more difficult to align due to more irregularities with respect to the standard carefully articulated production of the delivered phrases, it cannot be denied that it is immensely more natural, for example, through exhibiting the speaker's uncertainty that is associated as an attribute of natural speech and can be picked up by articulatory speech synthesis [BFS+]. In order to learn more about constructing articulatory speech databases in the future, it would be useful to evaluate the importance of spontaneous speech.

One of the major issues of the study was the problem of phonetic alignment. With the recording being quite noisy and the video being in asynchrony with the audio, it is of no doubt that obtaining a correct correspondence between the timing of the audio recording, the MRI frame and the linguistic annotation label is a tall order. Nevertheless, as it was commented upon in the section about the analysis of phonetic label inconsistencies, oftentimes articulation makes it quite clear that there is a time shift (the example of /b, b, b, b, i, i, i, i, i, i, i, i, i/ could produce $---++----------$, where $+$ stands for the lips closure). It would be expected to improve phonetic alignment dramatically if we integrated articulatory information in the alignment process. Having a separate clean recording of prompted speech, devoid of the noise of the MRI machine, and transferring its phonetic labeling onto our denoised samples, could also be of help, provided that the noise does not hinder the alignment between the two versions of the utterance.

Finally, a very promising—and more ambitious—direction of follow-up research would be to represent articulatory information of the whole vocal tract rather than a few selected areas. The challenge in that is that we would have most likely to do without precise outlines of the articulators, since they could be too prone to mistakes. For this, a potential solution could be

to represent the vocal tract configuration with methods such as [LSNQ18] and then to encode it with statistical methods such as PCA or deep learning methods such as autoencoders. There is, however, a growing body of research to track full articulator outlines with various amount of detail with supervised, semi-supervised and unsupervised machine learning, such as [TGH+19], so eventually it should be possible to have precisely segmented images as well.

5

Static targets versus running speech for articulatory speech synthesis

In the previous chapters, we have seen what can be done for articulatory speech synthesis using static or dynamic MRI, with quite different methods.

The difference between the two data types is inherent.

On the one hand, MRI can capture the position of a vocal tract that was held stable over the acquisition time (typically a dozen or more 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 with a high resolution, but due to the extended acquisition time, this picture is frozen.

On the other hand, the protocol of RT-MRI selects only one piece of 3D volume; for speech production research, typically the mid-sagittal one. It captures the tissues within that slice in real time [LZL$^+$19], which enables us to analyze rapid-paced speech movements. The speech observed with such a method is unrestricted and therefore highly natural, allowing for a deep understanding of the dynamics of the articulators [NTR$^+$14, TN16, RTP$^+$18] and, as Chapter 4 showed, for speech synthesis based on them. However, it cannot be denied that in the attempt to gain enough temporal coverage in RT-MRI we lose a lot of image sharpness and clarity. If the slice is not thin enough, the intricate geometry of the articulators gets projected on a single plane (there are phonemes with quite complex three-dimensional behavior, such as the lateral /l/); if the speaker moves too fast, no position will be manifested for long enough to be captured by the machine (since each elementary acquisition lasts 2 or 3 ms and the last radius gets acquired around 18 ms after the first). Both of these points can result in ghosting (for example, the presence of two outlines of the tongue tip, which is an especially rapid articulator), image blurring or other artifacts, subsequently affecting the analysis and rendering image segmentation especially difficult.

As shown in the previous chapters, our case was not an exception to the general rule. Table 5.1 summarized that while being an optimal solution for capturing consciously controlled vocal tract positions with a high resolution, static MRI captures had the disadvantage of demonstrating unrealistic, probably overarticulated vocal tract shapes. The speaker had apparent trouble avoiding excessive nasalization as there is no conscious control of the velum when it is

	MRI	RT-MRI
Advantages	Image resolution, sharpness and quality "Distilled" manifestation of coarticulation-influenced vocal tract configuration	Unrestrained, natural speech in real time Speaker's comfort
Disadvantages	No temporal information Incorrect vocal tract positions	Reduced image quality No (entirely) conscious control over articulation

Table 5.1: Advantages and disadvantages in regular (static) MRI and RT-MRI, summarized.

not actively used in speech production; also, at showing an appropriate context when articulating particular phonemes (for example, not raising the tongue dorsum enough for anticipating /i/—see Fig. 5.1,—while /i/ is *defined* as a close front unrounded vowel; using this target to guide an articulatory speech synthesizer would naturally lead to a misinterpretation of anticipatory coarticulation for /i/, making it more similar to the effects of /a/, an open central unrounded vowel in French). There is evidence that phonemes that are intrinsically dynamic in articulation, such as liquids, were also misrepresented by the static protocol of MRI acquisitions, leading again to unrealistic shapes [LETV18].

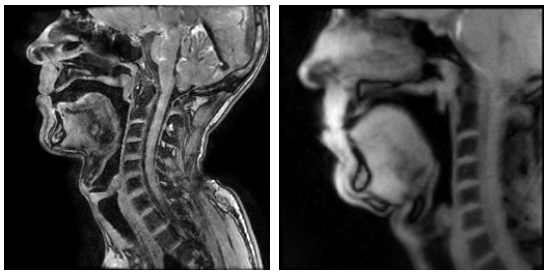


Figure 5.1: Static (left) and a typical dynamic (right) recording of /p(i)/ articulation by speaker S_A. Note how the tongue dorsum is not raised high enough in the static dataset; this tongue position would be much better explained by anticipating a more open sound like /a/. This could have been helped with a special voice and accent training preliminary to the acquisitions.

This brings us to the need to build a bridge between the two directions of work. More precise details follow below.

5.1 Objectives

In our two approaches to treating the mechanics of speech production , the pivoting point was manifesting or circumventing the notion of an elementary unit of articulation.

Within the static approach, we explicitly built the system around some key articulatory positions revealing coarticulation and serving as articulatory targets, every articulatory motion generated as a transition from the previous target to the next in a given amount of time.

As for the dynamic one, there articulatory parameter sequences were generated similarly to acoustic parameters, statistically driven by the relation identified in the data between linguistic specification and articulatory parameter values. It does imply that the speaker's intended utterance guides their articulation, but does not specify how. Any generated value is a result of the fact that the system has had enough examples of when this linguistic specification was seen together with this value and its derivative estimation; there is no strategy involved.

To further illustrate the last point, let us explore the following example: if we look at the parameter of lips opening in the synthesis of the syllable /ba/, the production of the stop /b/ can only happen if the lips collide with force, involving the muscles of the lips, the pressure in the oral cavity builds up until the certain point at which the lips part, both because of the air forcing them to do so and because of the jaw movement that proceeds to the open vowel /a/. If the contact is fleeting, with no force, then what will be produced is a bilabial flap: /b̆a/; if there is no contact at all, just the lips coming closer to each other, it is a bilabial fricative (or approximant): /βa/; if there is contact, but it does not get released during /a/, no /a/ will be produced, since /a/ is a vowel and not an obstruent. So, whether the lips close or not, is something that defines whether the synthetic sound will be recognized as what was intended or not.

This means that from the articulatory perspective lips closure is less of a question of precision (whether the opening is 2 cm^2 or 2.2 cm^2, or 0.0 cm^2 versus 0.2 cm^2) and rather whether it is present or not. In other words, the closure is a categorical value, which the resulting speech intelligibility entirely hinges upon. Thus it would ideally require some kind of planning, and that would lead us to go back to modeling the organization of articulatory movements.

This motivates us to see whether we can find any trace of such coarticulation-aware targets as we had in the static case present in the dynamic data. Not only will this serve as validation to the static-target approach—show whether it is possible to use this kind of static articulatory representations as points of reference for the dynamics of speech—and serve as a stepping stone to potential hybrid methods, it will also advance us in RT-MRI data treatment in general: for example, it can serve as a base for the decision on which RT-MRI captures to annotate manually.

The aim of the part of work is to look for such static, frozen articulatory targets that were constructed in the static MRI corpus in RT-MRI data and give an interpretation either of their presence, possibly to some extent or only in some cases, or absence, effectively drawing a link between the two types of data. The objective is to employ measures that are proven to be efficient in image processing and computer vision feature extraction to compare the static and dynamic MRI datasets, and to draw conclusions from the dynamics and distributions of these similarity measures.

This work was a joint effort of Ioannis Douros, Anastasia Shimorina and myself: the methodology was preliminarily validated by Ioannis Douros and me, Ioannis Douros prepared the processing protocol of the initial set of input images (that were later replaced by the vocal tract windows processed and cut as described in the previous chapter), the computation was set up by me and run with the help of Ioannis Douros and Anastasia Shimorina, the results were aggregated by Anastasia Shimorina and me, and evaluation and analysis were done by me, though Anastasia Shimorina provided helpful discussions about their methodology.

5.2 Data and methods

The present study used two MRI corpora: classic static MRI, the same as in Chapter 3, and RT-MRI, as in Chapter 4. The outline of the work was as follows:

- Treat the static and the dynamic frames to make them comparable;
- Utilize structural similarity, Earth mover's distance, and SIFT matches to compare them;
- Verify the validity and consistency of the obtained measures;
- Study the temporal behavior of each measure and interpret it;
- Analyze the identified similarities.

The following sections will cover these steps in greater detail.

5.2.1 Treating MRI and RT-MRI captures

When matching the images of these two datasets, one has to face several issues:

(1) The resolution and quality of the images is not the same: 256×256 pixels against 189×189. Furthermore, MRI is very sensitive to movement, resulting in a certain amount of blurring in the dynamic images.

(2) The images do not depict exactly the same areas of the subjects' vocal tracts, nor do the subjects take exactly same positions or posture. Moreover, three years passed between acquiring the static and dynamic datasets, resulting in some minor physical changes in speaker S_A; and naturally, there are differences between speakers S_A and S_B.

(3) Static acquisitions may produce shapes that will never be observed in dynamic data since they involve no phonation and there are phonemes whose sustained imitation of articulation is either difficult or impossible (liquids due to their dynamic nature—a point raised in [LETV18]; stops, whose burst is a result of pressure building up in the vocal tract; it is difficult to control nasality).

(4) The static dataset is rather small and should not be expected to be able to cover all the images in the dynamic dataset.

(5) While being larger, the dynamic dataset still remains relatively small as far as speech resources go. When breaking down into specific contexts, phoneme classes, syntactic structures, or speaking styles, data sparsity quickly becomes an issue.

Taking all the points above into account, we created rectangular windows that only contained the vocal tract information (Fig. 4.2a), from the laryngeal region to the lips, as was explained in Chapter 4.2.1. This cropping relied on the reference points of the tip of the nose and the corner of one of the vertebrae, which is less reliable than fitting the articulatory model derived from static images to the dynamic ones, but with the upside of being usable in the absence of articulatory contours. These windows were applied both to the static and dynamic data. Then the smaller windows were resized to images of size 84×82 pixels.

Static MRI

We applied histogram matching to each static image so that its image characteristics follow the ones of a sample dynamic image. As an aside, since the head position of speaker S_A was quite different in the static MRI dataset compared to the RT-MRI one and, out of all chosen techniques, only SIFT is rotation- and location-invariant, we turned the static images by $12.3°$ so that the vocal tract angles were aligned with those in the dynamic ones.

As was explained in Chapter 3.3.1, each of the 95 MRI captures corresponded either to a vowel V or to a blocked consonant-vowel $C(V)$ articulation.

RT-MRI

There were some RT-MRI images where the cropping algorithm for leaving out all articulatorily irrelevant information failed due to an incorrectly determined nose tip (this part of the work was done before the speaker- and sequence-specific improvements to selecting windows that were described in Chapter 4). These sequences were kept out.

For the purpose of analysis, each cropped RT-MRI frame needed to correspond to a phonetic label which was a force-aligned [YEG+02] [WWK16] corrected output of eLite HTS [RBBD14] (see Chapter 4.2.1 for details). Every sequence where the phonetization algorithm failed was thus excluded too.

In the end it left us with $368,848$ RT-MRI frames.

5.2.2 Image comparison measures

Ideally, an efficient algorithm for image comparison should be able to:

- Identify a shared manner of articulation such as: a contact for a stop or a nasal, a narrowing in the vocal tract for a fricative or a liquid, the absence of obstruction for a vowel, an open or close velopharyngeal port for a nasal or an oral sound. Any such shared feature should raise the resulting value of the similarity measure by a few points.

- Capture a shared place of articulation and the critical articulator: identify when both the static and the dynamic frames depict, for example, an alveolar consonant, and give points for that. This way, when the phonemes are identical up to the feature of voice (for example, static /k/ and dynamic /k/ or static /p/ and dynamic /b/), they should be recognized as even more similar than those that share only the place or only the manner.

- Be sensitive to coarticulatory effects: we would hope that because of the protruded lips /p(y)/ should turn out to be closer to /y/ than, say, /p(a)/ is; or, to make another example, the tongue dorsum in /t(i)/ should already be raised to anticipate the close vowel /i/ and thus /t(i)/ should be quite similar to /i/, which should not be the case for /t(ɛ)/ where the tongue is already preparing for an open /ɛ/ position.

The same criteria were then applied when developing a method to evaluate the results of this work.

We chose to stay as rigorous in our approach as possible and to have a coherent measure between each of the static images and each of the dynamic images. We cut out the rectangular of

the vocal tract and resized the resulting images to 84×82 pixels. Then three techniques which are known to perform well in image processing and computer vision feature extraction were used: structural similarity (SSIM) [WBSS04, Ava09], Earth mover's distance (EMD) [RTG00] (Wasserstein distance on the histograms of the images), and scale-invariant feature transform (SIFT) [Low99, Low04].

EMD measures the difference between two probability distributions, calculating the work it would take to transform one of them into the other. When applied to pixel intensity histograms, it produces a measure of image similarity. If $f_{i,j}$ is the optical flow between clusters p_i and q_j from P and Q respectively and $d_{i,j}$ is the ground distance between them,

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}} \tag{5.1}$$

Lower values of EMD mean more similar images.

SSIM [WBSS04, Ava09] is a measure that originally quantified perceived image degradation when given an original image and its compressed version, but can be used to quantify similarity between any two images. It is calculated on windows of the image. SSIM between two windows x and y—in our implementation, windows of size 7—is a ratio that depends on the windows' averages, variances and the covariance [WSB03]:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{5.2}$$

where μ_x and σ_x^2 are the average and the variance respectively of window x, μ_y and σ_y^2 of y, σ_{xy} is the covariance of x and y, and consonants c_1 and c_2 are computed as follows:

$$c_s = (k_s \times R)^2,$$

R being the dynamic range of the pixel-values.

The final SSIM index is the average over all the windows. Its values range from -1 to 1, 1 standing for identical images, and -1 for an inverted local structure [BZP08].

SIFT [Low99, Low04] is a feature detection algorithm. It describes features as histograms of oriented gradient in a neighborhood. First, a certain number of features are detected and computed. The larger this number is, the lower the threshold local-contrast score is for a feature to be retained. Most weak features in low-contrast regions are filtered out, and edge-like features are preferred. Then the computed features in the source and target images are compared and matched. We used feature matching with 5 k-d trees [Ben80] using FLANN (Fast Library for Approximate Nearest Neighbors) [ML17], which is suitable for nearest neighbor search in large datasets and high-dimensional features. Each match was associated with a distance: the lower this value was, the better the match. If two images depict similar objects, their matches are supposed to be close, which brings in the idea to measure image similarity by the proportion of close matches out of all matches considered—Equation 5.3:

$$\mathrm{SIFT}(x, y) = \frac{|matches(x, y) : match.dist < threshold|}{|matches(x, y)|} \tag{5.3}$$

Before running the comparisons on the entire dataset, I took a small subset of different vocal tract configurations and studied the quality of matches identified by SIFT when adjusting the

image preprocessing features (bilateral filter [TM98] parameters, as explained in Chapter 4.2.1), the number of features and the match distance threshold. The configuration with the fewest noisy matches that still yielded enough informative ones was when the bilateral filter had diameter d of each pixel neighborhood set to 9 (a large filter—I checked values $5, 6, 7, 8, 9$: a larger filter was more sensitive to the global structure of the objects in the image) and when the filter's σ parameter both in the color and coordinate space was reduced to 20 in contrast to 75 from the setup in Chapter 4.2.1—a much more moderate change (I checked σ values $20, 40, 60$ and 80: too much blurring reduced the quality of matches). As for the number of features and their threshold distance, it proved to be a good idea to calculate a moderate number of features, 80 (I checked $5, 20, 35, 50, 65, 95, 110, 125$ and 140: too few features were not informative enough, and too many made them noisier). With this in mind, I set the limit $threshold$ to be 80 (I checked $60, 70, 80, 85, 90$ and 95: being too lenient with the distance threshold lead to false matches, and too strict caused the matches to become trivial or disappear altogether).

Since SIFT measure is a proportion of all matches, its values range from 0 to 1, and the greater the value, the closer two images are.

SIFT transformation is invariant to rotation and feature location in the image. To discount the matches drawn between different articulators, I experimented with a modified version of this measure as well, $SIFT_l$, which is a simplified version of an additional geometric test:

$$SIFT_l(x, y) = \frac{1}{\sum_{n=1}^{M} dist(keypoint_n^{(x)}, keypoint_n^{(y)})} \times \frac{1}{M}, M = |matches(x, y) : match.dist < 70|, \tag{5.4}$$

the idea being that the smaller the displacement is between two matched features, the more probable it is that it corresponds to the same articulator (hence the first ratio), and that image pairs with a lot of strong matches should not be penalized for having too many components in the sum (hence the second ratio).

5.3 Experiments

5.3.1 Temporal behavior

First the chosen measures needed to be validated through studying their temporal behavior: where their extrema were found and whether there were any patterns in the relations between the extrema.

Figures 5.4, 5.5 and 5.6 show an example of the change of the EMD, SSIM, SIFT and $SIFT_l$ values over a sequence "ma ville natale" /ma.vil.na.tal/ (see Figure 5.3 for the spectrogram). In particular, Figure 5.6c shows SIFT and $SIFT_l$, that are based on the same method SIFT, next to each other.

According to our criteria, an ideal similarity measure would have major peaks when comparing the phonemes with the same articulation (/m(a)/ to /m(a)/, /a/ to /a/, /f(i)/ to /v(i)/...), smaller peaks for the same place of articulation (shared between /t(a)/, /n(a)/ and also /l(a)/), moderate increases when comparing any two vowels, and very minor increases when comparing a vowel to a consonant anticipating it.

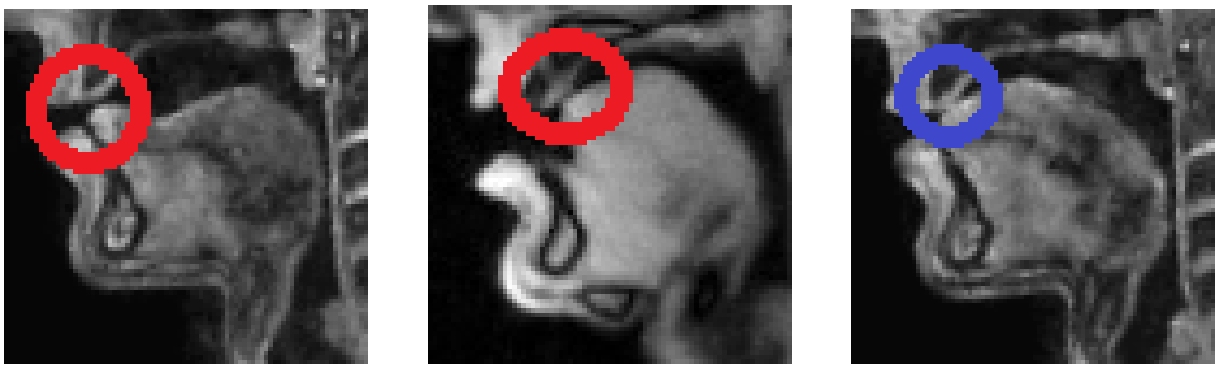
What can be identified from these figures is that EMD comparisons (Figure 5.4) do not seem to be informative, as they are highly correlated and do not depend on articulation.

SSIM (Figure 5.5) consistently peaks on /a/-to-/a/ comparisons; the same holds for /i/, /n(a)/ and /t(a)/. Shapes resembling /f(i)/ are identified at /i/, and also in transition periods, such as from /m/ to /a/ and /l/ to /n/ to /a/. One can also remark that quantitative analysis of SSIM values, just by treating the minimum and maximum values, is less promising than analysis of local extrema, as the relative positioning of each curve rarely changes. In our example, it usually is /t(a)/ on the top and /n(a)/ at the bottom. Meanwhile, the local temporal behavior of each curve seems to be more informative.

The peaks of SIFT(/m(a)/, x), SIFT(/a/, x), SIFT(/f(i)/, x) are consistent with the occurrences of /m(a)/, /a/, /v(i)/ respectively. There is a certain confusion of vowels: the static /i/ resembles the collection of frames of the dynamic /a/; additionally, the shapes of /l(a)/ and /n(a)/ are activated on /t(a)/ as well. A shape resembling /t(a)/ is encountered not only in /t/, but also during the transitions between /m(a)/ and /a/, /i/ and /l/ and /a/ and /l/.

The change of SIFT_l (Equation 5.4) over time reveals that generally the values do not diverge much from 0. Again, there is a certain confusion regarding vowels: shapes resembling /a/ are associated with instances of /i/; of nasal consonants: /m/ is confused for /n/; shapes similar to /l(a)/ appear at /a/; /t(a)/ is correctly identified at /t/, but also during the transition from /v/ to /i/.

(a) A shared place of articulation is a ground to increase the similarity value: static /t(u)/ (left) and a frame of dynamic /s(ã)/ (center) sharing the alveolar region, while different places should not: see the /s(ã)/ and static /k(ɛ)/ with the velar closure (right).



(b) A shared manner of articulation is a ground to very slightly increase the similarity value as an indication of a similar distance at the constriction: static /s(a)/ (left) and a frame of dynamic /ʃ(a)/ (center), being both sibilants, are somewhat similar, while different manners are not: see the /ʃ(a)/ and static /t(a)/ (right).

(c) Coarticulatory effects are a ground to increase the similarity value: static /t(i)/ (left) and a frame of dynamic /i/ (center), whereas a difference in them is not: consider static /t(ɛ)/ (right).

Figure 5.2: Phoneme comparison criteria: what articulation features should contribute as a factor of similarity (left third of the page-center), and how not being shared should lead us to the conclusion that the phonemes are different (right half of the page-center).

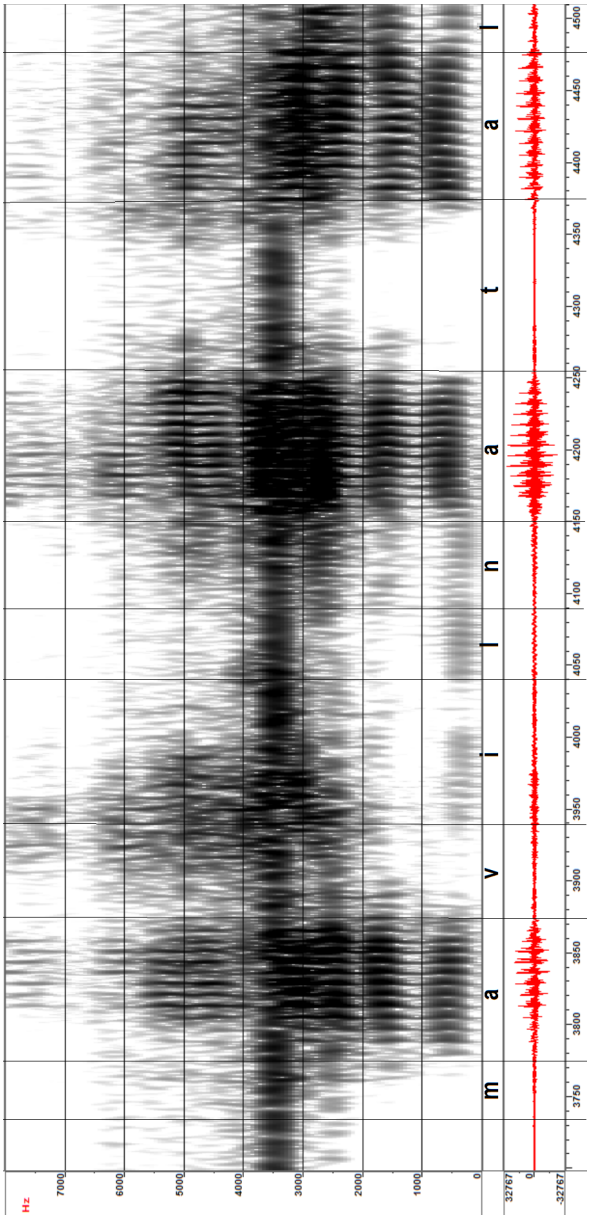


Figure 5.3: The labeled spectrogram of the denoised sequence that was used for the temporal analysis of EMD, SSIM, SIFT and SIFT_l in Figures 5.4–5.6, "ma ville natale" /ma.vil.na.tal/.

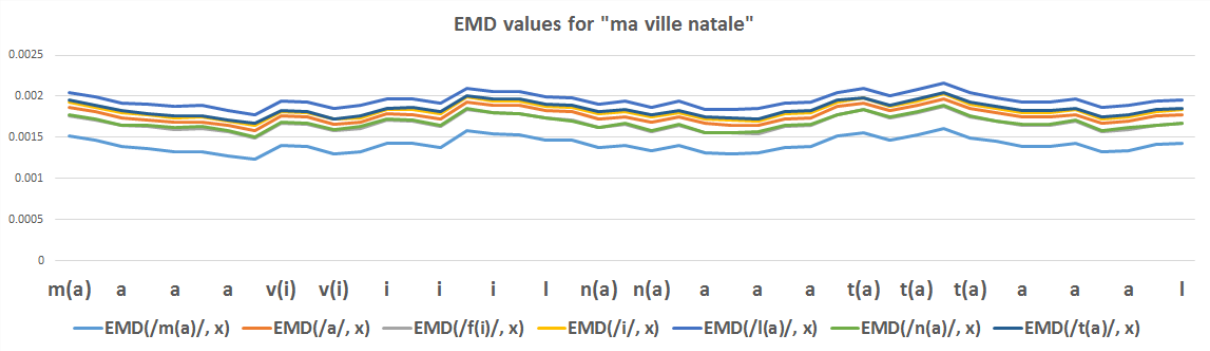


Figure 5.4: The change of EMD—Earth mover's distance—values over time when comparing the static frames /m(a)/, /a/, /f(i)/, /i/, /l(a)/ and /t(a)/ to the same sequence as in Figure 5.3, "ma ville natale" /ma.vil.na.tal/. The lower the value, the more similar the frames are considered to be. The signals are strongly correlated and do not seem to be informative.

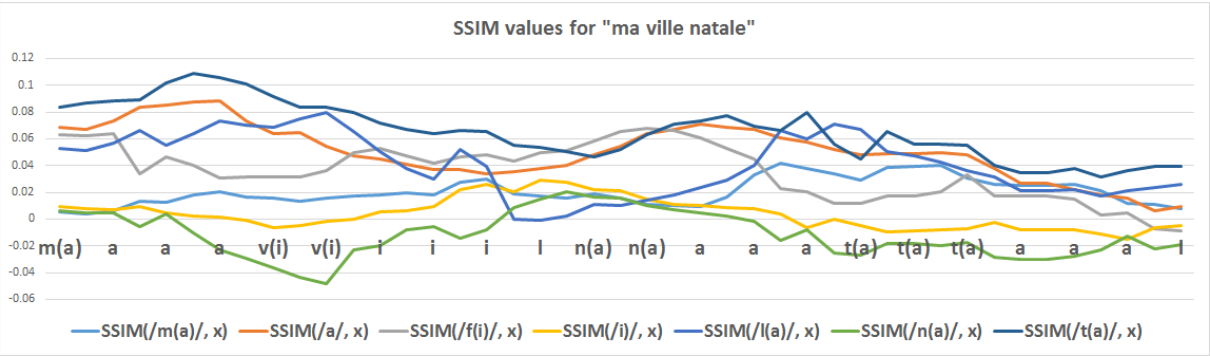


Figure 5.5: The change of SSIM—structural similarity—values over time when comparing the static frames /m(a)/, /a/, /f(i)/, /i/, /l(a)/ and /t(a)/ to the same sequence as in Figure 5.3, "ma ville natale" /ma.vil.na.tal/. The greater the value, the more similar the frames are considered to be. The peaks of SSIM(/a/, x) are consistent with the occurrences of /a/; the same holds for /i/, /n(a)/ and /t(a)/. /m(a)/, /f(i)/ and /l(a)/ do not follow the labeling.

Figure 5.7 shows the similarity measures between seven static images of /f/ (/f(i), f(ɛ), f(a), f(o), f(u), f(y), f(ø)/) and the consecutive images of the dynamic sequence /fi, fe, fɛ, fa, fɔʁ, fo, fu, fy, fø, fœʁ, fɑ̃, fɔ̃, fɛ̃/.

One can see that, just like in Figure 5.4, the values of EMD in Figure 5.7a are strongly correlated and do not seem to capture any articulatorily relevant information.

Same as in Figure 5.5, Figure 5.7b demonstrates that the ranking of SSIM values when comparing a particular set of static images to the real-time sequence typically stays quite stable. In our example, the values of SSIM(/f(a)/, x) are almost always the lowest, and SSIM(/f(o)/, x) the highest. The shapes most resembling the static samples of /f/ appear in the transitions from /f/ to vowels. Distinguishing the vocalic anticipatory coarticulation is impossible.

Figure 5.7c shows that SIFT values anticipate the beginning of /f/, the peaks occurring right before the beginning of the phoneme. Again, distinguishing the vocalic anticipatory coarticula-

tion is impossible.

Figure 5.7d, like Figure 5.6b, shows that most SIFT_l values are very close to 0. The spikes do not seem to be very informative, though it can be said that they generally occur at the end of the fricative and can last through the vowel and the subsequent beginning of silence. Minor peaks follow the behavior of SIFT: they pre-empt the beginning of /f/. Distinguishing the vocalic anticipatory coarticulation is impossible.

Overall, the analysis of the measure's temporal behavior in the examples above leads to the following conclusions:

- EMD does reflect some changes in the image sequences, but they are not specific to any articulatory configuration. Likely, the static and dynamic images needed to be much more similar to each other for EMD to be informative due to the optical flow assumptions that a contour corresponding to an object in the source image can only slightly move in the target image, while our images are both dissimilar in terms of image quality and representing differently shaped contours.

- In terms of interpretability, the performance of SSIM and SIFT is quite similar. Both do encode some articulatory information: we can identify both vowels and consonants. Above we discussed mismatches as, for example, according to SIFT static /l(a)/ and /n(a)/ were similar not only to their dynamic counterparts (taken as entire collections of frames spanning over the occurrences of a particular phonetic label, /l(a)/ and /n(a)/ in this case) but also to /t(a)/. Such mismatches and similar demonstrate that in the case of consonants SIFT is rather sensitive to the place of articulation (where the constriction occurs), which makes it better at getting consonants right but creates difficulties to identify how open or close and front or back a vowel is, and SSIM is better at capturing the general positioning of the articulators, which makes it better at dealing with vowels and anticipating a certain vocalic context.

- The concerns regarding how well the static dataset models the real-time one seem to be justified: the especially problematic effects are those that we expected in point **(3)** at the beginning of this chapter (Chapter 5.2.1): the production of nasals and liquids and realistic anticipatory coarticulation. Likewise, point **(4)** about the poor coverage of the variety of shapes occurring in speech production is a valid concern, considering the frequent mismatches at the transition periods between any two phonemes.

- Adding the feature match location into SIFT to create SIFT_l (Equation 5.4) does have some merit, but the interpretability of this measure is severely affected by the unevenness of its peaks: in comparison to the very strong spikes that do not actually contain much articulatory information, the moderate peaks have the potential to be useful when merged with other values.

It should be noted that the jerks in the measures' plots occur more frequently and are greater in magnitude than the changes in actual articulation can occur. Therefore we find it imprudent to look for patterns one image at a time, for example, at the image or images in the center

of a given phoneme, like in the approach of [LSNQ18]. Instead, we resolve ourselves to using averaging and looking for general patterns.

5.3.2 Distributions and correlations

All the measures reflected that the static dataset, being done with speaker S_A, resembled S_A's part of the real-time one better than that of S_B (Figure 5.8): in case of EMD, the mass of histogram for S_A compared to S_B is shifted to the left, and in case of the rest of the measures, to the right. The only exception to this is SSIM and its S_A-S_B distributions (Figure 5.8, middle pair): in the frames of S_B, SSIM identified more images that had a similarity value higher than the mode than in those of S_A.

When breaking down by phonemes, this pattern of S_A's frames being closer to the static dataset than those of S_B are holds. The shapes of the distributions, however, stay the same—see Figure 5.9.

The calculated distances were aggregated by speakers, by phonemes, by phonemes in vocalic context (what vowel V is anticipated in the dynamic dataset according to the phonetic labeling), by the phoneme's phonetic classes and by speaking styles (spontaneous speech or not). To reduce the memory load, it was done in 10 randomly split blocks, 6 times over.

Table 5.2 shows the mean and standard deviation of each metric across the entire volume of speech by speakers S_A and S_B. Overall the relationship between the measures is not strong: see correlations in Table 5.3.

	E(EMD) ± SD(EMD)	E(SSIM) ± SD(SSIM)	E(SIFT) ± SD(SIFT)	E($SIFT_l$) ± SD($SIFT_l$)
S_A	0.0016 ± 0.0003	0.0350 ± 0.0331	0.0665 ± 0.0317	0.0115 ± 0.0989
S_B	0.0018 ± 0.0002	0.0367 ± 0.0297	0.0613 ± 0.0299	0.0080 ± 0.0346

Table 5.2: Means (E) and standard deviations (SD) of the image similarity measures, speakers S_A and S_B.

Measure	EMD		SSIM		SIFT		$SIFT_l$	
	S_A	S_B	S_A	S_B	S_A	S_B	S_A	S_B
EMD			0.42	-0.02	0.07	0.00	0.01	-0.01
SSIM					0.08	0.06	0.01	0.00
SIFT							-0.09	-0.25

Table 5.3: Correlations between EMD, SSIM, SIFT and $SIFT_l$ values across the entire set of image comparisons. Between any two measures not involving EMD, a positive correlation is indicative of an agreement in judgment. With EMD, it is the reverse since EMD is lower for similar images while the others are higher.

None of the similarity measures followed the normal distribution:

- EMD:

  - Shapiro-Wilk's test (ShW): statistic 0.963 for S_A and 0.975 for S_B, p-value 0.000 for both,
  - D'Agostino's K^2 test (DA): $131,032.833$ for S_A, $218,447.082$ for S_B, p-value 0.000;

- SSIM:
  - Shapiro-Wilk's test (ShW): statistic 0.999 for S_A and 0.998 for S_B, p-value 0.000 for both,
  - D'Agostino's K^2 test (DA): $2,272.411$ for S_A, 4492.322 for S_B, p-value 0.000;

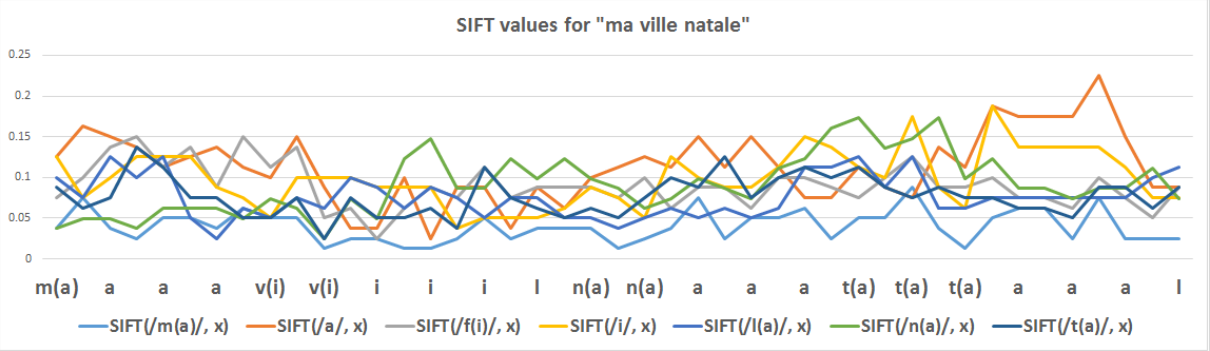
- SIFT:
  - Shapiro-Wilk's test (ShW): statistic 0.966 for S_A and 0.977 for S_B, p-value 0.000 for both,
  - D'Agostino's K^2 test (DA): $107,402.547$ for S_A, $108,058.162$ for S_B, p-value 0.000;

- $SIFT_l$:
  - Shapiro-Wilk's test (ShW): statistic 0.039 for S_A and 0.135 for S_B, p-value 0.000 for both,
  - D'Agostino's K^2 test (DA): $13,912,105.949$ for S_A, $11,653,781.164$ for S_B, p-value 0.000;

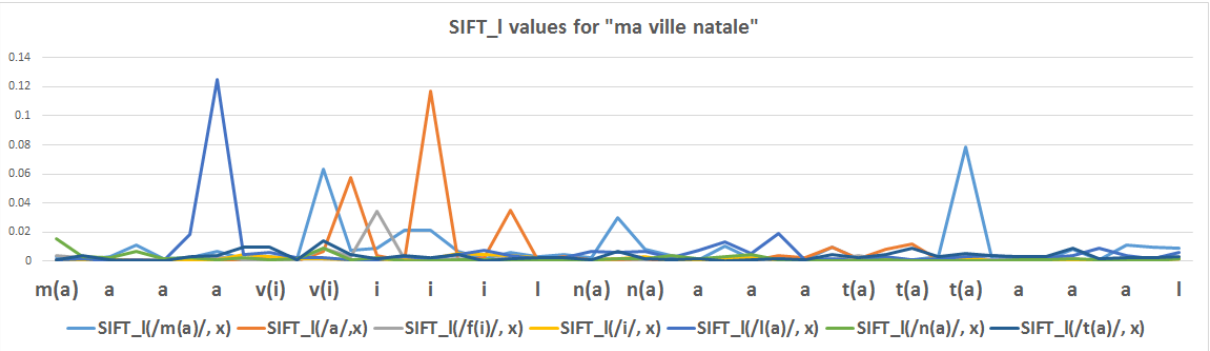
Thus, in order to determine whether the observed differences between phonemes are statistically significant, I needed to use non-parametric tests. The samples made by different measures were assumed independent.

5.3.3 Analyzing the speakers

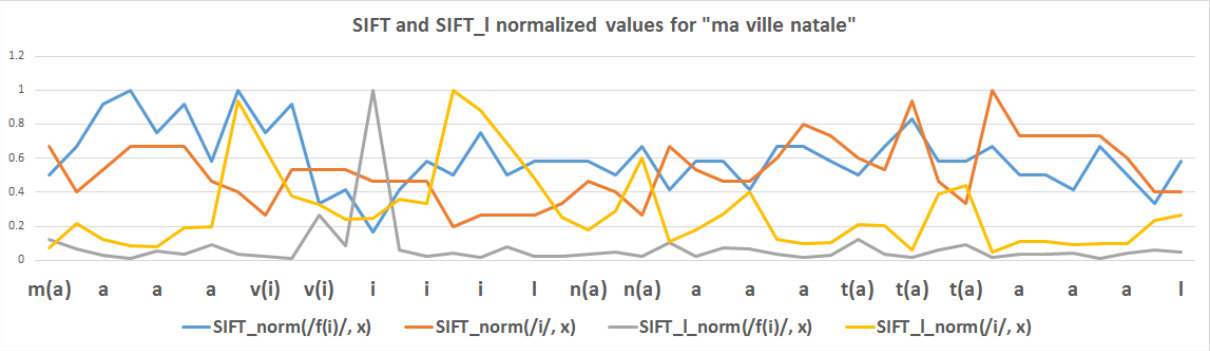
For every measure, its distribution for speaker S_A is different from the one for S_B are different; furthermore, quite obviously, the distributions of any two measures for a given speaker are different: see Table 5.4 (Kolmogorov-Smirnov and Mann-Whitney tests).



(a) The change of SIFT (Equation 5.3) over time. The peaks of SIFT$(/\mathrm{m(a)}/, \mathrm{x})$, SIFT$(/\mathrm{a}/, \mathrm{x})$, SIFT$(/\mathrm{f(i)}/, \mathrm{x})$ are consistent with the occurrences of /m(a)/, /a/, /v(i)/ respectively. The static /i/ resembles the collection of frames of the dynamic /a/. /l(a)/ and /n(a)/ are activated on /t(a)/ as well. A shape resembling /t(a)/ is encountered not only in /t/, but also during the transitions between /m(a)/ and /a/, /i/ and /l/ and /a/ and /l/.

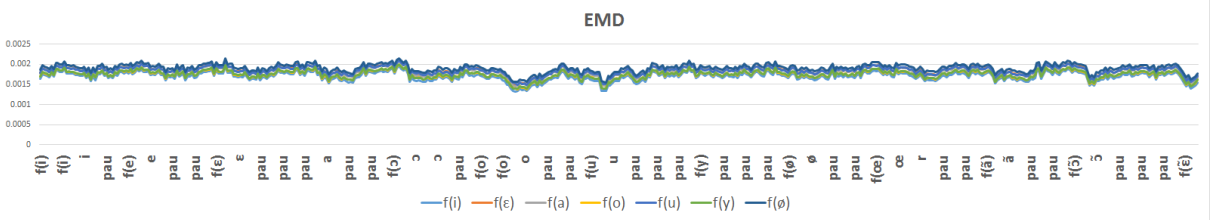


(b) The change of SIFT$_\mathrm{l}$ (Equation 5.4) over time. Most of the time the values are close to 0. Shapes resembling /a/ are associated with instances of /i/; /m/ is confused for /n/; shapes similar to /l(a)/ appear at /a/; /t(a)/ is correctly identified at /t/, but also during the transition from /v/ to /i/.

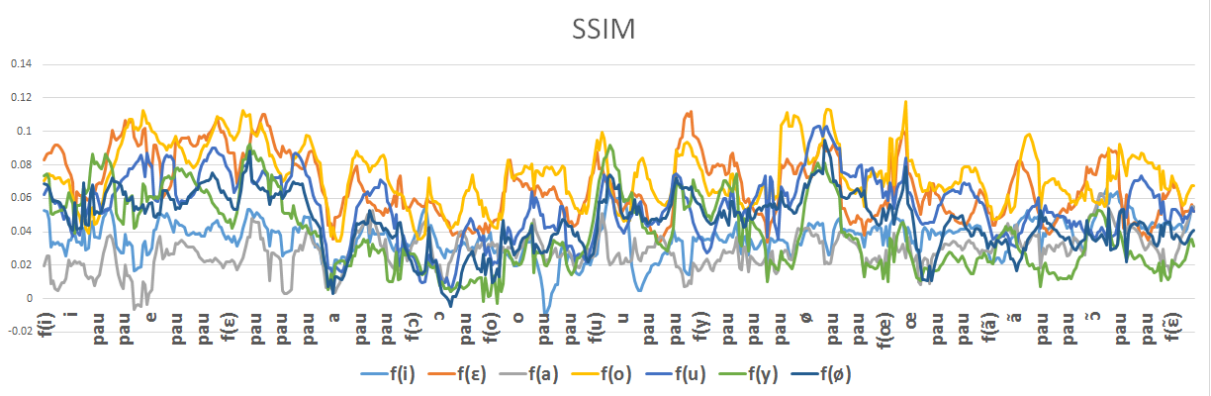


(c) The comparison of SIFT and SIFT$_\mathrm{l}$ measures: SIFT identifies the similarity between /f(i)/ and /v(i)/. SIFT$_\mathrm{l}$ does, too, but it demonstrates a higher peak already in the section of /i/. Both SIFT and SIFT$_\mathrm{l}$ find that the static /i/ resembles the collection of frames of the dynamic /a/ (/i/-like shapes are also found during the transitioning time, for example, between /a/ and /t/; the top peak of SIFT$_\mathrm{l}(/\mathrm{i}/, \mathrm{x})$ is, indeed, at /i/).

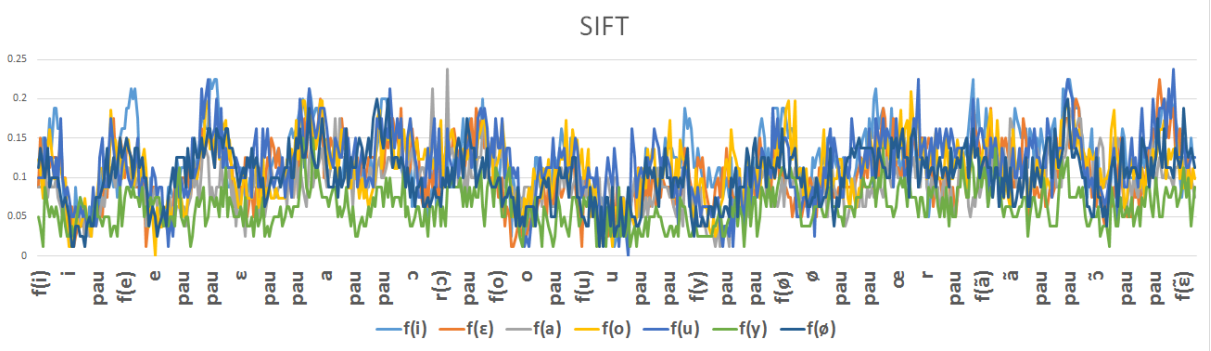
Figure 5.6: The change of SIFT (scale-invariant feature transform)-based SIFT and SIFT$_\mathrm{l}$ values in time when comparing the static frames /m(a)/, /a/, /f(i)/, /i/, /l(a)/ and /t(a)/ to the same sequence as in Figure 5.3, "ma ville natale" /ma.vil.na.tal/. An ideal similarity measure would have major peaks when comparing the phonemes with the same articulation (/m(a)/ to /m(a)/, /a/ to /a/, /f(i)/ to /v(i)/...), smaller peaks for the same place of articulation (shared between /t(a)/, /n(a)/ and also /l(a)/), moderate increases when comparing any two vowels, and very minor increases when comparing a vowel to a consonant anticipating it.



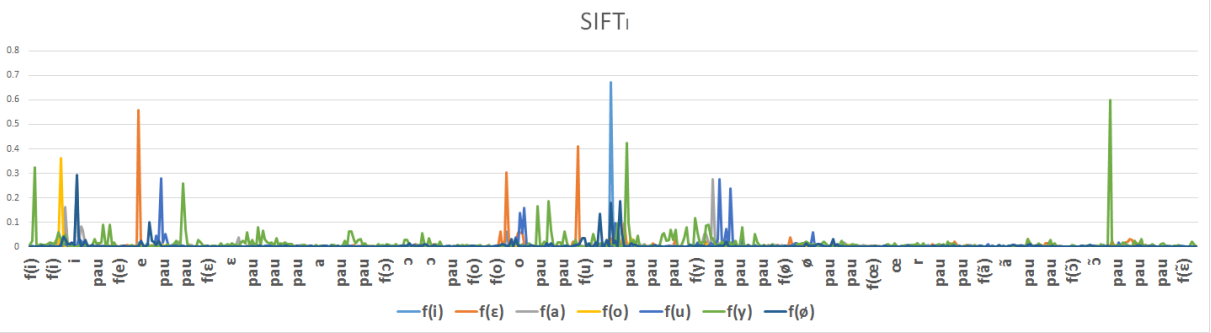
(a) The temporal behavior of EMD over the course of the sequence. The lower the value, the more similar the images. No evidence of articulatory information.



(b) The temporal behavior of SSIM over the course of the sequence. The higher the value, the more similar the images. The shapes most resembling the static samples of /f/ appear at the transitions from /f/ to vowels.



(c) The temporal behavior of SIFT over the course of the sequence. The higher the value, the more similar the images. The values anticipate the beginning of /f/, the peaks occurring right before the beginning of the phoneme.



(d) The temporal behavior of $SIFT_l$ over the course of the sequence. The higher the value, the more similar the images. The spikes generally occur at the end of /f/ and can last through the vowel and the subsequent beginning of silence. Minor peaks follow the behavior of SIFT (Figure 5.7c): they pre-empt the beginning of /f/.

Figure 5.7: The global view of the plots of the similarity measures between seven static images of /f/ (/f(i), f(ɛ), f(a), f(o), f(u), f(y), f(ø)/) and the consecutive images of the dynamic sequence /fi, fe, fɛ, fa, fɔʁ, fo, fu, fy, fø, fœʁ, fɑ̃, fɔ̃, fɛ̃/.

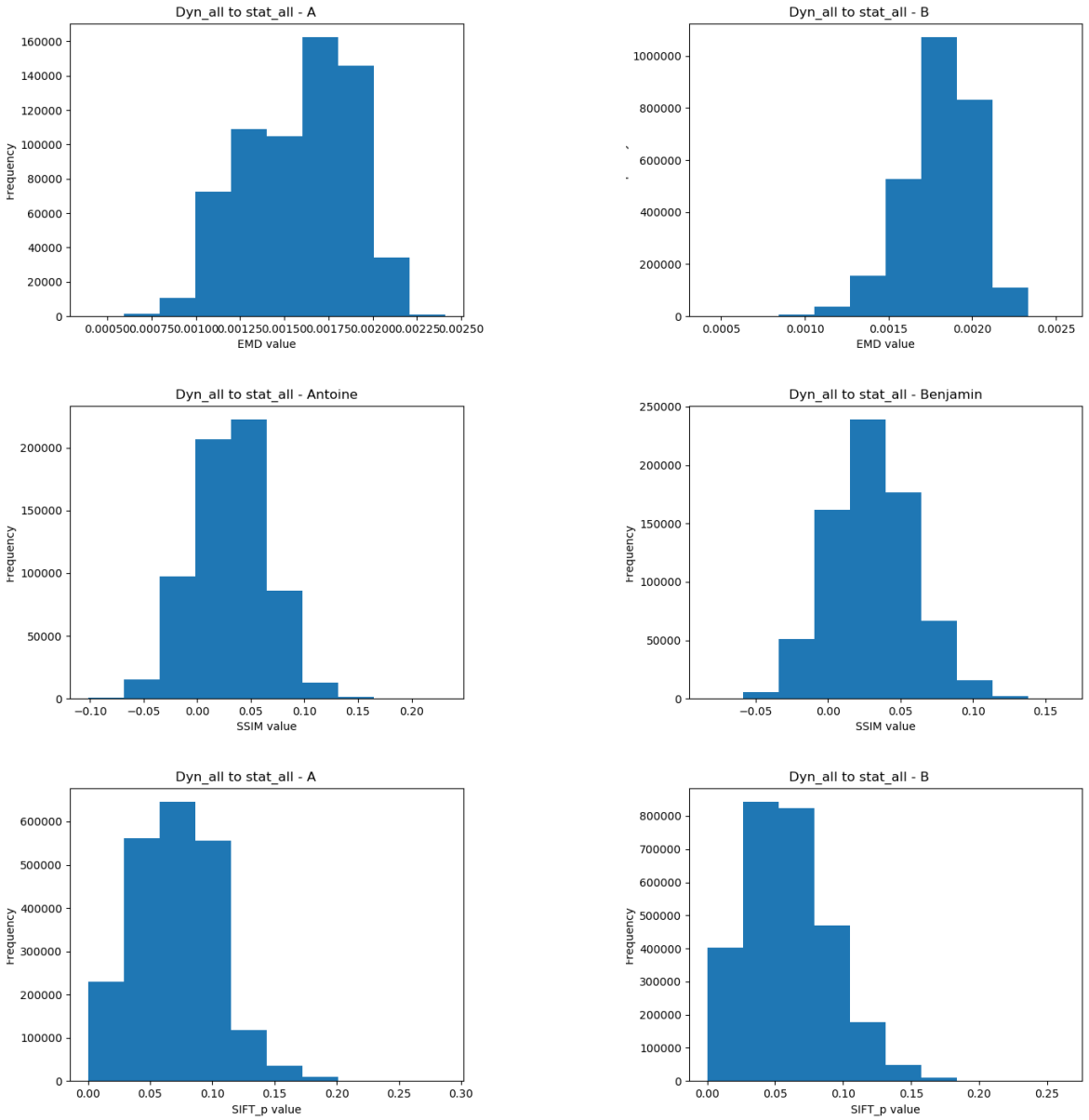


Figure 5.8: The overall distribution of the EMD measure for the distance between all the dynamic and all the static captures for speakers S_A (left) and S_B (right) (the pair at the top); SSIM (middle pair); and SIFT (bottom). One can see that the static dataset was closer to S_A's part of the dynamic one. The overall shapes of the distributions are similar, except that SSIM has more image pairs whose similarity was greater than the mode for S_B than for S_A.

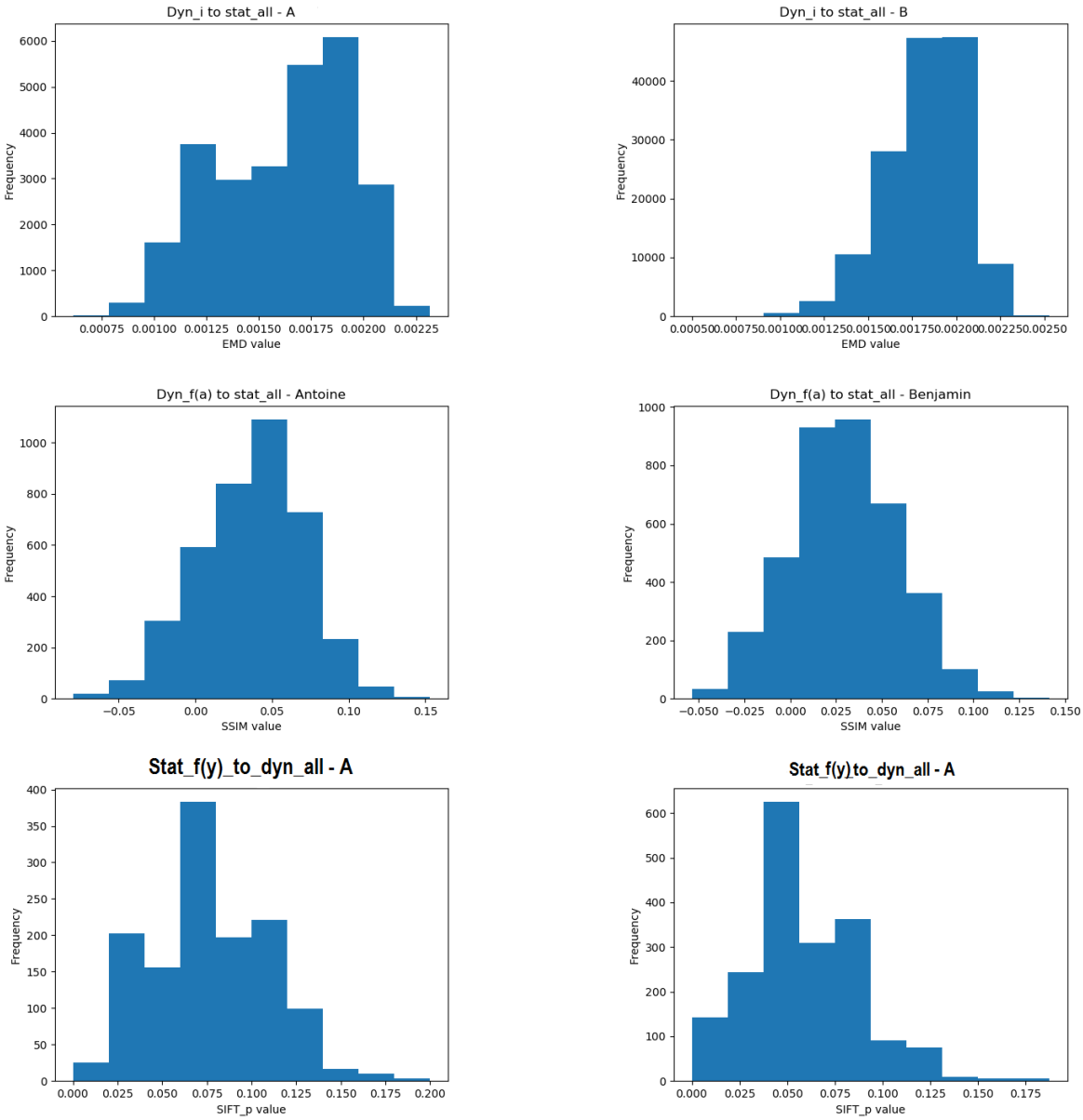


Figure 5.9: The overall distribution of the EMD measure for the distance between all the dynamic captures labeled as /i/ and the bulk of static captures for speakers S_A (left) and S_B (right) (the pair at the top); SSIM, the dynamic /f(a)/ and the entire corpus of static captures (middle pair); and SIFT, /f(y)/ (the bottom). One can see that the static dataset was closer to S_A's part of the dynamic one.

Measure	Test	EMD		SSIM		SIFT		$SIFT_l$	
		S_A	S_B	S_A	S_B	S_A	S_B	S_A	S_B
EMD	KS	0.874795		0.837491	0.873411	0.993261	0.989590	0.493535	0.443662
	MW	7×10^{11}		8×10^{11}	9×10^{11}	3×10^{10}	8×10^{10}	2×10^{12}	4×10^{12}
SSIM	KS			0.659491		0.440215	0.382946	0.637211	0.691701
	MW			1×10^{12}		1×10^{12}	2×10^{12}	1×10^{12}	1×10^{12}
SIFT	KS					0.993261		0.896325	0.894488
	MW					7×10^{10}		2×10^{11}	3×10^{11}
$SIFT_l$	KS							0.482467	
	MW							3×10^{13}	

Table 5.4: For S_A and for S_B, each measure's distribution is different from the others, and for each measure, the distributions for S_A and S_B are different: Kolmogorov-Smirnov (KS) and Mann-Whitney (*MW*) equality tests. p-value is 0.000.

Statistic \ Measure	EMD		SSIM		SIFT		$SIFT_l$	
	S_A	S_B	S_A	S_B	S_A	S_B	S_A	S_B
KS	0.298121	0.331552	0.206611	0.213495	0.068425	0.049031	0.005574	0.037007
MW	3×10^{11}	5×10^{11}	4×10^{11}	6×10^{11}	5×10^{11}	8×10^{11}	5×10^{11}	9×10^{11}

Table 5.5: Verifying whether the distribution of each measure is different or not when comparing the static images to the spontaneous-speech real-time sequences or the other equality tests: Kolmogorov-Smirnov (KS) and Mann-Whitney (*MW*). p-value is 0.000.

Speaking styles affect the distributions: spontaneous speech is found to be different from non-spontaneous (Table 5.5, Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) equality tests, $\alpha = 0.05$).

5.3.4 Phoneme comparisons

To recapitulate, I had the similarity values for each of the 95 MRI captures compared to each of the $368,848$ RT-MRI frames, according to four similarity measures: EMD, SSIM, SIFT and SIFT_l. Each capture, MRI and RT-MRI alike, corresponded to a phonetic label, from which I could extract the phoneme in production as well as its anticipated vowel, where applicable.

So, for any phoneme pair encountered, I aggregated all instances of their comparisons in the corpus.

For example, when studying a measure sim on the static MRI capture of /a/ and any instances of /ɔ/ in the RT-MRI corpus, if we have twenty frames of /ɔ/ in the first recording, five in the third and five in the tenth, it would provide thirty values of sim. Or, when studying phonemes in context, I could find the comparison values for the static MRI capture of, for example, /k(i)/ and all the dynamic instances of, say, /t(a)/.

Every phoneme was associated to a list of its phonetic classes, such as *vowel* and *consonant*, and, more specifically, *oral vowel*, *nasal vowel*, *back vowel*, *central vowel*, etc. or *labial*, *alveolar*, etc. and *plosive*, *fricative*, etc. This way the aggregation could be carried out not only over all separate phonemes, but also over the phonetic classes. To stay within the range of comparable features, the phonetic classes were divided into the categories of place and manner of articulation, nasality and lip roundedness.

Once all comparison values for a particular pair of phonemes, coarticulated phonemes or phonetic classes were gathered, and I calculated their average and standard deviation. Note that it was done in 10 randomly selected blocks, as in 10-fold cross-validation tests, to reduce the memory load and repeated 6 times to make sure the conclusions were not due to the randomness of the data split. Additionally, I tested V (vowel) and C(V) (consonant C anticipating a vowel V) distributions for being different.

The averages and standard deviations were then used to identify which static phonemes, phonemes in context or phonetic classes were the closest to each phoneme, phoneme in context or phonetic class encountered in the dynamic corpus, and vice versa, to what dynamic phonemes, etc. the static ones came the closest. This part of the work is presented below in section 5.3.4— "Matching static and real-time phonemes, phonemes in context and phonetic classes".

The distributions were used to verify the presence of coarticulatory effects. This part of the work, too, is presented below, in section 5.3.4: "Testing for phoneme-specific distribution equality".

Testing for phoneme-specific distribution equality

So, in order to do the following:

- To see whether we can see the same distributions for phonemes differing only in voicing, e.g. /p/ and /b/;
- To identify differences caused by coarticulation, e.g. /p(a)/ and /p(ã)/;

- To tell the distributions of vowels apart,

I tested the distributions formed by all comparisons of V's and $C(V)$'s with a same consonant C for being different with Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) equality tests, $\alpha = 0.05$.

When comparing the distributions that apeared from the comparison between all rtMRI frames labeled as a given vowel and the bulk of static MRI frames to the similar distribution but for another vowel, all measures were able to distinguish the distributions of almost all vowels. There were instances where a vowel pair would produce two indistinguishable measure distributions. This does not disqualify the result as most of these vowel pairs are rather close phonetically (for example, /o/ and /u/); however, since our ideal measure would pick up on phonological differences as well, such indistinguishable results are important to consider. Specifically, EMD could not distinguish /a/ and /ɔ/ for S_A, but was able to distinguish all vowels for S_B; SSIM could not distinguish /o/ and /u/ for S_B; SIFT /e/ and /ɔ̃/, /ɛ̃/ and /œ/, /ɛ̃/ and /ɛ/ for S_A, /ɑ̃/ and /ɔ/, /o/ and /ɛ/, /o/ and /ə/, /œ/ and /œ̃/ for S_B. $SIFT_l$ was the most indiscriminate of them all: the distributions of /ɑ̃/ and /œ/, /a/ and /œ/, /ɔ̃/ and /ə/, /u/ and /y/ for S_A and /e/ and /y/, /ɛ̃/ and /o/, /ɛ̃/ and /œ̃/ were found equal. Some confusions were frequent across data splits, some were not. It should also be noted that almost all of these pairs are cases when at least one vowel of the pair is relatively infrequent in the data.

When comparing distributions appearing from calculating a similarity measure between a static image depicting a certain vowel or phoneme and the entire real-time dataset, there were almost no instances when they would be equal. The exact list of indistinguishable distributions varied with data splits, but usually there were fewer than five pairs per measure per speaker. For example, EMD could not distinguish /ʁ(a)/ from /ʁ(ɑ̃)/ for S_A, /ʁ(ɑ̃)/ from /ʁ(ɛ)/ for S_B, SSIM /k(ɛ)/ from /k(ɔ̃)/, /k(a)/ from /k(u)/ for S_B, SIFT /ʃ(o)/ from /ʃ(u)/, /ʃ(a)/ from /ʃ(u)/, /ʁ(ɑ̃)/ from /ʁ(i)/, /t(ɛ)/ from /t(i)/ for S_A/ and /n(ɛ)/ from /n(i)/, /ʁ(u)/ from /ʁ(y)/ for S_B, and, finally, $SIFT_l$ /s(ø)/ from /s(u)/, /s(i)/ from /s(u)/, /ʃ(a)/ from /ʃ(ɛ)/, /ʃ(ɛ)/ from /ʃ(y)/, /l(e)/ from /l(o)/ for S_A and /k(i)/ from /k(y)/, /k(i)/ from /k(ɔ̃)/, /ʃ(a)/ from /ʃ(ɛ)/, /ʁ(ɛ)/ from /ʁ(ɔ̃)/ for S_B, etc. To summarize, typically the indistinguishable distributions concern the phonemes /s/, /ʃ/, /ʁ/, /l/, which together with the nasals are on the list of the phonemes that were suspected to be problematic to capture in a static setting since the very beginning, as the aerodynamics of producing a fricative is quite crucial to adjust its articulation and therefore the static acquisition with its silent articulation was at a disadvantage. This is an indication of the fact that, first, for most of the static images there is no evidence for them to be produced incorrectly or inconsistently, since their distributions appear to be unequal; and second, some sounds are certainly difficult to sustain for an extended time in an MRI machine.

Matching static and real-time phonemes and phonemes in context

As each of our measures has a relation of order, i.e. any two phonemes can be considered more or less similar according to it (in SIFT, $SIFT_l$ and SSIM, the greater the value, the more similar the phonemes; in EMD, the reverse), that meant that for any static capture I had the means to see what phonemes were considered to be the most similar to it, on average, according to each of the measures. This direction of analysis is encoded as "static to dynamic" in the figures and text of this chapter. And vice versa, for any label encountered in the dynamic corpus, I could identify the static capture that came to be the closest on average (dynamic to static).

The exact list of suggestions varied with every program run, but repetitions appeared consistently. I analyzed the two most frequent top matches across all the data splits, considering the following factors to be of influence:

- Whether the most frequent top matches across the cross-fold checks were correct;
- How large was the variance of the particular phoneme pair comparison—with respect to the average value;
- How various the top matches across the cross-fold checks were: if, according to a measure sim, the closest phoneme to ph changes with every data split, then having a match of ph_1 and ph_2 does not bear the same importance as consistently finding ph_2 to be the closest to ph_1.

The behavior of similarity measures differed in the following two scenarios: (**a**) *stat-to-dyn*: compare a given static MRI image depicting a consonant-vowel /C(V)/ or vowel /V/ blocked articulation to all the RT-MRI images, broken down by their coarticulated phoneme labels (/C(V)/ or /V/ again), or compare a set of static MRI images depicting the same consonant /C/ in different vocalic contexts to all RT-MRI images as per their phoneme labels, with no coarticulation information; and (**b**) the other direction *dyn-to-stat*: compare the bunch of dynamic images labeled as a particular phoneme[17], with coarticulation or not, to static images, with coarticulation or not.

When performing image comparisons in the *dyn-to-stat* direction, EMD consistently lists /o/ and /ɔ/ as the best candidates for any dynamic image when discarding coarticulatory information, and /o/, /ɔ/ or /m(a)/ when keeping it in. This holds for both of the speakers. Its implication is that, on average, the statistics of the RT-MRI frame dataset matched most closely those of these three static MRI captures.

In the reverse direction, *stat-to-dyn*, EMD is very noisy and inconsistent when dealing with coarticulated phonemes (/C(V)/). When disregarding coarticulation, the best matches start repeating over different data splits, and some of them become right, but this improvement is not dramatic and the matches stay sporadic.

We can thus conclude that there is little hope to capturing articulatory information with the EMD measure, because of the datasets being too different.

SSIM also has a problem of giving same best matches for all dynamic phonemes with no coarticulation: in the *dyn-to-stat* direction, in the case of sounds with little or no obstruction (vowels, approximants) /ɛ/, /ɔ/ and /a/ are the most frequent matches for $\mathrm{S_A}$, and /u/, /w/ and /e/ for $\mathrm{S_B}$. When additionally breaking down by anticipated vowels, I additionally get /ʁ/ and /f/ in different vocalic contexts that indiscriminately appear as top frequent matches for $\mathrm{S_A}$, and /s/, /ʃ/, /p/ for $\mathrm{S_B}$. This could be due to the fact that in smaller groupings, broken down by coarticulation, the impact of articulatory transitions appearing in RT-MRI but labeled just the same, as vowels or consonants anticipating a vowel, becomes large enough to get a high SSIM score on average (or it could be indicative of incorrect timing in the phonetic labels). As further evidence of this, when comparing RT-MRI vowel frames to static images, I can identify some relation between the original vowels and the matches' anticipated vowels: for example,

[17] I will call such image sets "dynamic images" and similarly hereon, even though I do not mean any particular RT-MRI frame; the approach is to generalize over all the frames that received a particular label

the most frequent best matches for /ɛ/ are /ʁ(ɛ)/, /l(e)/; for /ɑ̃/ the best matches are /ʁ(ɛ)/, /ʁ(ɑ̃)/; for /k(i)/ it is /k(i)/ or /l(i)/. This means that the openness of the anticipated vowel is usually correct. The feature of nasality is not captured in the comparisons in this direction: SSIM matches /f(ɛ)/ and /p(o)/ to the dynamic images of /m(ɛ)/; or, another example, /p(y)/, /l(y)/, /p(i)/ to the dynamic images of /m(y)/.

As for **stat-to-dyn** SSIM, in most cases its best match for sounds with little or no obstruction will indiscriminately be /fi/, /ø/ or /u/ for S_A and /fi/, /o/ and /w/ for S_B—sounds that fall into the same category, but are no precise matches (both /u/ and /o/ are back rounded vowels, and the difference between them is that /u/ is close and /o/ is close-mid; /w/ is a labial approximant, the closest consonantal equivalent of /u/). It both disqualifies SSIM from being used as evidence of the static frames validity in this case, and, regarding the consistency and similarity across speakers (labial approximants and close-mid rounded vowels), does show that some articulatory information is indeed captured. The precision increases for fricatives, both with vocalic context and not (/f, v, z, s, ʒ, ʃ, ʁ/—appearances of correct matches are numerous and frequent over different data splits), and even more so in the case of plosives: both /k/ and /g/ get matches of /k/ and /g/, /p, b/ are mostly matched to /b/ in case of S_A or /m/ in case of S_B, which captures the labial articulation, but raises questions about voicing and nasality. /t, d/ get matches of fricatives with close places of articulation: /f, v/, /s, z/, /ʃ, ʒ/. When applied to nasal stops, however, almost all **stat-to-dyn** SSIM's top frequent matches involve nasality: for example, for /m(ɔ̃)/, they are /b(e) (same place of articulation) or /v(ɔ̃)/ for S_A, or /n(e) or /m(e)/ for S_B, or for /n(u)/, it is /j(u)/ (close place of articulation, same vocalic anticipation) or /b(ɛ̃)/ (nasality) for S_A and /j(u)/ or /m(u)/ (nasality, same vocalic anticipation) for S_B.

What can be concluded is that with SSIM it is easier to find the RT-MRI images that are most similar to a given static image or a set of them rather than vice versa, to find the best static image to explain a collection of RT-MRI images. When the articulators are close enough, SSIM begins to capture vocalic context of consonants, and when there is a contact such as in plosives, SSIM is capable of identifying the place of articulation. The asymmetry between SSIM's treatment of nasality—no handling nasality in the *dyn-to-stat* direction, capturing it in the *stat-to-dyn* direction and even favoring nasal matches for some oral consonants—is an indicator that the concern regarding the correctness of nasalization in the static dataset may be quite justified: the oral sounds may be produced as if slightly nasalized because of the lack of tension in the articulators, and in the simulation of nasal sounds the velopharyngeal port may be not open enough because there is no actual speech production occurring during the static MRI acquisition.

In the *dyn-to-stat* direction, SIFT's matches of vowels with no regard of coarticulation are rather close: its best candidates for /a/ are /a/ and /e/ (both are front and unrounded) for S_A and only /a/ for S_B; for /e/, /e/ and /i/ (both are front and unrounded) for S_A and /a/ and /e/ for S_B; for /i/, /e/ and /i/ (both are front and unrounded) for S_A and /y/ (also close and front as /i/, but rounded) for S_B; for /u/, /w/ for both speakers, /w/ being the closest consonant to /u/; for /y/, /e/ and /w/ for S_A, /y/ and /w/ for S_B. When I break the cases of RT-MRI and static MRI consonants by the vowels they anticipate, the top frequent best matches of vowels start including consonants, with the same effects as in the similar scenario of SSIM: there is a relation between the (dynamic) vowel and the vowel anticipated by the (static) consonant. For example, SIFT matches /a/ and /k(a)/ to the dynamic /a/, or /w(a)/, /t(o)/ and /p(o)/ to the dynamic /o/. The explanation could be the same as in SSIM: either this is the influence of the

transition times, or it is an indicator of temporal shifts in the phonetic labeling.

In the *stat-to-dyn* direction, SIFT's matches of vowels with no regard of coarticulation are rather oftentimes nasal: the static /a/ is matched to /ø/ or /ɛ̃/ of S_A or to /œ̃/ or /ø/ of S_B; /ø/ to /ø/ or /œ/ for S_A and /ø/ for S_B; /ɛ/ to /œ̃/ or /ɛ̃/ for S_A and /œ̃/ for S_B; /œ/ to /ø/ or /œ̃/ of S_A or /œ̃/ or /ã/ of S_B. Just like it was with SSIM of the same scenario, /ɦ/, /w/ and /u/ are unusually frequent and indiscriminate matches as well.

When applied to nasal vowels and consonants, *dyn-to-stat* SIFT is not capable of matching them to the nasal vowels in the static dataset: all its matches are oral. As for the *st-to-dyn* direction, SIFT does pick either nasal RT-MRI phonemes or phonemes with velar articulation, e.g. according to SIFT, if coarticulation is taken into account, the closest matches to the static /ã/ are /b(ã)/ or /v(ã)/ for S_A and /r(o)/ for S_B; and with no coarticulation, no consonants appear among the matches, and the matches are correct: for example, the static /ɛ̃/ gets matched to the dynamic /ɛ̃/ or /œ̃/ for S_A or /œ̃/ or /ɛ̃/ for S_B.

In fricatives, with no regard for anticipated vowels both *dyn-to-stat* and, to a lesser extent, *stat-to-dyn* SIFT matches dynamic (static) fricatives to static (dynamic) vowels. *Stat-to-dyn* works better: for example, its matches for /ʁ/ are nasal vowels, and top matches for /f, s, ʃ/ do include not only vowels, but also /f, v/, /s, z/ and /ʃ, ʒ/ respectively. However, when we take into account the vocalic context, the best matches drastically improve for /f, v/ and /s, z/: the best match for the dynamic /f(a)/ is the static /f(u)/ or /f(i)/ for S_A, /f(u)/ for S_B; for /f(o)/, it is /w(a)/ or /f(o)/ for S_A and /f(u)/ or /f(o)/ for S_B; for /s(o)/, it is /s(u)/ or /s(ø)/ for S_A and /s(u)/ or /f(u)/ for S_B. There is no improvement for /ʁ/ and /ʃ, ʒ/: the dynamic /ʁ/ gets mostly confused with static /k/, and the static /ʁ/ straight out does not correspond to any reliable matches in the dynamic datasets (basically, every data split changes the best candidate, and they do not repeat). One possible explanation could be that the liquid sound /ʁ/ is produced through a fleeting contact between the velum and the tongue in RT-MRI data and through a clearly visible, full-on contact, which is more appropriate for a stop or a trill, in the static data. As for /ʃ, ʒ/, their dynamic instances are matched to the static /n/, /s/, /l/, /t/—again, and sounds with a similar place of articulation, but a different manner. The SIFT matches of the static /ʃ/ are extremely noisy. Whenever SIFT does matches it correctly, it is to the dynamic /ʒ/ rather than /ʃ/.

Just like in fricatives, in stops and nasal stops, with no regard for anticipated vowels *dyn-to-stat* SIFT matches them to static vowels, but when we take into account the vocalic context, it identifies the approximate place of articulation: aside from correct matches, we may have frequent confusions like /n/-/t/-/f/-/s/ and /p/-/m/-/v/. It should be noted, however, that no vocalic context of the static /l/ ever appears among the best matches of the dynamic /l/. Also, in the *stat-to-dyn* direction, SIFT matches the static /n/ not only to the dynamic /n/, but also to /t/, /v/ and /z/, thus favoring dynamic matches with a close place of articulation, but no contact and no nasality. The matches of the static /l/ in its various vocalic contexts most often are the instances of /j/, though /n/, /v/, /z/ appear too.

This means that SIFT is efficient at identifying the place of articulation, not only for the consonants, but also for the vowels (whether they are front, central or back). Putting together the evidence of mismatches, the results of SIFT indicate that the static-MRI simulation of the liquids /l, ʁ/ and treatment of nasality may have indeed been incorrect: the static and dynamic samples of /l, ʁ/ almost never match, and the asymmetry of the matches in the nasals means that, in the case of nasal sounds, the velopharyngeal port was not open enough in the static

dataset to correctly reflect the dynamic one, and in the case of oral sounds, they were slightly nasalized, which prevented the dynamic dataset's oral sound samples from being matched to them.

The analysis of *dyn-to-stat* SIFT_l matches shows that it makes indiscriminate matches in the case of vowels, matching almost all of them to /o/ and /ɔ/ in the case of S_A and /u/ and /o/ of S_B. In approximants and liquids, dynamic /j/ are matched to static /l/, and static samples of /l/ are matched to dynamic ones of /j/. *Stat-to-dyn* SIFT_l is very noisy, either producing new top matches at almost every data split, or matching indiscriminately to /u, œ, fi/.

Before concluding that SIFT seems to be most sensitive to articulation, let us investigate its methodology in greater detail. Figure 5.10 shows what kind of matches can be made correctly by SIFT and how it reflects the interpretability of these results.

In particular, Figure 5.10a and Figure 5.10b are examples of very reasonably identified matches. Both of them depict the SIFT transformation from a static vowel—/a/ in the case of Figure 5.10a, /ɛ/ in the case of Figure 5.10b—to a dynamic one, /a/. The algorithm picks up on all the key points of articulation. The question of interpretability comes up when comparing the resulting values of these comparisons. The expected result, naturally, would be that the static /a/ resembles the dynamic /a/ at least slightly better than the static /ɛ/ does, despite the fact that the gesture of /ɛ/ passes over a shape very similar to /a/. This is not, however, what happens. Due to the multiple matches at the articulatorily less relevant chin and a match of the constriction of the vocal tract despite its incorrect location, the static /ɛ/ gains a higher match count and is thus considered more similar to /a/ than the static /a/ is ($\text{SIFT}(/\text{a}/_{\text{st}}, /\text{a}/_{\text{dyn}}) = 0.1, \text{SIFT}(/\text{a}/_{\text{st}}, /\text{ɛ}/_{\text{dyn}}) = 0.1125$). This behavior is the perfect example of the motivation to analyze the values over time and on average, since the change in time proves to be more illustrative of the real similarities—see Figure 5.6.

Sometimes the matches are indeed correct, but we have to admit that this success does not do justice to what is happening in the frame from the articulatory point of view, as is shown in Fig. 5.10c where the algorithm picks up on the shared place of articulation between /n(i)/ and /z/ rather imprecisely; alternatively, even if the matches do identify a certain similarity as in Figure 5.10d between /p(a)/ and /p(i)/, it may not be significant enough to show up in the numerical results due to too few matches in total.

It also goes without saying that the algorithm may also make mistakes, the extreme cases of which are shown in Figure 5.11: the mistakes may be made in such a way that does not disqualify the final result, as in Figure5.11a, or such that cannot be treated in any other way than noise, as in Figure 5.11b. As mentioned in the SIFT configuration above in Chapter 5.2.2, I visually analyzed a set of frame pairs to find the right balance between selectivity and productivity of SIFT matches through variations in its parameters. The final set of parameters ensured that cases like in Figure 5.11 did not happen often.

Thus, it is reasonable to conclude that EMD and SIFT_l do not provide (much) articulatorily relevant information, while SIFT definitely does, especially on average, which enables us to discard noise. SSIM can also provide some indirect evidence in the **stat-to-dyn** direction, once the articulators are close enough. In the case of vowels, SIFT is more sensitive to the front-central-back distinction, while SSIM to open-mid-close.

The qualitative analysis of matches and mismatches brings us to the following conclusions:

- We do not dispose of the means to validate or discredit the simulation of the quality of

vowel production in the static MRI dataset.

- There is an issue of nasality representation in the static MRI dataset: the oral sounds are slightly too nasalized, and the nasal sounds are not nasalized enough.
- The liquids /l, ʁ/ produced for the static MRI dataset do not match those seen in the real-time data.
- The phoneme class that is matched the best is the fricatives barring /ʃ, ʒ/, which seem to have been problematic to either produce in the static MRI setting or to match by our methods.
- Judging by matches like /a/-/k(a)/ or /o/-/p(o)/, what may be happening could be **(a)** the contribution of consonant-vowel transition periods in the real-time data that is not represented by the static frames, or **(b)** the phonetic label timing, obtained through forced alignment, could be incorrect, thus affecting the analysis.
- A curious effect is that it was common that voiced consonants from real-time data matched more often than their unvoiced counterparts, which were actually recorded in the static dataset. It is not necessarily due to their vocal tract shapes being a better match, but could be simply due to the procedure of averaging, since voiceless fricatives typically last longer than their voiced counterparts, thus allowing for a greater articulatory variation.

5.4 Evaluation

The qualitative analysis in the previous section above was performed on the most frequent and closest matches. There is a need to also evaluate the results quantitatively, to cover a larger expanse of data rather than the top most frequent matches and to make sure their interpretation was not affected by human bias.

5.4.1 Articulatory similarity measure

Let us consider a function that can compare any two images depicting vocal tract configurations and give a value of how similar they are. Let us say, we compare /l/ and a number of images: /l/, /u/, /k/ and /f/, and according to our analysis, the image that is closest to /l/ is /f/. This is incorrect, but these phonemes actually have a lot in common: both are articulated at the front of the vocal tract (alveolar and labiodental consonants), both are oral, and the lip position is very similar unless affected by coarticulation. It would be reasonable to penalize this error less than for assuming that the closest phoneme to /l/ is /k/, which has a different place of articulation.

This motivated me to create a "golden standard" similarity measure based on the similarity of articulatory features, according to the principles laid out above in Chapter 5.2.2.

This measure, $\mathrm{RS}(\mathrm{ph}_1(V_1), \mathrm{ph}_2(V_2))$ standing for *reference similarity* (Equation 5.10) between phonemes ph_1 and ph_2, optionally anticipating V_1 and V_2 respectively, has several components.

The first component is CF (Equation 5.5), for *common features*, and it is based on the lists of phonetic classes each phoneme is associated to. For example, /n/ is a *nasal consonant*

(which refers to the class of nasal stops),*alveolar*, *nasal* (which refers to all phonemes with nasality), *consonant*, and /t/ is a *plosive*, *alveolar*, *oral consonant*, *oral* and *consonant*. The features they share are *alveolar* and *consonant*—two features in total,—while the union of the lists of features is six features long. Thus the proportion of the shared features is two over six.

$$CF(ph_1(V_1), ph_2(V_2)) = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}, F_k = \{\text{articulatory features of } ph_k\} \tag{5.5}$$

The values of CF thus vary from 0 to 1, 0 standing for two phonemes that do not have a single feature in common and 1 for those that are identical from the articulatory point of view (such as /p/ and /p/ or /k/ and /g/).

One disadvantage of CF is that it has no information on how close or far to each other two places of articulation are, or how similar or distinct two manners of articulation are. This information is used in the next two components, PlA and MA, standing for the *places* and *manners of articulation* respectively.

PlA works on the ordering of all possible places of articulation within the vocal tract: *labial*, *alveolar*, *palatal*, etc. The distance between element positions on this list, divided by the maximal distance possible (6 since my implementation has 7 categories of places of articulation), can be used to measure the similarity of the places of articulation—Equation 5.6:

$$PlA(ph_1(V_1), ph_2(V_2)) = \frac{|pl_1 - pl_2|}{6}, pl_k = \text{place of articulation of } ph_k \tag{5.6}$$

Similarly to PlA, MA works on the ordering of all possible manners of articulation: *plosive*, *nasal consonant*, *fricative*, *approximant*, *close vowel*, *close-mid vowel*, etc.—Equation 5.7:

$$MA(ph_1(V_1), ph_2(V_2)) = \frac{|m_1 - m_2|}{7}, m_k = \text{manner of articulation of } ph_k \tag{5.7}$$

The final two components, CVsim (Equation 5.8) and VVsim (Equation 5.9), are values that relate to the anticipated vowels V_1 and V_2, taking into account the contribution of coarticulation into the similarity between ph_1 and ph_2. Since V_1 and V_2 are optional, if one or both of them are missing, the respective components do not apply. They are also omitted when dealing with cases such as /p(a)/ and /b(a)/—consonants that vary only in voice and anticipate the same vowel—in order not to penalize them for dissimilarity between /p/ and /b/ and /a/.

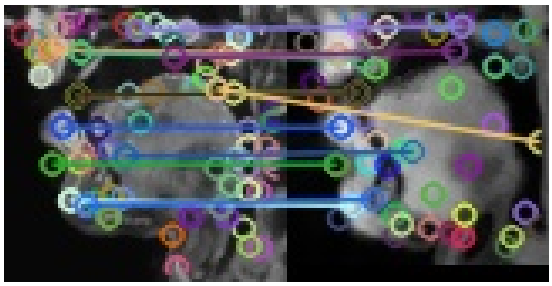
$$CVsim(ph_1(V_1), ph_2(V_2)) = RS(ph_1, V_2) + RS(ph_2, V_1) \tag{5.8}$$

$$VVsim(ph_1(V_1), ph_2(V_2)) = RS(V_1, V_2) \tag{5.9}$$

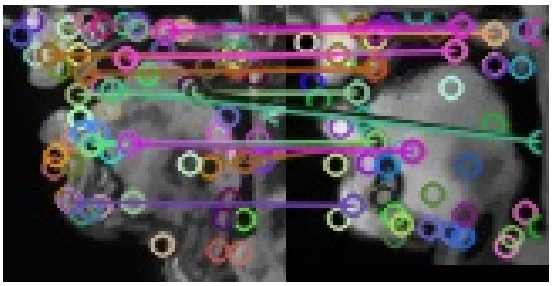
Then RS can be defined as follows:

$$RS(ph_1(V_1), ph_2(V_2)) = \frac{1}{\sum_f w_f} \times \sum_{f \in \{CF,PlA,MA,CVsim,VVsim\}} w_f \times f(ph_1(V_1), ph_2(V_2)) \tag{5.10}$$

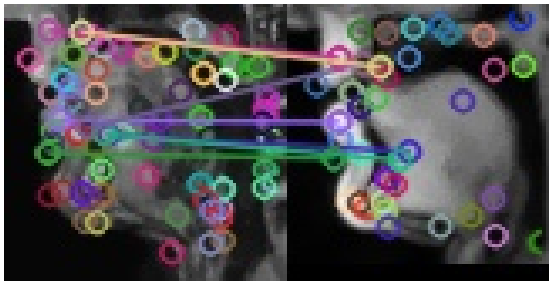
I empirically used $w_{CF} = 0.5, w_{PlA} = 1, w_{MA} = 1, w_{CVsim} = 0.3$ and $w_{VVsim} = 0.1$.
An illustrative excerpt of the values that can be obtained with RS is given in Table 5.6.



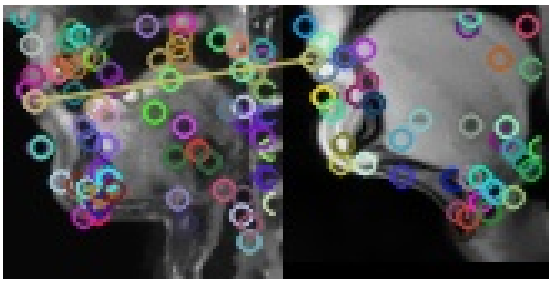
(a) An example of correct results produced by SIFT when comparing the static /a/ with the dynamic /a/: correctly matched the alveolar region, the palate, the upper and lower lips, two matches at the chin, the tongue, the opening between the tongue dorsum and the velum, the velum itself. The only imprecise match is the space between the uvula and the back of the tongue being mapped further down the pharyngeal wall of the speaker. The total number of matches is 8, which produces the value $8/80 = 0.1000$. Compare this SIFT result to the one in Fig. 5.10b, where the same dynamic frame is compared to /ɛ/ rather than /a/.



(b) An example of correct results produced by SIFT when comparing the static /ɛ/ with the same dynamic frame as in Fig. 5.10a, /a/: just like in that example, SIFT correctly matched the palate, the upper and lower lips, bottom of the tongue and the velum. One more match at the chin appeared. Despite the fact that /a/ is central and /ɛ/ is front, the frontal narrowing between the tongue and the palate in the static frame was matched to the central one in the dynamic. The total number of matches is 9, which produces the value $9/80 = 0.1125$, which is greater than 0.1000 from comparing the theoretically closer static frame of /a/ to this dynamic frame.

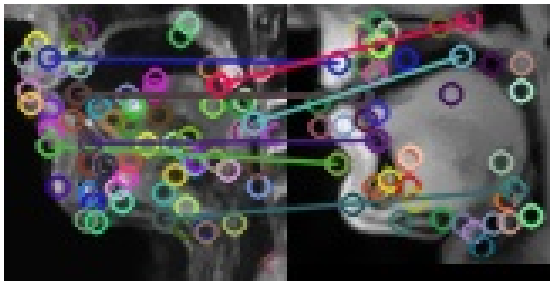


(c) An example of correct results produced by SIFT when comparing the static /n(i)/ with S_A's dynamic /z/: The algorithm correctly matched the upper and lower lips, the chin and the palate, and it is correct not to match the alveolar contact of /n/ with the mere narrowing of the vocal tract for the alveolar fricative /z/. However, these matches only very approximately reflect the shared place of articulation for the two sounds.

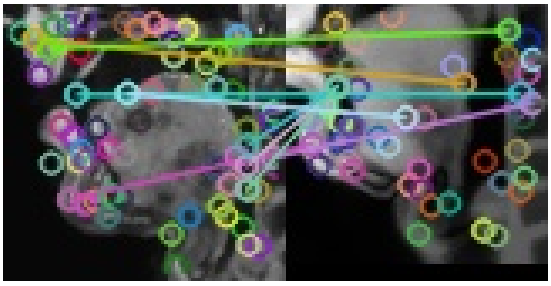


(d) An example of correct results produced by SIFT when comparing the static /p(a)/ with S_B's dynamic /p(i)/: The algorithm correctly matched the place of articulation, the lips, but since this is the only match identified, it will not be significant enough to consider the sounds similar.

Figure 5.10: Examples of quite reasonable matches done by SIFT when comparing static images (left halves) to the dynamic ones (right halves). Circles indicate areas that were features, i.e. that were considered as potential matches. Whenever a circle at the left is joined to a circle at the right, it means the respective areas matched and, from the algorithm's point of view, the configuration of the articulator in that region is similar.



(a) An example of mixed results produced by SIFT when comparing the static /m(5̃)/ with the dynamic pause (S_A will eventually start pronouncing /ku/): two correct matches at the lips, one of the absence of alveolar contact, one of the sublingual cavity, and yet mismatching the chin with the epiglottis, the back of the tongue with the dorsum and the constriction between the tongue and the velum with a similar shape in the nasal cavity. This is an example of reasonably noisy matches.



(b) A worst-case scenario for SIFT, where all the matches are very noisy: matching the static /a/ against the dynamic /l(e)/ and mapping the upper lip to the pharyngeal wall and to the back of the tongue, the teeth to the vertebra, the tongue dorsum to its root and the chin to the vertebra. The situation was probably aggravated by the incorrect window.

Figure 5.11: Examples of incorrect matches done by SIFT when comparing static images (left halves) to the dynamic ones (right halves). Circles indicate areas that were features, i.e. that were considered as potential matches. Whenever a circle at the left is joined to a circle at the right, it means the respective areas matched and, from the algorithm's point of view, the articulator's configuration there is similar.

	/a/	/ã/	/b(a)/	/b(ã)/	/d(a)/	/d(i)/	/ø/	/e/	/f(ɛ)/	/m(a)/	/o/	/ʁ(ɔ)/	/s(e)/	/ə/	/y/
/a/	1.00	0.73	0.37	0.34	0.42	0.39	0.73	0.79	0.40	0.36	0.58	0.38	0.44	0.73	0.68
/ã/	0.73	1.00	0.22	0.25	0.28	0.25	0.55	0.56	0.26	0.24	0.68	0.48	0.30	0.67	0.50
/b(a)/	0.37	0.22	1.00	0.85	0.78	0.79	0.48	0.48	0.78	0.75	0.36	0.46	0.71	0.38	0.52
/d(a)/	0.42	0.28	0.78	0.76	1.00	0.89	0.53	0.53	0.71	0.69	0.41	0.51	0.80	0.43	0.57
/d(ɛ)/	0.41	0.27	0.78	0.76	0.88	0.90	0.53	0.54	0.71	0.69	0.41	0.52	0.81	0.44	0.57
/ø/	0.73	0.55	0.48	0.46	0.53	0.54	1.00	0.92	0.53	0.46	0.81	0.51	0.59	0.74	0.91
/e/	0.79	0.56	0.48	0.46	0.53	0.55	0.92	1.00	0.53	0.47	0.75	0.51	0.60	0.78	0.84
/ɛ̃/	0.74	0.66	0.42	0.41	0.47	0.47	0.78	0.79	0.47	0.43	0.65	0.45	0.52	0.79	0.73
/f(u)/	0.36	0.26	0.77	0.77	0.70	0.72	0.52	0.52	0.88	0.71	0.44	0.57	0.80	0.42	0.57
/f(y)/	0.38	0.24	0.79	0.77	0.71	0.74	0.54	0.53	0.89	0.73	0.42	0.56	0.81	0.42	0.59
/o/	0.58	0.68	0.36	0.37	0.41	0.42	0.81	0.75	0.41	0.34	1.00	0.64	0.47	0.74	0.73
/œ/	0.83	0.65	0.40	0.38	0.45	0.44	0.86	0.79	0.45	0.38	0.68	0.44	0.49	0.74	0.81
/ɔ/	0.68	0.78	0.28	0.29	0.33	0.32	0.68	0.63	0.33	0.26	0.86	0.56	0.37	0.74	0.63
/p(a)/	0.37	0.22	1.00	0.85	0.78	0.79	0.48	0.48	0.78	0.75	0.36	0.46	0.71	0.38	0.52
/ʁ(i)/	0.39	0.45	0.49	0.49	0.54	0.56	0.53	0.54	0.57	0.44	0.62	0.90	0.64	0.53	0.59
/ʁ(ɔ)/	0.38	0.48	0.46	0.47	0.51	0.53	0.51	0.51	0.55	0.41	0.64	1.00	0.60	0.53	0.55
/s(ø)/	0.44	0.30	0.71	0.69	0.80	0.82	0.60	0.59	0.80	0.66	0.48	0.61	0.91	0.48	0.63
/ʃ(ɛ)/	0.50	0.36	0.67	0.65	0.72	0.73	0.63	0.64	0.76	0.62	0.51	0.65	0.81	0.53	0.67
/t(e)/	0.40	0.26	0.78	0.76	0.89	0.91	0.55	0.56	0.72	0.70	0.43	0.52	0.82	0.44	0.58
/u/	0.53	0.63	0.40	0.41	0.45	0.47	0.73	0.68	0.45	0.38	0.91	0.68	0.51	0.69	0.81

Table 5.6: Some reference similarity (RS) values for phoneme pairs.

As it was discussed above, for each static MRI frame and for each RT-MRI frame's phonetic label each of the four measures had its "best candidates"—RT-MRI phonetic labels or static MRI frames respectively that, on average, were the closest. RS enabled me to carry out a formal evaluation of the matches and mismatches in these best candidates.

Furthermore, I generalized PlA (Equation 5.6) and MA (Equation 5.7) to be able to evaluate the best candidates when dealing not with phonemes, but with articulatory classes—RPhClD standing for reference phonetic class distance, Equation 5.11:

$$\mathrm{RPhClD}(\mathrm{phcl}_1, \mathrm{phcl}_2) = \frac{|\mathrm{index}(\mathrm{phcl}_1) - \mathrm{index}(\mathrm{phcl}_2)|}{\mathrm{maximal\ index\ in\ this\ category}} \tag{5.11}$$

When applied to the *oral-nasal* or *rounded-unrounded* distinctions, RPhClD is 0 when the phonetic classes are the same and 1 otherwise. When applied to the places and manners of articulation, RPhClD can take intermediate values between 0 and 1 as well, to show that a *fricative* is as close to an *approximant* as an *approximant* is to a *close vowel*, and that is much closer than to an *open vowel*.

5.4.2 Articulatory error

For each block of aggregated data and each phoneme pair encountered in it, I produced the averages and standard deviations as explained above, thus, given a phoneme in one corpus (static or real-time), obtaining a ranking of phonemes of the complementary corpus by their similarity to the initial phoneme.

To begin, let us consider only the best candidate $\mathrm{cand}_{\mathrm{ph}}^{(1)}$ for each phoneme ph and each of the four similarity measures. The phoneme in question could be compared to the candidate:

$$\mathrm{RS}^{(1)} = \mathrm{RS}(\mathrm{ph}, \mathrm{cand}_{\mathrm{ph}}^{(1)})$$

If the candidate was a perfect match, then $\mathrm{RS}^{(1)}$ was equal to 1. If there was nothing in common between the two phonemes, $\mathrm{RS}^{(1)}$ was 0. This way, for $\mathrm{cand}_{\mathrm{ph}}^{(1)}$, the measure that produced it produced an error $1 - \mathrm{RS}^{(1)}$.

When evaluating a measure (EMD, SIFT, SIFT_l or SSIM) only on the first best candidate, the error AE, standing for articulatory error, would accumulate over the entire set of initial phonemes and their best candidates:

$$\mathrm{AE}_1 = \sum_{\mathrm{ph}} 1 - \mathrm{RS}(\mathrm{ph}, \mathrm{cand}_{\mathrm{ph}}^{(1)})$$

Let us consider further best candidates. The measure expectation for the $(\mathrm{ph}, \mathrm{cand}_{\mathrm{ph}}^{(1)})$ pair was $\mathrm{E}[\mathrm{sim}(\mathrm{ph}, \mathrm{cand}_{\mathrm{ph}}^{(1)})]$; for, say, the second best candidate it was lower in the case of SIFT, SIFT_l and SSIM and higher in the case of EMD: $\mathrm{E}[\mathrm{sim}(\mathrm{ph}, \mathrm{cand}_{\mathrm{ph}}^{(2)})]$. We can count this difference in and discount the error contributed by a second- or third-best candidate proportionally to the

difference between their averages—Equation 5.12:

$$AE_k = \begin{cases} \sum_{ph} \sum_k (1 - RS(ph, cand_{ph}^{(k)})) \times \frac{E[sim(ph, cand_{ph}^{(k)})]}{E[sim(ph, cand_{ph}^{(1)})]} & \text{if } sim = SIFT \text{ or } SIFT_l, \\ \sum_{ph} \sum_k (1 - RS(ph, cand_{ph}^{(k)})) \times \frac{E[sim(ph, cand_{ph}^{(k)})]+1}{E[sim(ph, cand_{ph}^{(1)})]+1} & \text{if } sim = SSIM, \\ \sum_{ph} \sum_k (1 - RS(ph, cand_{ph}^{(k)})) \times \frac{E[sim(ph, cand_{ph}^{(1)})]}{E[sim(ph, cand_{ph}^{(k)})]} & \text{if } sim = EMD \end{cases} \quad (5.12)$$

This way, if the average value for the second best candidate was $\frac{2}{3}$ times the average value for the winner, this RS-related error will be counted within the final AE_k error value with the weight $\frac{2}{3}$.

Table 5.7 presents the best and worst similarity measures, according to the articulatory error from Equation 5.12, across the entire dataset, aggregated by speakers, directions of comparisons (*stat-to-dyn* or *dyn-to-stat*), the number of best matches counted into the formula (one or two), the best and worst similarity measures and their average error across the entire corpus in that speaker-direction-coarticulation modality setting. It shows that, in fact, on average the matches on the frames of S_A were better with SIFT, and those of S_B (and consequently, both of them pulled together) with SSIM (but for *stat-to-dyn* comparisons when coarticulation is taken into account, when SIFT is better for S_B too—though actually their error values are almost equal there, 0.2406 for SIFT and 0.2421 for SSIM). The worst matches are EMD and $SIFT_l$.

Nb_best	Speaker	Direction	Mode	Best measure	Error of the best measure	Worst measure	Error of the worst measure
1	S_A	*St-to-dyn*	C, V	SIFT	0.227	EMD	0.326
			C(V), V	SIFT	0.228	$SIFT_l$	0.264
		Dyn-to-st	C, V	SIFT	0.272	$SIFT_l$	0.374
			C(V), V	SIFT	0.247	EMD	0.344
	S_B	*St-to-dyn*	C, V	SSIM	0.261	$SIFT_l$	0.313
			C(V), V	SIFT	0.241	$SIFT_l$	0.256
		Dyn-to-st	C, V	SSIM	0.277	EMD	0.342
			C(V), V	SSIM	0.194	EMD	0.329
	Both	*St-to-dyn*	C, V	SSIM	0.261	EMD	0.335
			C(V), V	SSIM	0.233	$SIFT_l$	0.265
		Dyn-to-st	C, V	SSIM	0.247	$SIFT_l$	0.353
			C(V), V	SSIM	0.224	EMD	0.340
2	S_A	*St-to-dyn*	C, V	SIFT	0.213	EMD	0.322
			C(V), V	SIFT	0.220	EMD	0.245
		Dyn-to-st	C, V	SIFT	0.266	EMD	0.337
			C(V), V	SIFT	0.237	EMD	0.321
	S_B	*St-to-dyn*	C, V	SSIM	0.270	EMD	0.304
			C(V), V	$SIFT_l$	0.221	EMD	0.243
		Dyn-to-st	C, V	SSIM	0.266	EMD	0.343
			C(V), V	SSIM	0.200	EMD	0.330
	Both	*St-to-dyn*	C, V	SIFT	0.251	EMD	0.329
			C(V), V	$SIFT_l$	0.221	EMD	0.249
		Dyn-to-st	C, V	SSIM	0.248	EMD	0.341
			C(V), V	SSIM	0.224	EMD	0.324

Table 5.7: The best and worst similarity measures according to the AE_1, AE_2 error functions (Equation 5.12).

5.5 Towards bringing the two directions together

Of the two approaches I have studied, one driven by static articulatory targets and one synthesizing the articulatory parameter sequence just like the acoustics, both had their strong and weak points, as discussed above. It would be highly beneficial to create articulatory targets with static captures and transitions between them with dynamic data. The work presented above can thus serve as a point to build upon when working towards bringing the two approaches together in a hybrid articulatory synthesizer.

First off, we showed how SSIM and SIFT can help us filter out MRI captures where the static simulation of sound production did not succeed: these are the sound categories (liquids, sibilants) where both SSIM and SIFT repeatedly fail to match the dynamic frames to the their static counterparts.

Then, one could augment the dataset of blocked articulations through combining the ensemble of static images (that would bring in the speaker's vocal tract dimensions, preserving the homogeneity of the database) with multiple dynamic frames labeled with the phonetic label of interest (that would bring in the articulatory information) with techniques such as [DTI+19].

As for extracting the articulatory strategies for transitions, SSIM and SIFT do not offer a straightforward way to do that with RT-MRI data. A promising prospect would be to go back to individual windows in the case of SSIM and feature matches in the case of SIFT and to try to identify those that have a high SSIM score or that are close matches. Such a specific focus on one place of articulation may be less informative for the global picture, but it could contribute more to our understanding of the articulators' navigation to and from the hypothetical articulatory target.

5.6 Conclusion

5.6.1 Overview of results

To match our static MRI dataset to the dynamic RT-MRI one, we have presented four image similarity measures. All four of them were appropriate candidates for investigation: they were known to be used in the field of computer vision, they were sensitive to the changes in the data, and their distributions were able to distinguish the speakers and even speaking styles.

Two of them, EMD and $SIFT_l$, were subsequently disqualified from drawing conclusions regarding articulation because of their temporal behavior (not informative peaks in the considered examples, high correlation between different comparisons in the case of EMD and most values being too close to 0 in comparison to occurring spikes in the case of $SIFT_l$—Figures 5.4 and 5.6b), their distributions (no discrimination between the similarity measure distributions for very different sounds), through the qualitative analysis of the phonemes of one dataset most often found to be on average the closest to a given phoneme of the other dataset (no articulatory interpretation for the confusions), and through the quantitative analysis with the AE_1, AE_2 error functions (Equation 5.12, Table 5.7), where they were ranked as the worst similarity measures in all modalities of the experiment.

The other two, SIFT and SSIM, were found to be capable of capturing some articulatory information, which was validated first with the temporal analysis on several examples, then, circumstantially, through having different measure distributions for sufficiently different sounds,

then through a qualitative analysis, where despite using a different methodology, they reached rather consistent conclusions regarding what sounds can be and cannot be matched through delving into the computation process of SIFT, and finally, through a quantitative analysis, where their performance was consistently the best judging by the AE_1, AE_2 error functions. SIFT performed better on matching the frames of S_A—same speaker as in the static dataset,— and SSIM fared better with a change of speaker for S_B. This could possibly relate to the shape of the distributions of SSIM comparison values, where while the mode of S_B's comparisons was lower than the mode of S_A's ones, the mass of images with a similarity value greater than the mode was higher for S_B than for S_A.

As the measures were rather noisy, a single comparison between just two images was usually not informative enough. Patterns apparent enough to let us draw conclusions appeared only through averaging and selecting the most frequently identified best matches in the case of qualitative analysis and all the best matches in the case of the quantitative one.

The quality of the static MRI dataset seems to be appropriate: in general, it can serve both as a reference for the place of articulation produced in the dynamic data and as a manifestation of coarticulation. The problematic sounds and features that I was able to identify through the analysis of measure distributions and mismatches were the liquids /l, ʁ/, whose dynamic production could not be matched by their static simulation, the alveolar fricatives /s, ʃ/, again, simulated unrealistically in the static setting, and the feature of nasality: apparently, the oral sounds in the static corpus were slightly too nasalized, and in the nasal sounds, vice versa, the velopharyngeal port did not open enough.

These problems tie well with the known issues of blocked articulation with no phonation (the precision of articulation of the sounds with a complicated temporal scenario), aggravated by the supine position in an MRI machine.

Furthermore, the peaks in the temporal behavior analysis and the mismatches found in the qualitative analysis are evidence of a great influence of transition periods in the RT-MRI data on the results. The shapes appearing at the time of transition to and from a sound may fail to be explained with the static image that, according to the phonetic label, should be the closest.

This means that, if we were to build a hybrid articulatory speech synthesizer both on MRI and RT-MRI data, MRI data could indeed be used on the condition that it would be cleaned out from the sounds whose simulation was not successful.

As for finding articulatory targets in the RT-MRI data to use them in a similar way as the static MRI frames, we have to conclude that the impact of coarticulatory effects and the rapidness of real-time articulation would prevent us from singling out one particular image of, for example, /b(i)/ to use it as a target. Since we were only able to find patterns through averaging and the analysis of most frequent values, the same approach would need to be used to create articulatory targets from RT-MRI frames: we would need to combine numerous samples labeled with /b(i)/ in one single picture. Static pictures could then serve as a source of information on the place of articulation, to help choose the best articulated frames out of the entire phoneme. Another approach could be to assume phonetic labeling to be correct and to indiscriminately take the middle of each phoneme as the target, as it was done in [LQSN17].

5.6.2 Future work

A direct continuation of this study would be to repeat it focusing on specific articulators, by taking windows as defined in Chapter 4. For example, a window of the velum (as in Figure 4.2b) is expected to let us zero in on the feature of nasality and disregard everything that would be irrelevant to it.

When building a joint articulatory synthesizer using both static and dynamic articulatory data, in order to benefit from the precision and little variation of the static data and from the actual speech production information in real time as manifested in the dynamic data, it seems feasible to extend the collection of static frames to cover more phonetic contexts, which could be done with image combination techniques such as [DTI+19].

Furthermore, the similarity measures could be integrated into forced alignment of RT-MRI data to improve the quality of phonetic labeling. It would be especially fruitful to combine the information of SIFT and SSIM, to be able to handle both similar and different speakers and to benefit from the strengths of both measures.

6

Conclusions

6.1 Global overview

The last three chapters presented the results for an articulatory speech synthesizer driven by coarticulation-aware static vocal tract configurations (Chapter 3), an articulatory speech synthesizer that functioned as a regular DNN-based parametric speech synthesizer expanded with articulatory parameters (Chapter 4) and a bridge between the data types underlying the two (Chapter 5).

It was shown that, in both methods used in articulatory speech synthesis, we can benefit from some advantages but have to face certain limitations inherent to the method.

Rule-based articulatory speech synthesis boasted a comprehensive control over the entire set of the articulators—the jaw, the tongue, the lips, the velum, the epiglottis and the larynx—as well as F0, glottal opening, subglottal and supraglottal pressure; with a varying quality of the result, it could cover the entirety of French phonology thanks to an extended set of vocal tract configurations, coming from static MRI images, that served as building blocks for the system. Applied as is, it used to have a problem with attaining constrictions for stops and nasal stops, reaching too close constrictions for fricatives and vowels, and mistreating the feature of nasality. This was efficiently, if brutally, solved by resetting those tubes of area functions that contradicted the phoneme in production, implemented at a post-processing stage, right before simulating the acoustics.

The movements and the speech generated by the rule-based system were correct enough to showcase the general validity of the approach but far from natural or intelligible enough to become usable in any real-case scenario. This was due to two major factors: timing of the system, that, too, was managed with rules, and the interplay between different components. The said interplay was not as decisive in the case of vowels, which have a fairly simple acoustics of production, and of stops, which have a very recognizable acoustic cue of a burst, but quite crucial for fricatives, where the place of articulation, the pressure and the timing need to be especially well coordinated, and for liquids, that were synthesized closer either to semi-vowels or to stops at the corresponding places of articulation.

The DNN-based parametric articulatory speech synthesizer that was presented then addressed the issue of the subpar timing control that was so conspicuous in the previous system. It was based on a standard deep-learning parametric speech synthesizer developed with the default "build your own voice" method in Merlin [WWK16] and trained with a feed-forward

network on the denoised audio recorded simultaneously with RT-MRI acquisitions. Then the original acoustic model of that base speech synthesizer was augmented with articulatory parameters automatically extracted from RT-MRI images. Those parameters did not represent as comprehensive a picture of articulation as the first synthesizer, but should give us the position of the lips, indirect evidence of the position of the tongue, nasality and key constriction values between the velum and the tongue and the tongue and the pharyngeal wall. The quality of speech and movements attained by this articulatory speech synthesizer is considerably higher than that of the first one. Articulation-wise, compared to the first synthesizer, it deals well with vowels and fricatives but struggles with those sounds that require the articulators to come in contact. As for articulatory trajectories, they are handled considerably more naturally that those in the first synthesizer.

The common point between the two systems was the use of MRI, albeit of different kinds: static and dynamic. It was important to explore the relation between the two types of data and to try to identify some key vocal tract configurations among those recorded in RT-MRI similar to what was captured in the static setting. I concluded that the static MRI dataset was in general valid, although it struggled with the representation of fricatives, where the aerodynamics of the production is an important factor to being able to pronounce the sound correctly, and liquids, the production of which requires the knowledge of their temporal behavior. The consequence of that is, possibly, an inferior quality of synthesis of those two classes of sounds by the rule-based synthesizer that relied on static MRI data was not only due to a misaligned strategy of all its components control, but also due to the shortcomings of the original data as well.

6.2 Future work

Both articulatory speech synthesizers had their strong and weak points. Aside from potential study-specific directions to explore that were mentioned at the end of each chapter, overall, it would be quite pertinent to unite the two approaches in a hybrid system. This hybrid system could be relatively simple, such as just using the duration model from the DNN-based articulatory speech synthesizer to inform the timing strategy of the rule-based synthesizer. It could also be much more intertwined. For example, we could make use of a dramatically extended library of vocal tract configurations using dynamic data, or apply an automatic delineation algorithm on RT-MRI data, represent the contours with the same articulatory model as in the rule-based synthesizer and train the DNN-based speech synthesizer with those articulatory parameters. Another promising prospect is to define gestures or targets as elements from a joint sensory-motor space rather than only sensory or only the other; this outlook is supported by psycholinguistic research [DR19].

Considering the long-term applications envisioned for articulatory speech synthesis, it will be necessary to focus not only on the correctness, comprehensiveness, naturalness and intelligibility of systems in development, but also on their ability to serve their purpose. This would require them to be quite flexible and highly reactive, for example, to allow the user to adjust the vocal tract geometry and promptly provide the updated acoustic output.

As for the exploration of articulatory speech synthesis in application to speech production studies, with the quality of synthesis such as in the DNN-based synthesizer, it is already within the realm of possibility, even more so when taking into account the astounding advances of deep

learning that should be able to help boost the output quality even more.

A

Prompts for spontaneous speech in the RT-MRI corpus

French	English
Comment avez vous choisi votre parcours professionnel et l'université ? Parlez nous de vos études.	How did you decide upon your career and university? Talk about your studies.
Parlez nous de ce que vous faites en recherche.	Talk about your research.
Qu'est-ce que vous aimez dans votre travail ?	What do you like in your work?
Racontez votre expérience de l'apprentissage des langues étrangères.	Talk about your experience with learning foreign languages.
Si vous deviez ouvrir un commerce, qu'est-ce que ça serait ?	If you were to start a business, what would it be?
Qu'est-ce que vous achetez au supermarché d'habitude ?	What do you usually buy at a supermarket?
Qu'est ce qui fait un petit déjeuner parfait pour vous ? Et pour les déjeuner et dîner ?	What is the best breakfast for you? Lunch? Dinner?
Pouvez-vous expliquer comment vous cuisinez votre plat préféré ?	Explain how to make your favorite dish.
Où préférez vous manger : dans un restaurant, chez vous ?	Do you prefer eating out or at home?
Que pensez vous des fast-food ?	What is your opinion on fast food?
Veuillez décrire votre ville natale et votre vie là bas.	Describe your home town and what your life there was like.
Racontez votre dernier voyage.	Talk about your last trip.
Quels sont vos loisirs préférés ?	What are your favorite pastimes?
Racontez un film ou un livre qui vous a laissé une impression durable.	Talk about a movie or a book that made a lasting impression on you.
Quelles sont vos principales sources d'informations (Internet, journaux, télé...) et de quel genre sont ces informations ?	What are your main information sources (the Internet, newspapers, TV...) and what kind of information is it?
Quelles tâches domestiques faites vous ? Bricolez vous ?	What household chores do you do? Do you tinker?
Quelle est votre destination de voyage de rêve et pourquoi ?	What is your dream traveling destination and why?
Que pensez vous du système de santé en France ?	What do you think of the healthcare system in France?
Que pensez vous des grèves en France ?	What is your opinion on strikes in France?
Quels sont les avantages et les inconvénients d'habiter à Paris ?	What are pros and cons of living in Paris?
Pensez vous que vous êtes économe ? Quelles dépenses vous permettez vous facilement ou non ?	Do you think yourself sparing? What expenses do you allow yourself easily and what not?

Table A.1: Prompts for spontaneous speech used to acquire the RT-MRI part of the ArtSpeechMRIfr [DFF+19] corpus.

B

Detailed summary in French

Cette thèse se situe dans le domaine de la synthèse articulatoire de la parole, ce qui est la tranformation du texte en sa réalisation vocale combinée avec l'évolution des mouvements des organs de l'articulation, des articulateurs, nécessaire pour produire l'énoncé.

La thèse est organisée en trois grandes parties :

- Celle consacrée au développement d'un synthétiseur articulatoire de la parole basé sur le concepte des cibles articulatoires ;
- Celle consacrée au développement d'un synthétiseur articulatoire de la parole qui s'appuie sur des données enregistrées en temps réel ;
- Celle qui traite des liens que l'on peut établir entre les deux approches au-dessus utilisées.

B.1 Synthèse articulatoire de la parole à partir des données IRM statiques

Le premier synthétiseur est issu d'une approche à base de règles. Celle-ci visait à obtenir le contrôle complet sur les articulateurs (mâchoire, langue, lèvres, vélum, larynx et épiglotte). Elle s'appuyait sur des données statiques du plan sagittal médian obtenues par IRM (Imagerie par Résonance Magnétique) correspondant à deux types d'articulations sans phonation :

- Des articulations bloquées de voyelles du français,
- Des articulations bloquées des syllabes de type consonne-voyelle où le locuteur devait imaginer être sur le point de prononcer la syllabe.

Cette partie du travail était composée de plusieurs étapes :

- L'encodage de l'ensemble des données grâce à un modèle du conduit vocal basé sur l'ACP (analyse en composantes principales) qui sépare l'influence des articulateur de l'un à l'autre et puis traite des contours d'un articulateur à la fois ;
- L'extension de l'ensemble des données pour compenser l'absence des enregistrements pour les syllabes non traitées dans le corpus. Cette estimation a été faite grâce à l'identification

de la rélation entre les vecteurs de /a/, /i/, /u/, /y/ et les vecteurs des autres voyelles et à l'application de cette rélation aux configurations du conduit vocal pendant l'articulation des syllabes traitées ;

- L'utilisation des configurations articulatoires obtenues comme sources de positions à atteindre et destinées à piloter le synthétiseur à base de règles qui est la contribution principale de cette première partie. Les facteurs qui ont été pris en compte dans le choix des cibles était des contraintes temporailes, spatiales et catégoriques. Une fois les cibles choisis, j'ai testé trois stratégies d'interpolation entre les vecteurs cibles :
  - Linear : l'interpolation entre les vecteurs cibles est linéaire, avec des virages serrés aux nœuds ;
  - Cosine : transitions lisses ;
  - Piecewise 1-d monotonic cubic Hermite interpolation : des transitions lisses, l'amplitude de chaque section de transition limitée par ses nœuds d'interpolation correspondants ;
  - Complex : les transitions se font avec l'interpolation précédente de Hermite cubique, mais le timing varie selon les articulateurs. Ceux qui sont critiques atteignent leur position cible plus rapidement que les autres, tandis que les articulateurs dont la contribution à l'intelligibilité sonore résultante n'est pas aussi grande se déplacent plus lentement (par exemple, la langue peut être dans plusieurs positions pour le son /b/, mais les lèvres doivent entrer en contact). De plus, les articulateurs composés de tissus plus lourds (comme le dos de la langue) se déplacent plus lentement que ceux légers et très mobiles (comme les lèvres).
- L'ajustement des conduits vocaux obtenus selon une perspective phonétique, par exemple, par l'imposition du contact entre les lèvres lors de la production d'une occlusive labiale ;
- La simulation acoustique permettant d'obtenir un signal acoustique, controlée par une séquence des parametres acoustiques qui doivent être en accord avec l'évolution du conduit vocal.

Les résultats de cette synthèse ont été évalués de manière visuelle, acoustique et perceptuelle, et les problèmes rencontrés ont été identifiés et classés selon leurs origines, qui pouvaient être : les données, leur modélisation, l'algorithme contrôlant la forme du conduit vocal, la traduction de cette forme en fonctions d'aire, ou encore la simulation acoustique.

Les résultats montrent que les formes des articulateurs changent dans le temps en fonction des trajectoires produites du conduit vocal, et celles-ci sont phonétiquement correctes. Elles évoluent en synchronisme avec d'autres paramètres de la production vocale : F0, pression sous-glottale et supraglottale. Ensemble, le timing et les valeurs du système doivent être dans un équilibre délicat afin de ne pas produire d'artefacts dans le pipeline de synthèse. Nous trouvons qu'ils sont suffisamment bien accordés pour les voyelles et les occlusives ; quant aux fricatives, l'interaction entre le lieu d'articulation, le contrôle de la pression et l'évolution temporelle est si complexe qu'elle se résume essentiellement à la modélisation de chaque fricative séparément.

J'en conclus que, dans le cadre de l'ensemble des données du conduit vocal, un modèle statistique pour le coder et un ensemble de règles pour le manipuler, ce travail représente une

exploration approfondie de l'approche. Les résultats montrent que de telles données et méthodes ne sont pas inadaptées à la construction d'un synthétiseur de parole articulatoire ; cependant, étant donné les limites de l'approche, il serait plus prometteur de poursuivre ce travail en incorporant des modèles provenant de certaines données dynamiques réellement enregistrées plutôt que de continuer à chercher un meilleur ensemble de règles par modélisation théorique, essais et erreurs.

B.2 Synthèse articulatoire de la parole à partir des données IRM en temps réel

La seconde approche a été développée en s'appuyant sur un synthétiseur de référence constitué d'un réseau de neurones feed-forward entraîné à l'aide de la méthode standard du système Merlin [WWK16] sur des données audio composées de parole en langue française enregistrée par IRM en temps réel.

Ces données ont été segmentées phonétiquement et linguistiquement avec eLite HTS [RBBD14], un outil que j'ai complété par des corrections des erreurs récurrentes. Les étiquettes ont été alignées avec les données audio avec HVite de HTK [YEG+02] et Merlin comme frontend [WWK16]. Ces données audio, malgré un débruitage, étaient fortement parasitées par le son de la machine à IRM, ce qui a causé des problèmes de synchronisation.

Nous avons complété le synthétiseur de référence en ajoutant dix paramètres représentant de l'information articulatoire :

- L'ouverture des lèvres et leur protrusion ce qui donne les informations essentielles sur le comportement des lèvres,
- La distance entre la langue et le vélum, ce qui nous donne les informations sur l'occlusion entre ces deux articulateurs et fournie une indice latente de la position verticale de la langue ;
- La distance entre le vélum et la paroi pharyngale, ce qui contrôle l'accès de l'air dans la cavité nasale et, par conséquent, la nasalisation,
- La distance entre la langue et la paroi pharyngale, ce qui represente une information sur la position horizontale de la langue.

Ces paramètres ont été extraits automatiquement à partir des images et alignés au signal et aux spécifications linguistiques.

Les séquences articulatoires et les séquences de parole, générées conjointement, ont été évaluées à l'aide de différentes mesures acoustiques :

- la distortion mel-cepstrum moyenne,
- l'erreur de prédiction de l'apériodicité,
- trois mesures pour F0 : RMSE (root mean square error), CORR (coéfficient de corrélation) and V/UV (frame-level voiced/unvoiced error).

J'ai conclus que l'ajout des paramètres articulatoires n'a pas dégradé le modèle acoustique original.

Un analyse de la pertinence des paramètres articulatoires par rapport aux labels phonétiques a également été réalisée. L'erreur la plus récurrente a été l'achèvement du contact dans les cas où le contact est indispensable pour produire le son.

Puis, j'ai vérifié la similitude entre les séquences originelles et celles qui étaient générées, avec la distance de déformation temporelle dynamique. Cette évaluation permet de conclure que les paramètres articulatoires générés s'approchent de manière acceptable des paramètres originaux.

B.3 Cibles statiques et la parole en temps réel pour la synthèse de la parole articulatoire

Les deux approches présentées ci-dessus ont en commun l'utilisation de deux types de données IRM. Ce point commun a motivé la recherche, dans les données temps réel, des images clés, c'est-à-dire les configurations statiques IRM, utilisées pour modéliser la coarticulation. Afin de comparer les images IRM statiques avec les images dynamiques en temps réel, nous avons utilisé plusieurs mesures :

- La similarité structurelle,
- La distance du "terrassier",
- SIFT ;

Après avoir vérifié la pertinence et la validité de ces mesures, j'ai étudié qualitativement et quantitativement, puis interprété leur comportement; j'ai ensuite analysé leur similarités.

J'en ai conclu que SIFT et la similarité structurelle capturaient bien les informations articulatoires et que leur comportement, de manière générale, validaient les données d'IRM statiques.

Les phonèmes et traits phonétiques problématiques que j'ai pu identifier à travers les analyses des distributions et incompatibilités de mesures étaient les liquides /l, ʁ/, dont la production dynamique ne pouvait être imitée par leur simulation statique, les fricatives alvéolaires /s, ʃ/, elles aussi simulées de manière non réaliste dans le contexte statique, et les caractéristiques de la nasalité.

Il semblerait que les sons du corpus statique aient été légèrement trop nasalisés, et que réciproquement les sons nasalisés présentaient une ouverture vélopharyngée insuffisante. Enfin, j'ai discuté l'impact de cette étude pour de futurs synthétiseurs articulatoires hybrides de la parole.

B.4 Conclusion globale

La synthèse vocale articulatoire basée sur des règles permettait un contrôle complet de l'ensemble des articulateurs - la mâchoire, la langue, les lèvres, le vélum, l'épiglotte et le larynx - ainsi que du F0, de l'ouverture glottale, de la pression sous-glottale et supraglottale ; avec une qualité variable du résultat, elle pouvait couvrir la phonologie française grâce à un jeu étendu de configurations

du conduit vocal, provenant des images IRM statiques qui constituaient la structure du système. Appliqué tel quel, il avait un problème pour atteindre des contacts pour les occlusives et les occlusives nasales, atteindre des approchements trop étroits pour les fricatifs et les voyelles, et maltraiter la caractéristique de la nasalité. Ceci a été résolu efficacement, bien que brutalement, en réinitialisant ces tubes de fonctions de zone qui n'était pas en accord avec le phonème en production, mis en œuvre à l'étape de post-traitement, juste avant de simuler l'acoustique.

Les mouvements et les énoncés générés par le système fondé sur des règles étaient suffisamment corrects pour démontrer la validité générale de l'approche, mais loin d'être assez naturels ou intelligibles pour devenir utilisables dans un scénario réel. Cela s'explique par deux facteurs majeurs : le timing du système, qui lui également a été géré selon des règles, et l'interaction entre les différentes composantes. Cette interaction n'a pas été aussi décisive dans le cas des voyelles, qui ont une acoustique de production assez simple, et des occlusives, qui sont acoustiquement très reconnaissables, mais tout à fait cruciale pour les fricatives, où le lieu d'articulation, la pression et le timing doivent être particulièrement bien coordonnés, et pour les liquides, qui sont synthétisés soit près des semi voyelles soit aux points d'articulation correspondants.

Le synthétiseur de parole articulatoire paramétrique basé sur DNN qui a été présenté a ensuite abordé la question du contrôle de la synchronisation subpar qui était si visible dans le système précédent. Il était basé sur un synthétiseur de parole paramétrique standard à apprentissage profond, développé avec la méthode par défaut "build your own voice" de Merlin [WWK16] et formé avec un réseau de feed-forward sur l'audio dénoisé enregistré simultanément aux acquisitions IRM en temps réel. Ensuite, le modèle acoustique original de ce synthétiseur de parole de base a été complété par des paramètres articulatoires extraits automatiquement des images IRM en temps réel. Ces paramètres ne représentaient pas une image de l'articulation aussi complète que le premier synthétiseur, mais devaient nous donner la position des lèvres, la preuve indirecte de la position de la langue, la nasalité et les valeurs clés de constriction entre le vélum et la langue et la langue et la langue et la paroi pharyngeale. La qualité de la parole et des mouvements atteints par ce synthétiseur de parole articulatoire est considérablement supérieure à celle du premier. Du point de vue de l'articulation, par rapport au premier synthétiseur, il traite bien les voyelles et les fricatives, mais a du mal avec les sons qui nécessitent le contact des articulateurs. Quant aux trajectoires articulatoires, elles sont traitées beaucoup plus naturellement que celles du premier synthétiseur.

Le point commun entre les deux systèmes était l'utilisation de l'IRM, quoique de nature différente : statique et dynamique. Il était important d'explorer la relation entre les deux types de données et d'essayer d'identifier certaines configurations clés des voies vocales parmi celles enregistrées dans le IRM en temps réel similaire à ce qui a été saisi dans le cadre statique. J'ai conclu que l'ensemble des données statiques de l'IRM était en général valide, bien qu'il ait eu des difficultés avec la représentation des fricatifs, où l'aérodynamique de la production est un facteur important pour pouvoir prononcer correctement le son, et des liquides, dont la production exige la connaissance de leur comportement dans le temps. La conséquence de cela est, peut-être, une qualité inférieure de synthèse de ces deux classes de sons par le synthétiseur basé sur des règles qui s'appuyait sur des données IRM statiques n'était pas seulement due à une stratégie mal alignée de tous ses composants de contrôle, mais aussi aux lacunes des données originales aussi bien.

Bibliography

[ACFS17] Francesco Avanzini, Piero Cosi, Rolando Füstös, and Andrea Sandi. When fantasy meets science: An attempt to recreate the voice of ötzi the "iceman". *StudiAISV*, 2017.

[AE17] Sasan Asadiabadi and Engin Erzin. Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors. In *Interspeech*, pages 636–640, 2017.

[AHM+15] Peter Anderson, Negar M Harandi, Scott Moisik, Ian Stavness, and Sidney Fels. A comprehensive 3D biomechanically-driven vocal tract model including inverse dynamics for speech research. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[And82] Stephen R Anderson. The analysis of French schwa: or, how to get something for nothing. *Language*, pages 534–573, 1982.

[ATB+09] Michael Aron, Asterios Toutios, Marie-Odile Berger, Erwan Kerrien, Brigitte Wrobel-Dautcourt, and Yves Laprie. Registration of multimodal data for estimating the parameters of an articulatory model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taiwan Taipei, 2009. URL: `http://hal.inria.fr/inria-00350298/en/`.

[Ava09] Alireza Nasiri Avanaki. Exact global histogram specification optimized for structural similarity. *Optical review*, 16(6):613–621, 2009.

[Bac10] Jolanta Bachan. Efficient diphone database creation for mbrola, a multilingual speech synthesiser. In *XII International PhD Workshop (OWD 2010). Conference Archives PTETiS*, volume 28, pages 303–308, 2010.

[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.

[Ben80] Jon Louis Bentley. Multidimensional divide-and-conquer. *Communications of the ACM*, 23(4):214–229, 1980.

[BFS+] Charlotte Bellinghausen, Thomas Fangmeier, Bernhard Schröder, Johanna Keller, Susanne Drechsel, Peter Birkholz, Ludger Tebartz van Elst, and Andreas Riedel. On the role of disfluent speech for uncertainty in articulatory speech synthesis. In *The 9th Workshop on Disfluency in Spontaneous Speech*, page 39.

[BG92a] C. P. Browman and L. Goldstein. Articulatory phonology: an overview. Status report on speech research, Haskins Laboratory, 1992.

[BG92b] Catherine P Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180, 1992.

[BGHV01] Allen R Braun, Andre Guillemin, Lara Hosey, and Mary Varga. The neural organization of discourse: An h215o-pet study of narrative production in english and american sign language. *Brain*, 124(10):2028–2044, 2001.

[BH07] David J Brenner and Eric J Hall. Computed tomography—an increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007.

[BHG+16] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS computational biology*, 12(11):e1005119, 2016.

[BHS14] Tara McAllister Byun, Elaine R Hitchcock, and Michelle T Swartz. Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research*, 57(6):2116–2130, 2014.

[Bir07] Peter Birkholz. Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In *INTERSPEECH*, pages 2865–2868, 2007.

[Bir13a] Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

[Bir13b] Peter Birkholz. VocalTractLab 2.0: A tool for articulatory speech synthesis. Technische Universität Dresden. Online version: `http://www.vocaltractlab.de/index.php?page=vocaltractlab-about`, 2013. [Online; accessed 10-October-2019].

[BJ03] P. Birkholz and D. Jackel. A three-dimensional model of the vocal tract for speech synthesis. In *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain*, pages 2597–2600, Aug 2003.

[BJK06a] P. Birkholz, D. Jackèl, and B. J. Kröger. Construction and control of a three-dimensional vocal tract model. In *Proc. Intl. Conf. Acoust., Spch., and Sig. Proc. (ICASSP 2006)*, pages 873–876, 2006.

[BJK06b] Peter Birkholz, Dietmar Jackèl, and Bernd J 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.

[BLDO] Théo Biasutto-Lervat, Sara Dahmani, and Slim Ouni. Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning.

[BN08] Erik Bresch and Shrikanth Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE transactions on medical imaging*, 28(3):323–338, 2008.

[BSH07] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing*. Springer, 2007.

[BW18] Paul Boersma and David Weenink. TextGrid file formats. `http://www.fon.hum.uva.nl/praat/manual/TextGrid_file_formats.html`, 2018. [Online; accessed 30-June-2019].

[BZP08] Alan C Brooks, Xiaonan Zhao, and Thrasyvoulos N Pappas. Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions. *IEEE Transactions on image processing*, 17(8):1261–1273, 2008.

[Cal89a] Calliope. Description acoustique. In *La parole et son traitement automatique*, chapter 3. Masson, Paris, 1989.

[Cal89b] Calliope. *La parole et son traitement automatique*. Masson, Paris, 1989.

[CH68] Noam Chomsky and Morris Halle. *The sound pattern of English.* ERIC, 1968.

[DCGO19] Sara Dahmani, Vincent Colotte, Valérian Girard, and Slim Ouni. Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. *Proc. Interspeech 2019*, pages 2598–2602, 2019.

[DFF+19] Ioannis Douros, Jacques Felblinger, Jens Frahm, Karyna Isaieva, Arun A. Joseph, Yves Laprie, Freddy Odille, Anastasiia Tsukanova, Dirk Voit, and Pierre-André Vuissoz. A multimodal real-time MRI articulatory corpus of French for speech research. In *InterSpeech-20th Annual Conference of the International Speech Communication Association-2019*, 2019.

[DPIZC07] Nina F Dronkers, Odile Plaisant, Marie Therese Iba-Zizen, and Emmanuel A Cabanis. Paul broca's historic cases: high resolution mr imaging of the brains of leborgne and lelong. *Brain*, 130(5):1432–1441, 2007.

[DPP+96] Thierry Dutoit, Vincent Pagel, Nicolas Pierret, François Bataille, and Olivier Van der Vrecken. The mbrola project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1393–1396. IEEE, 1996.

[DR19] Maya Davis and Melissa A Redford. The emergence of discrete perceptual-motor units in a production model that assumes holistic phonological representations. *FRONTIERS IN PSYCHOLOGY*, 10, 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. In *InterSpeech-20th*

Annual Conference of the International Speech Communication Association-2019, 2019.

[EB11] Abdulkadir Eryildirim and Marie-Odile Berger. A guided approach for automatic segmentation and modeling of the vocal tract in mri images. In *2011 19th European Signal Processing Conference*, pages 61–65. IEEE, 2011.

[EL16a] 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.

[EL16b] Benjamin Elie and Yves Laprie. A glottal chink model for the synthesis of voiced fricatives. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5240–5244. IEEE, 2016.

[ELVO16] Benjamin Elie, Yves Laprie, Pierre-André Vuissoz, and Freddy Odille. High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data. In *Eusipco, Budapest*, pages 1353–1357, August 2016.

[Eng00] Olov Engwall. Are static MRI measurements representative of dynamic speech? results from a comparative study using MRI, EPG and EMA. In *Sixth International Conference on Spoken Language Processing*, 2000.

[EPZ+11] Byron D Erath, Sean D Peterson, Matías Zañartu, George R Wodicka, and Michael W Plesniak. A theoretical model of the pressure field arising from asymmetric intraglottal flows applied to a two-mass model of the vocal folds. *The Journal of the Acoustical Society of America*, 130(1):389–403, 2011.

[ESe19] David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). Ethnologue: Languages of the world. Twenty-second edition. Dallas, Texas: SIL International. Online version: `http://www.ethnologue.com`, 2019. [Online; accessed 22-June-2019].

[Eur17] Special Eurobarometer. 2012. Europeans and their languages. *European Commission*, 2017.

[FA04] Angela D Friederici and Kai Alter. Lateralization of auditory language functions: a dynamic dual pathway model. *Brain and language*, 89(2):267–276, 2004.

[Fan60] G. Fant. *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960.

[Fan71a] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[Fan71b] Gunnar Fant. The F-patterns of compound tube resonators and horns. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, 1971.

[FD00] Dan Foygel and Gary S Dell. Models of impaired lexical access in speech production. *Journal of Memory and Language*, 43(2):182–216, 2000.

[FDR04] Cécile Fougeron and Elisabeth Delais-Roussarie. Liaisons et enchaînements: «fais_en á fez_en parlant». *Actes des Journées d'Etudes sur la Parole*, pages 221–224, 2004.

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

[FL68] James L Flanagan and Lorinda L Landgraf. Self-oscillating source for vocal-tract synthesizers. *Audio and Electroacoustics, IEEE Transactions on*, 16(1):57–64, 1968.

[Fla13] James L Flanagan. *Speech analysis, synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

[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.

[Fou59] Pierre Fouché. *Traité de prononciation française.* Klincksieck, Paris, 1959.

[Fou01] Cécile Fougeron. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of phonetics*, 29(2):109–135, 2001.

[FS97] Cécile Fougeron and Donca Steriade. Does deletion of French schwa lead to neutralization of lexical distinctions? In *EUROSPEECH*, 1997.

[GB88] Terry L Gottfried and Patrice Speeter Beddor. Perception of temporal and spectral information in French vowels. *Language and Speech*, 31(1):57–75, 1988.

[GB11] Frank H Guenther and Jonathan S Brumberg. Brain-machine interfaces for real-time speech synthesis. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5360–5363. IEEE, 2011.

[GBW+09] Frank H Guenther, Jonathan S Brumberg, E Joseph Wright, Alfonso Nieto-Castanon, Jason A Tourville, Mikhail Panko, Robert Law, Steven A Siebert, Jess L Bartels, Dinal S Andreasen, et al. A wireless brain-machine interface for real-time speech synthesis. *PloS one*, 4(12):e8218, 2009.

[GGGG11] Maurice Grevisse, André Goosse, Maurice Grevisse, and Maurice Grevisse. *Le bon usage: grammaire langue française.* De Boeck., 2011.

[GLM12] Randall Gess, Chantal Lyche, and Trudel Meisenburg. *Phonological variation in French: Illustrations from three continents*, volume 11. John Benjamins Publishing, 2012.

[GN99] Fiona Gibbon and Katerina Nicolaidis. Palatography. *Coarticulation: Theory, data and techniques*, pages 229–244, 1999.

[Gra50] Maurice Grammont. *Traité de phonétique*. Librairie Delagrave, 1950.

[GWTPP06] Jean-Michel Gérard, Reiner Wilhelms-Tricarico, Pascal Perrier, and Yohan Payan. A 3D dynamical biomechanical tongue model to study speech motor control. *arXiv preprint physics/0606148*, 2006.

[Har76] William J Hardcastle. *Physiology of speech production: an introduction for speech scientists*. Academic Press, 1976.

[Har99] WJ Hardcastle. Electromyography. *Coarticulation: theory, data and techniques. University Press, Cambridge*, pages 270–283, 1999.

[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.

[HB19] Ian S Howard and Peter Birkholz. Modelling vowel acquisition using the birkholz synthesizer. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pages 304–311, 2019.

[Hic12] Gregory Hickok. Computational neuroanatomy of speech production. *Nature reviews neuroscience*, 13(2):135, 2012.

[Hic14] Gregory Hickok. Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience*, 29(1):52–59, 2014.

[HM05] MS Howe and RS McGowan. Aeroacoustics of [s]. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 461, pages 1005–1028. The Royal Society, 2005.

[HM08] Kiyoshi Honda and Shinji Maeda. Glottal-opening and airflow pattern during production of voiceless fricatives: a new non-invasive instrumentation. *The Journal of the Acoustical Society of America*, 123(5):3738–3738, 2008.

[HM11] Ian S Howard and Piers Messum. Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, 15(1):85–117, 2011.

[HP00] Gregory Hickok and David Poeppel. Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, 4(4):131–138, 2000.

[HP04] Gregory Hickok and David Poeppel. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99, 2004.

[HP07] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393, 2007.

[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.

[IF72] Kenzo Ishizaka and James L Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell system technical journal*, 51(6):1233–1268, 1972.

[Int99] International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[Jac01] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[JF00] Sun-Ah Jun and Cécile Fougeron. A phonological model of French intonation. In *Intonation*, pages 209–242. Springer, 2000.

[JHdA+13] Johan Jansson, Andreas Holmberg, Rodrigo Vilela de Abreu, Cem Degirmenci, Johan Hoffman, Mikael Karlsson, and Mats Abom. Adaptive stabilized finite element framework for simulation of vocal fold turbulent fluid-structure interaction. In *Proceedings of Meetings on Acoustics*, volume 19, pages 035–041. Acoustical Society of America, 2013.

[JMPSA06] Christiane Jadelot, Mathieu Mangeot, Etienne Petitjean, and Susanne Salmon-Alt. Morphalou 2.0. `https://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php`, 2006. [Online; accessed 30-June-2019].

[Jon56] Daniel Jones. *The pronunciation of English*, volume 369. Cambridge University Press, 1956.

[KBC99] William F Katz, Sneha V Bharadwaj, and Burkhard Carstens. Electromagnetic articulography treatment for an adult with broca's aphasia and apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 42(6):1355–1366, 1999.

[KEB08] Victor Kuperman, Mirjam Ernestus, and Harald Baayen. Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. *The Journal of the Acoustical Society of America*, 124(6):3897–3908, 2008.

[KG18] Advait Koparkar and Prasanta Kumar Ghosh. A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5004–5008. IEEE, 2018.

[KMG10] William F Katz, Malcolm R McNeil, and Diane M Garst. Treating apraxia of speech (aos) with ema-supplied visual augmented feedback. *Aphasiology*, 24(6-8):826–837, 2010.

[KN99] Barbara Kühnert and Francis Nolan. The origin of coarticulation. *Coarticulation: Theory, data and techniques*, pages 7–30, 1999.

[Kra05] Michael H Krane. Aeroacoustic production of low-frequency unvoiced speech sounds. *The Journal of the Acoustical Society of America*, 118(1):410–427, 2005.

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

[KSKM13] Juhani Knuuti, Antti Saraste, Marko Kallio, and Heikki Minn. Is cardiac magnetic resonance imaging causing DNA damage? *European heart journal*, 34(30):2337–2339, 2013.

[Lav94] John Laver. *Principles of phonetics*. Cambridge University Press, 1994.

[LB02] Edward Loper and Steven Bird. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[LB11a] 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.

[LB11b] 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.

[LD12] Peter Ladefoged and Sandra Ferrari Disner. *Vowels and consonants*. John Wiley & Sons, 2012.

[LET15] Yves Laprie, Benjamin Elie, and Anastasiia Tsukanova. 2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes. In *International Congress of Phonetic Sciences*, 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.

[Lev65] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *USSR's Reports of Science Academies*, 163(4):845–848, 1965.

[Lin91] Qiguang Lin. Speech production theory and articulatory speech synthesis. *The Journal of the Acoustical Society of America*, 90(4):2203–2203, 1991.

[LJ14] Peter Ladefoged and Keith Johnson. *A course in phonetics*. Cengage learning, 2014.

[LL70] Pierre R Léon and Monique Léon. *Introduction á la phonétique corrective á l'usage des professeurs de français á l'étranger*. Hachette, 1970.

[LM98] Peter Ladefoged and Ian Maddieson. The sounds of the world's languages. *Language*, 74(2):374–376, 1998.

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

[Low99] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[LQSN17] Adam C Lammert, Thomas F Quatieri, Christine H Shadle, and Shrikanth S Narayanan. Speed accuracy tradeoffs in speech production. Technical report, MIT LIncoln Laboratory Lexington United States, 2017.

[LRM99] Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38, 1999.

[LRP+13] Adam C Lammert, Vikram Ramanarayanan, Michael I Proctor, Shrikanth Narayanan, et al. Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In *Interspeech*, pages 959–962, 2013.

[LSF12] John E Lloyd, Ian Stavness, and Sidney Fels. Artisynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In *Soft tissue biomechanical modeling for computer assisted surgery*, pages 355–394. Springer, 2012.

[LSG04] Romary L., Salmon-Alt S., and Francopoulo G. Standards going concrete : from LMF to Morphalou. In *Workshop on Electronic Dictionaries, Coling*, Geneva, Switzerland, 2004.

[LSNQ18] Adam C Lammert, Christine H Shadle, Shrikanth S Narayanan, and Thomas F Quatieri. Speed-accuracy tradeoffs in human speech production. *PloS one*, 13(9):e0202180, 2018.

[LVC14] Y. Laprie, B. Vaxelaire, and M. Cadot. Geometric articulatory model adapted to the production of consonants. In *10th International Seminar on Speech Production (ISSP)*, Köln, Allemagne, May 2014. URL: `http://hal.inria.fr/hal-01002125`.

[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.

[Mae90a] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[Mae90b] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle and A. Marschal, editors, *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990.

[Mar94] Petros Maragos. Fractal signal analysis using mathematical morphology. In P. Hawkes and B. Kazan, editors, *Advances in Electronics and Electron Physics*, volume 88, chapter 4, pages 199–246. Academic Press, 1994.

[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.

[McG88] Richard S McGowan. An aeroacoustic approach to phonation. *The Journal of the Acoustical Society of America*, 83(2):696–704, 1988.

[MCTO11] Utpala Musti, Vincent Colotte, Asterios Toutios, and Slim Ouni. Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. In *Auditory-Visual Speech Processing 2011*, 2011.

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

[MIGG99] Ferdinando A Mussa-Ivaldi, N Gantchev, and G Gantchev. Motor primitives, force-fields and the equilibrium point theory. *From Basic Motor Control to Functional Recovery. Academic Publishing House" Prof. M. Drinov", Sofia, Bulgaria*, pages 392–398, 1999.

[MJB12] R.S. McGowan, M.T. Jackson, and M.A. Berger. Analyses of vocal tract cross-distance to area mapping: an investigation of a set of vowel images. *Journal of the Acoustical Society of America*, 131(1):424–434, 2012.

[MKF+10] MR McNeil, WF Katz, TRD Fossett, DM Garst, NJ Szuminsky, G Carter, and KY Lim. Effects of online augmented kinematic and perceptual feedback on treatment of speech movements in apraxia of speech. *Folia Phoniatrica et Logopaedica*, 62(3):127–133, 2010.

[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*, Lyon, France, August 2013. URL: `http://hal.inria.fr/hal-00836829`.

[ML17] Marius Muja and David G Lowe. Fast library for approximate nearest neighbors. *Dosegljivo: https://github. com/mariusmuja/flann*, 2017.

[MPH+17] Stefania Marin, Marianne Pouplier, Philip Hoole, Manfred Pastätter, Lasse Bombien, Ioana Chitoran, and Alexei Kochetov. Towards a typology of consonant coarticulation: gauging the space between universal and language-specific patterns of consonant timing. In *IPS Workshop on Abstraction, Diversity, and Speech Dynamics*, 2017.

[MTS+10] H. Nam V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Epsy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *11th Annual Conference of the International Speech Communication Association - INTERSPEECH 2010*, Makuhari, Chiba, Japan, 2010.

[MVL58] Paul Moore and Hans Von Leden. Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica*, 10(4):205–238, 1958.

[MZF19] Debasish Ray Mohapatra, Victor Zappi, and Sidney Fels. An extended two-dimensional vocal tract model for fast acoustic simulation of single-axis symmetric three-dimensional tubes. *arXiv preprint arXiv:1909.09585*, 2019.

[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.

[NMHJ+12] H. Nam, V. Mitra, M. Hasegawa-Johnson, C. Epsy-Wilson, E. Saltzman, and L. Goldstein. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America*, 132(6):3080–3989, 2012.

[NMT+12] Hosung Nam, Vikramjit Mitra, Mark Tiede, Mark Hasegawa-Johnson, Carol Espy-Wilson, Elliot Saltzman, and Louis Goldstein. A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America*, 132(6):3980–3989, 2012.

[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.

[Öhm66] S.E. Öhman. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39(1):151–168, 1966.

[Öhm67] Sven EG Öhman. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2):310–320, 1967.

[Ope13] OpenStax College. Organs and structures of the respiratory system. `http://cnx.org/contents/t2sgkCQ-@8/Organs-and-Structures-of-the-R`, 2013.

[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.

[PB15] Simon Preuß and Peter Birkholz. Optical sensor calibration for electro-optical stomatography. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[PBKN10] Michael I Proctor, Daniel Bone, Athanasios Katsamanis, and Shrikanth S Narayanan. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[PBL13] Jonathan L Preston, Nickole Brick, and Nicole Landi. Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology*, 2013.

[Per17] Pascal Perrier. What goals for articulatory speech synthesis? In *The 11th International Seminar on Speech Production*, 2017.

[PLM17] Jonathan L Preston, Megan C Leece, and Edwin Maas. Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language & Communication Disorders*, 52(1):80–94, 2017.

[PMRC+14] Jonathan L Preston, Patricia McCabe, Ahmed Rivera-Campos, Jessica L Whittle, Erik Landry, and Edwin Maas. Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research*, 57(6):2102–2115, 2014.

[PP14] Manfred Pastätter and Marianne Pouplier. The articulatory modeling of german coronal consonants using tada. In *Proceedings of the 10th International Seminar on Speech Production*, pages 308–311, 2014.

[PP15] Manfred Pastätter and Marianne Pouplier. Onset-vowel timing as a function of coarticulation resistance: Evidence from articulatory data. In *ICPhS*, 2015.

[PVC+96] Xavier Pelorson, C Vescovi, E Castelli, A Hirschberg, APJ Wijnands, and HMA Bailliet. Description of the flow through in-vitro models of the glottis during phonation. application to voiced sounds synthesis. *Acta Acustica united with Acustica*, 82(2):358–361, 1996.

[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.

[RMC19] Qinwan Rabbani, Griffin Milsap, and Nathan E Crone. The potential for a speech brain–computer interface using chronic electrocorticography. *Neurotherapeutics*, 16(1):144–165, 2019.

[RPVH+07] Nicolas Ruty, Xavier Pelorson, Annemie Van Hirtum, Ines Lopez-Arteaga, and Avraham Hirschberg. An in vitro setup to test the relevance and the accuracy of low-order vocal folds models. *The Journal of the Acoustical Society of America*, 121(1):479–490, 2007.

[RRUC13] Zeynab Raeesy, Sylvia Rueda, Jayaram K Udupa, and John Coleman. Automatic segmentation of vocal tract MR images. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1328–1331. IEEE, 2013.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[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.

[RVSN16] Vikram Ramanarayanan, Maarten Van Segbroeck, and Shrikanth S Narayanan. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer speech & language*, 36:330–346, 2016.

[Rya13] Camille L Ryan. Language use in the united states: 2011. *Economics and Statistics Administration*, 2013.

[S+85] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985.

[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.

[SB16] Simon Stone and Peter Birkholz. Angle correction in optopalatographic tongue distance measurements. *IEEE Sensors Journal*, 17(2):459–468, 2016.

[Sch21] William L Schwartz. Syllabication in French and suggestions for accenting the letter E. *The Modern Language Journal*, 5(7):374–377, 1921.

[SGBR88] E Saltzman, L Goldstein, C Browman, and P Rubin. Modeling speech production using dynamic gestural structures. *The Journal of the Acoustical Society of America*, 84(S1):S146–S146, 1988.

[SK87] Elliot Saltzman and JA Kelso. Skilled actions: a task-dynamic approach. *Psychological review*, 94(1):84, 1987.

[SLMD02] A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function tranformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36(3-4):169–180, March 2002.

[SM89] Elliot L Saltzman and Kevin G Munhall. A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382, 1989.

[SMB18] Simon Stone, Michael Marxen, and Peter Birkholz. Construction and evaluation of a parametric one-dimensional vocal tract model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1381–1392, 2018.

[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.

[SSDW+01] Ronald C Scherer, Daoud Shinwari, Kenneth J De Witt, Chao Zhang, Bogdan R Kucinschi, and Abdollah A Afjeh. Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees. *The Journal of the Acoustical Society of America*, 109(4):1616–1630, 2001.

[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.

[Sto13] Brad H Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer speech & language*, 27(4):989–1010, 2013.

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

[SVH19] Ryan Stokes, Jonathan H Venezia, and Gregory Hickok. The motor system's [modest] contribution to speech perception. *Psychonomic bulletin & review*, 2019.

[SWF+18] Zhihua Su, Jianguo Wei, Qiang Fang, Jianrong Wang, and Kiyoshi Honda. Tongue segmentation with geometrically constrained snake model. In *Interspeech*, pages 3117–3121, 2018.

[Sze10] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2010.

[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.

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

[Tho86] TJ Thomas. A finite element model of fluid flow in the vocal tract. *Computer Speech & Language*, 1(2):131–151, 1986.

[Tit73] Ingo R Titze. The human vocal cords: a mathematical model. *Phonetica*, 28(3-4):129–170, 1973.

[TM98] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Iccv*, volume 1, page 2, 1998.

[TN15] Asterios Toutios and Shrikanth Narayanan. Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data. In *ICPhS*, 2015.

[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.

[TNT+13] Keiichi Tokuda, Yoshihiko Nankaku, Takechi Toda, Heishun Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013. URL: http://dx.doi.org/10.1109/JPROC.2013.2251852, doi:10.1109/JPROC.2013.2251852.

[Tra87] Bernard Tranel. *The sounds of French: An introduction*. Cambridge university press, 1987.

[TSS+16] Asterios Toutios, Tanner Sorensen, Krishna Somandepalli, Rachel Alexander, and Shrikanth S Narayanan. Articulatory synthesis based on real-time magnetic resonance imaging data. In *INTERSPEECH*, pages 1492–1496, 2016.

[TT90] Herbert M. Teager and Shushan M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.

[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.

[W+97] John C Wells et al. Sampa: computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4, 1997.

[WBSS04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[WECT19] Matthew Walenski, Eduardo Europa, David Caplan, and Cynthia K Thompson. Neural networks for sentence comprehension and production: An ale-based meta-analysis of neuroimaging studies. *Human brain mapping*, 2019.

[WH16] Grant M Walker and Gregory Hickok. Bridging computational approaches to speech production: The semantic–lexical–auditory–motor model (slam). *Psychonomic bulletin & review*, 23(2):339–352, 2016.

[WH18] GRANT WALKER and GREGORY HICKOK. Speech production integrating psycholinguistic, neuroscience, and motor control perspectives. *The Oxford Handbook of Psycholinguistics*, 2018.

[WHS19] Victor Wetzel, Thomas Hélie, and Fabrice Silva. Power balanced time-varying lumped parameter model of a vocal tract: modelling and simulation. 2019.

[WSB03] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003.

[WSB17] Benjamin Weitz, Ingmar Steiner, and Peter Birkholz. Gesture-based articulatory text to speech synthesis. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pages 324–331, 2017.

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

[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.

[YNW19] Tsukasa Yoshinaga, Kazunori Nozaki, and Shigeo Wada. A simplified vocal tract model for articulation of [s]: The effect of tongue tip elevation on [s]. *PLOS ONE*, 14(10):e0223382, 2019.

[Zem10] Willard R Zemlin. *Speech and Hearing Science, Anatomy and Physiology*. Pearson Education (US), Fourth Edition edition, 2010.

[Zen06] Heiga Zen. An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, 133, 2006.

Résumé

Cette thèse se situe dans le domaine de la synthèse articulatoire de la parole et est organisé en trois grandes parties : les deux premières sont consacrées au développement de deux synthétiseurs articulatoires de la parole ; la troisième traite des liens que l'on peut établir entre les deux approches utilisées.

Le premier synthétiseur est issu d'une approche à base de règles. Celle-ci visait à obtenir le contrôle complet sur les articulateurs (mâchoire, langue, lèvres, vélum, larynx et épiglotte). Elle s'appuyait sur des données statiques du plan sagittal médian obtenues par IRM (Imagerie par Résonance Magnétique) correspondant à des articulations bloquées de voyelles du français, ainsi que des syllabes de type consonne-voyelle, et était composée de plusieurs étapes : l'encodage de l'ensemble des données grâce à un modèle du conduit vocal basé sur l'ACP (analyse en composantes principales) ; l'utilisation des configurations articulatoires obtenues comme sources de positions à atteindre et destinées à piloter le synthétiseur à base de règles qui est la contribution principale de cette première partie ; l'ajustement des conduits vocaux obtenus selon une perspective phonétique ; la simulation acoustique permettant d'obtenir un signal acoustique. Les résultats de cette synthèse ont été évalués de manière visuelle, acoustique et perceptuelle, et les problèmes rencontrés ont été identifiés et classés selon leurs origines, qui pouvaient être : les données, leur modélisation, l'algorithme contrôlant la forme du conduit vocal, la traduction de cette forme en fonctions d'aire, ou encore la simulation acoustique. Ces analyses nous permettent de conclure que, parmi les test effectués, les stratégies articulatoires des voyelles et des occlusives sont les plus correctes, suivies par celles des nasales et des fricatives.

La seconde approche a été développée en s'appuyant sur un synthétiseur de référence constitué d'un réseau de neurones feed-forward entraîné à l'aide de la méthode standard du système Merlin [WWK16] sur des données audio composées de parole en langue française enregistrée par IRM en temps réel. Ces données ont été segmentées phonétiquement et linguistiquement. Ces données audio, malgré un débruitage, étaient fortement parasitées par le son de la machine à IRM. Nous avons complété le synthétiseur de référence en ajoutant huit paramètres représentant de l'information articulatoire : l'ouverture des lèvres et leur protrusion, la distance entre la langue et le vélum, entre le vélum et la paroi pharyngale, et enfin entre la langue et la paroi pharyngale. Ces paramètres ont été extraits automatiquement à partir des images et alignés au signal et aux spécifications linguistiques. Les séquences articulatoires et les séquences de parole, générées conjointement, ont été évaluées à l'aide de différentes mesures : distance de déformation temporelle dynamique, la distortion mel-cepstrum moyenne, l'erreur de prédiction de l'apériodicité, et trois mesures pour F0 : RMSE (root mean square error), CORR (coéfficient de corrélation) and V/UV (frame-level voiced/unvoiced error). Un analyse de la pertinence des paramètres articulatoires par rapport aux labels phonétiques a également été réalisée. Elle permet de conclure que les paramètres articulatoires générés s'approchent de manière acceptable des paramètres originaux, et que l'ajout des paramètres articulatoires n'a pas dégradé le modèle acoustique original.

Les deux approches présentées ci-dessus ont en commun l'utilisation de deux types de données IRM. Ce point commun a motivé la recherche, dans les données temps réel, des images clés,

c'est-à-dire les configurations statiques IRM, utilisées pour modéliser la coarticulation. Afin de comparer les images IRM statiques avec les images dynamiques en temps réel, nous avons utilisé plusieurs mesures : la similarité structurelle, la distance du "terrassier" et SIFT ; après avoir vérifié la pertinence et la validité de ces mesures, j'ai étudié qualitativement et quantitativement, puis interprété leur comportement; j'ai ensuite analysé leur similarités. J'en ai conclu que SIFT et la similarité structurelle capturaient bien les informations articulatoires et que leur comportement, de manière générale, validaient les données d'IRM statiques. Les phonèmes et traits phonétiques problématiques que j'ai pu identifier à travers les analyses des distributions et incompatibilités de mesures étaient les liquides /l, ʁ/, dont la production dynamique ne pouvait être imitée par leur simulation statique, les fricatives alvéolaires /s, ʃ/, elles aussi simulées de manière non réaliste dans le contexte statique, et les caractéristiques de la nasalité. Il semblerait que les sons du corpus statique aient été légèrement trop nasalisés, et que réciproquement les sons nasalisés présentaient une ouverture vélopharyngée insuffisante. Enfin, j'ai discuté l'impact de cette étude pour de futurs synthétiseurs articulatoires hybrides de la parole.

Mots-clés: synthèse articulatoire, articulation, conduit vocal, IRM, IRM dynamique

Abstract

The thesis is set in the domain of articulatory speech synthesis and consists of three major parts: the first two are dedicated to the development of two articulatory speech synthesizers and the third addresses how we can relate them to each other.

The first approach results from a rule-based approach to articulatory speech synthesis that aimed to have a comprehensive control over the articulators (the jaw, the tongue, the lips, the velum, the larynx and the epiglottis). This approach used a dataset of static mid-sagittal magnetic resonance imaging (MRI) captures showing blocked articulation of French vowels and a set of consonant-vowel syllables; that dataset was encoded with a PCA-based vocal tract model. Then the system comprised several components: using the recorded articulatory configurations to drive a rule-based articulatory speech synthesizer as a source of target positions to attain (which is the main contribution of this first part); adjusting the obtained vocal tract shapes from the phonetic perspective; running an acoustic simulation unit to obtain the sound. The results of this synthesis were evaluated visually, acoustically and perceptually, and the problems encountered were broken down by their origin: the dataset, its modeling, the algorithm for managing the vocal tract shapes, their translation to the area functions, and the acoustic simulation. We concluded that, among our test examples, the articulatory strategies for vowels and stops are most correct, followed by those of nasals and fricatives.

The second explored approach started off a baseline deep feed-forward neural network-based speech synthesizer trained with the standard recipe of Merlin [WWK16] on the audio recorded during real-time MRI (RT-MRI) acquisitions: denoised (and yet containing a considerable amount of noise of the MRI machine) speech in French and force-aligned state labels encoding phonetic and linguistic information. This synthesizer was augmented with eight parameters representing articulatory information—the lips opening and protrusion, the distance between the tongue and the velum, the velum and the pharyngeal wall and the tongue and

the pharyngeal wall—that were automatically extracted from the captures and aligned with the audio signal and the linguistic specification. The jointly synthesized speech and articulatory sequences were evaluated objectively with dynamic time warping (DTW) distance, mean mel-cepstrum distortion (MCD), BAP (band aperiodicity prediction error), and three measures for F0: RMSE (root mean square error), CORR (correlation coefficient) and V/UV (frame-level voiced/unvoiced error). The consistency of articulatory parameters with the phonetic label was analyzed as well. I concluded that the generated articulatory parameter sequences matched the original ones acceptably closely, despite struggling more at attaining a contact between the articulators, and that the addition of articulatory parameters did not hinder the original acoustic model.

The two approaches above are linked through the use of two different kinds of MRI speech data. This motivated a search for such coarticulation-aware targets as those that we had in the static case to be present or absent in the real-time data. To compare static and real-time MRI captures, the measures of structural similarity, Earth mover's distance, and SIFT were utilized; having analyzed these measures for validity and consistency, I qualitatively and quantitatively studied their temporal behavior, interpreted it and analyzed the identified similarities. I concluded that SIFT and structural similarity did capture some articulatory information and that their behavior, overall, validated the static MRI dataset. The problematic sounds and features that I was able to identify through the analysis of measure distributions and mismatches were the liquids /l, ʁ/, whose dynamic production could not be matched by their static simulation, the alveolar fricatives /s, ʃ/, again, simulated unrealistically in the static setting, and the feature of nasality: apparently, the oral sounds in the static corpus were slightly too nasalized, and in the nasal sounds, vice versa, the velopharyngeal port did not open enough. Finally, I commented on the repercussions of the study for potential hybrid articulatory speech synthesizers.

Keywords: articulatory speech synthesis, articulation, vocal tract, MRI, RT-MRI