A Coarticulation Model for Articulatory Speech Synthesis

Anastasiia Tsukanova

MSc. Dissertation

Department of Intelligent Computer Systems
Faculty of Information and Communication Technology
University of Malta
2016

Supervisor:
Yves Laprie, Centre National de la Recherche Scientifique, France
Local advisor:
Michael Rosner, Department of Intelligent Computer Systems, University of Malta

Submitted in partial fulfilment of the requirements for the Degree of European Master of Science in Human Language Science and Technology

Abstract

The state-of-the-art techniques for speech synthesis rely either on concatenation of acoustic units taken from a vast pre-recorded speech database noting the relevant linguistic information or on statistical generation of the necessary acoustic parameters and using a speech production model. These approaches yield synthesis of good quality, but are purely technical solutions which bring no or very little information about the acoustics of speech or about how the articulators (mandible, tongue, lips, velum...) are controlled.

In contrast, the articulatory approach generates the speech signal from the vocal tract shape and its modelled acoustic phenomena. The vocal tract deformation control comprises slow anticipation of the main constriction and fast and imperatively accurate aiming for consonants.

The system predicts the sequence of vocal tract consecutive configurations from a sequence of phonemes of the French language to be articulated and a model of the coarticulation effects in it. We use static magnetic resonance imaging (MRI) captures of the vocal tract shape when producing phonemes in various contexts, thus following an approach by Birkholz (2013). The evaluation of the model is done both on the animated graphics representing the vocal tract shape evolution (how natural and efficient the movement is) and on the synthesised speech signals that are perceptively and—in terms of formants—qualitatively compared to identical utterances made by a human.

Our results show that there are a lot of effects in the dynamic process of speech that manage to be reproduced by manipulating solely static data. We discuss generation of pure vowels, vowel-to-vowel and vowel-consonant-vowel transitions, and articulators' behaviour in phrases, report which acoustic properties have been rendered correctly and what could be the reasons for the system to fail to produce the desired result in other cases, and ponder how to reduce the after-effects of target-oriented moves to obtain a more gesture-like motion.

M.Sc. (HLST)

FACULTY OF INFORMATION
AND COMMUNICATION TECHNOLOGY
UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own work, of work of another person, whether or not such work has been published" (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master's dissertation submitted is my own work, except where acknowledged and referenced. No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

As a Master's student, as per Regulation 58 of the General Regulations for University Postgraduate Awards, I accept that should my dissertation be awarded a Grade A, it will be made publicly available on the University of Malta Institutional Repository.

Student Name:	Anastasiia Tsukanova
Student's ID:	1403343
Course Code:	CSA5310 HLST Dissertation
Title of work:	A Coarticulation Model for Articulatory Speech Synthesis

Signature of Student:

Date:
February 12, 2016

Acknowledgements

I would like to thank my supervisor, Yves Laprie, for his guidance, immense help and unwavering belief in me, as well as my local advisor in Malta, Michael Rosner, who has been very supportive and in particular helped me out with various organisational issues at different stages of the work.

Furthermore, I would like to thank Benjamin Elie for kindly allowing me to test my model in his acoustic simulation program (Elie and Laprie, 2015, 2016), and our subjects, Antoine Liutkus and Quentin Brabant, for the data.

I am also grateful to the European Masters Program in Language and Communication Technologies for granting me a chance to be a student here.

Finally, I thank my parents and friends: their love, support, and sense of humour were invaluable. Special thanks to Anastasiia Demikhovska for her help with proofreading.

Contents

List of Figures

List of Tables

1 Introduction

Being social animals as Aristotle put it, humans need to interact with each other, to share ideas and defend their views, to impress the others and get impressed themselves. As speech is one of the primary means of communication, the interest in its multiple facets naturally never fades.

Speech processing technology dates back to the 1940s. It deals with automatic speech understanding, generating, and coding in a variety of contexts ranging from interactions between humans (face-to-face communication, communication at a distance or in a group) to the ones between a human and a machine. The progress varies with particular tasks, yet there is an area which has proved to be especially challenging: *speech synthesis*.

The motivation for computerised generation of speech, first and foremost, lies in the realm of facilitating the experiences of visually impaired people. Then this research area prepares the ground for various commercial products such as call-centre automation or supplements to mobile applications—for example, readers of text messages, route directions, news articles, weather forecasts, etc. Finally, it is of scientific interest by itself and has the potential to be ubiquitously used for language learning and even for medical purposes to analyse and treat speech pathologies or consequences of operations on the vocal tract.

Over the years of research, there were numerous attempts made to generate human-like speech. One governing line of thought was to concentrate on the biological mechanism of speech production, study the articulatory movements and, given a particular geometric shape of the vocal tract, model the acoustic phenomena that give rise to the corresponding speech signal. The other one took notice that there are physiological effects which do not matter perceptually; we may obtain perceptually identical utterances in too many ways, or we even may overarticulate. Therefore it was suggested to disregard the knowledge of human body and merely mimic the final speech product instead, i.e. devise terminal analog synthesis models.

For a long time the research yielded no results that would sound natural until the late 1980s brought the idea to concatenate basic speech units—usually clippings of consecutive phoneme pairs that are called *diphones*. By the early 1990s the diphone selection techniques developed into the so-called "unit selection" method of concatenative speech synthesis which was able to spread on a larger number of phonemes and took into account numerous relevant speech features: lexical stress, pitch, part-of-speech information, position within a speech segment. To allow for more major transformations of the recorded signals into the resulting waveform, the unit selection method served as a basis for statistical parametric speech synthesis (late 1990s). One such kind of speech synthesis has turned out to be especially fruitful: HMM-based speech synthesis, where the generative model is a hidden Markov model (HMM). It has allowed to produce highly intelligible and natural synthetic speech signals. As yet the unit selection method and the HMMs are the state of the art in the field, depending on the task at hand.

However, even when the quality reached by HMM-based speech synthesis is good, this approach still has a flaw: being a rather technical solution, in its direct application

it loses all or almost all information on the acoustics of the synthetic speech or the way the *articulators* (speech organs—the lips, teeth, tongue, alveolar ridge, hard palate, velum, uvula, epiglottis, and glottis) are configured, how the active articulators move to produce a particular utterance.

As a consequence, many goals of computerised speech synthesis that were discussed above cannot be reached, and if a usage scenario requires a speech synthesis system to modify the characteristics of the recorded voice (e.g. add shrilling or produce the effect of a hoarse voice), mirror some anatomical changes after surgery on a patient, or change the speaker completely, the number of tools to do this and still retain the high quality of the synthetic speech is rather limited. Basically it is beyond the resources of a speech synthesiser to reliably produce a signal that would be more than a mere concatenation of the smaller units.

Alternatively, the articulatory approach offers a full control of the vocal tract, generating the voice from the temporal evolution of the vocal tract geometry. If any parameter needs changing within the natural limits, it should be attainable in no matter of time. In fact, this becomes another facet of the problem: there may be too many parameters to control; but implementations which resolve this by modelling the speech process in a simplified way still belong to the realm of this approach.

Apart from this advantage, the articulatory approach can give interesting insights into speech production and assist in other speech-related studies.

1.1 Problematics of the Articulatory Speech Synthesis. Aims and Objectives of the Present Work

When modelling the vocal tract transition for the articulatory approach, we need to learn how to progress through the sequence of phonemes that we would like the system to articulate: what the units in speech articulation are, how they are related to the phonemes in the sequence, and how to join them.

The problem is, while written speech is sequential together with the transcription of an utterance to be articulated, the oral speech is quasi-sequential at best. Its phases, usually referred to as respiration, phonation, articulation, and resonance, do not occur one by one. When positioned close, the allegedly separate phonemes influence each other heavily because of the temporal overlaps in speech production and the way the structures of the speech system influence each other. So, the same phonological segment, but in different contexts will be articulated differently. (For instance, the English phoneme /k/ is not realized the same way when followed by /i/ and when followed by /ɔ/: in the first case (/kiː/), the place of vocal tract constriction is brought forward as in comparison to the second one (/kɔː/).) This phenomenon, called *coarticulation*, is a non-trivial process with its own mechanism; there are both patterns and irregularities in it. Especially pronounced are the effects of anticipation of the phonemes to come; the carryover impact of the previous ones is considered to be reduced to passive inertia.

So the aim of this work is to model anticipatory coarticulation for articulatory speech synthesis.

The objectives are to translate the recorded database of ca. 100 static magnetic resonance imaging (MRI) scans—each of them captures the vocal tract configuration of a French native speaker when he produces a particular phoneme in a particular vowel context—to a database of parameter sequences representing the articulators' positioning in each case; extrapolate the database to estimate the missing samples; implement a coarticulation-aware algorithm for sequencing the samples to form utterances; use it in an existing speech synthesis system; analyse the results brought by the tested approaches.

1.2 Structure of the Report

Chapter 2 is dedicated to a more detailed and in-depth survey of the research area. In this chapter, Section 2.1 covers the mechanism of speech production within the source-filter theory, going through speech production in the anatomical down-top order, just as speech is produced: the human mechanism of respiration is described in Section 2.1.1, proceeding to phonation in Section 2.1.2 and articulation in Section 2.1.3. Sections 2.2 and 2.3 relate how speech sounds are produced by the articulators overviewed back in Section 2.1.3 and what information is essential to French phonetics. Within the latter section, Sections 2.3.1, 2.3.2, and 2.3.3 are dedicated to an overview of what kinds of sounds are found in the French language, what temporal patterns in the spoken French are there, and what prosodic cues and speech segments can be found in a spoken utterance respectively. Then we proceed to an overview of articulatory speech synthesis: Section 2.4.1 briefly reviews what kinds of articulatory data are available for research; Section 2.4.2 explains how this collected data can be modelled and then used in speech synthesis systems, and Section 2.4.3 explains the notions of coarticulation and assimilation and how coarticulation models can be implemented in speech synthesis.

Chapter 3 relates the work that has been done: Section 3.1 explains the approach by means of an illustrative example, and Section 3.2 is dedicated to the dataset—modelling it (Section 3.2.1) and extrapolating from it to the missing samples (Section 3.2.2). Section 3.3 brings all aspects of the work together and describes processing of an utterance: parsing the user's input (see Section 3.3.1), modelling coarticulation (Section 3.3.2), obtaining the area functions (Section 3.3.3), controlling subglottal and supraglottal pressure (Section 3.3.4) and vocal folds oscillations (Section 3.3.5) along with the voice pitch (Section 3.3.6). Section 3.4 summarises the achieved results.

These results are evaluated in the next chapter, Chapter 4. First it discusses the collected data and its modelling (Section 4.1) and reviews the quality of dataset expansion (Section 4.1.2). Then we select evaluation criteria and an evaluation set (Section 4.2), and analyse the synthesis results in Section 4.3.

This is followed by conclusions found in Chapter 5 where we briefly overview what has been achieved and what problems have been encountered, which brings us to a discussion in Section 5.1 on what can be done in future work.

The appendix gives detailed information on the main program structure.

Bibliography lists the resources cited in the report.

2 Background and Context

2.1 Mechanism of Speech Production

Speech is produced by operating the speech organs. This ability is not inborn; in fact, by nature, even though other species have also developed their kinds of informative verbal communication, the speech organs were not meant to be used to produce meaningful sounds: the system is constructed for breathing and eating. This explains why speech requires cognitive control (otherwise it is not speaking but babbling) and why it must be learned and adapted to every newly acquired language (Steinberg et al., 2013). 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. As for auditory feedback, the speaker is more conscious of it, 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 (Zemlin, 2010), 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.

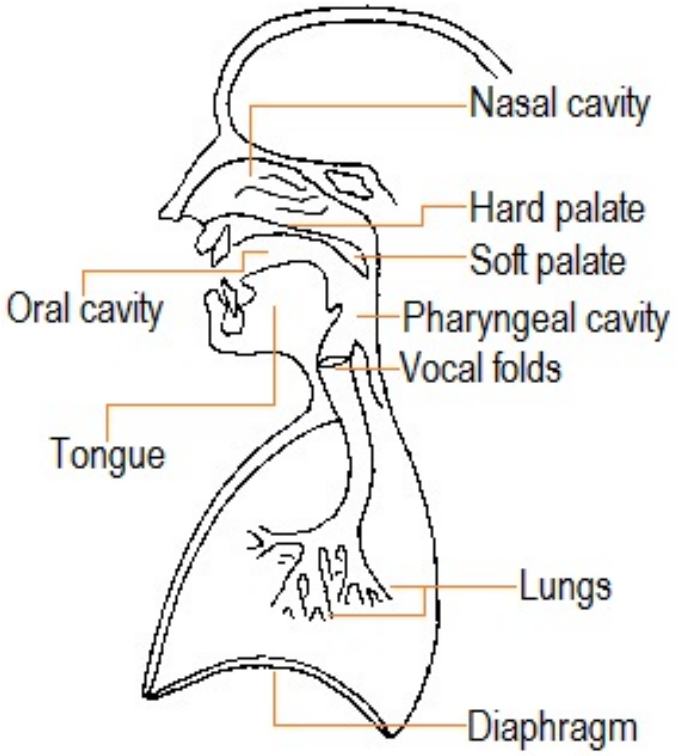


Figure 1: Human speech mechanism (Zemlin, 2010)

Fig. 1 schematically illustrates the organs involved in speech production. This is a mid-saggital section of an adult's vocal tract.

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:

$$| P(f) | = | U(f) | \cdot | H(f) | \cdot | R(f) |, \tag{1}$$

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 (Fant, 1971a).

2.1.1 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 (Zemlin, 2010).

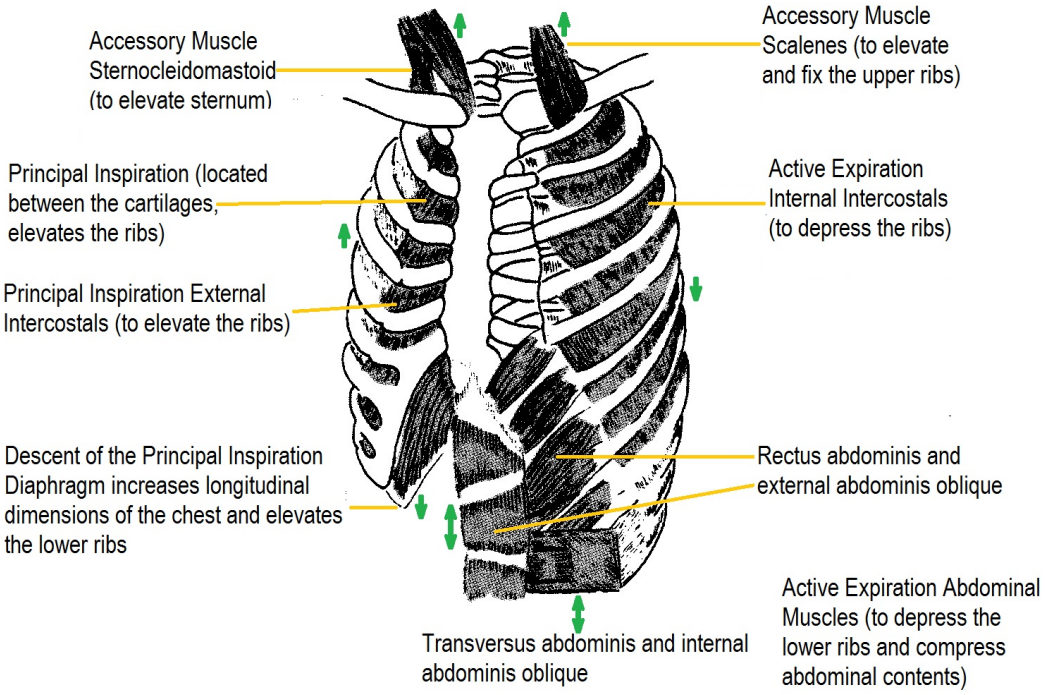


Figure 2: The principal muscles of respiration (Zemlin, 2010)

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 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). 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 (Zemlin, 2010). 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 (Flanagan, 2013). 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) (Zemlin, 2010). Most of the utterances are pronounced on expiration; ingressive sounds, for which the airstream flows inward through the mouth or nose, are rare (Ladefoged and Maddieson, 1998). So in order to avoid exhausting the drawn air amidst 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.

2.1.2 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; normally this frequency is smaller than the harmonic frequency of the vocal folds. 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, or as fundamental frequency measured in Hz—cycles per second. 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 (Zemlin, 2010).

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 (Moore and Von Leden, 1958).

Mathematical models of the larynx for speech simulation include a single-degree-of-freedom model by Flanagan and Landgraf (1968) 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 behaviour of the real larynx; two-degree-of-freedom models such as by Ishizaka and Flanagan (1972) 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 artefacts and unrealistic consequences for the parameters; and the sixteen-mass model by Titze (1973) 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 (Zemlin, 2010):

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 build of the further vocal tract rather than the larynx, which brings us to the next section.

2.1.3 Articulation

So, phonation involved vibrations of the vocal folds that can essentially be summarised 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 pitch 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. It comes just above the glottis and is an acoustic tube, around 17 cm long for an adult male, with a varying cross-sectional area. The glottis end of this tube has to be represented as closed, because in comparison to the mouth opening and nostrils, the resistance at the glottis is very high. 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. 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 Thomas (1986). Then these findings were theoretically supported by Teager and Teager (1990) and McGowan (1988). 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 (Maragos, 1994).

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:

$$Re = \frac{intertial\ forces}{viscous\ forces} = \rho \frac{UL}{\mu}, \tag{2}$$

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. Section 2.2 is devoted to the phonetic details (Ladefoged and Disner, 2012).

Fig. 3 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).

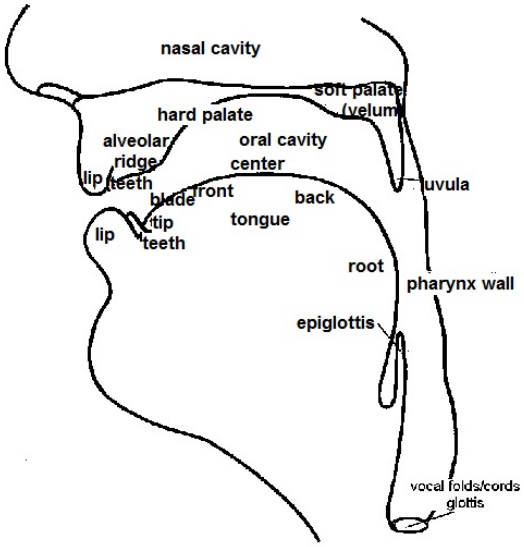


Figure 3: Human vocal organs (Ladefoged and Johnson, 2014)

- 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.

[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.

- 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 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. 4), 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. 5).

- 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, centre, and back further then. The tongue moves thanks to a 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. Hardcastle (1976) proposes only seven parameters that allow for most tongue movements: 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.

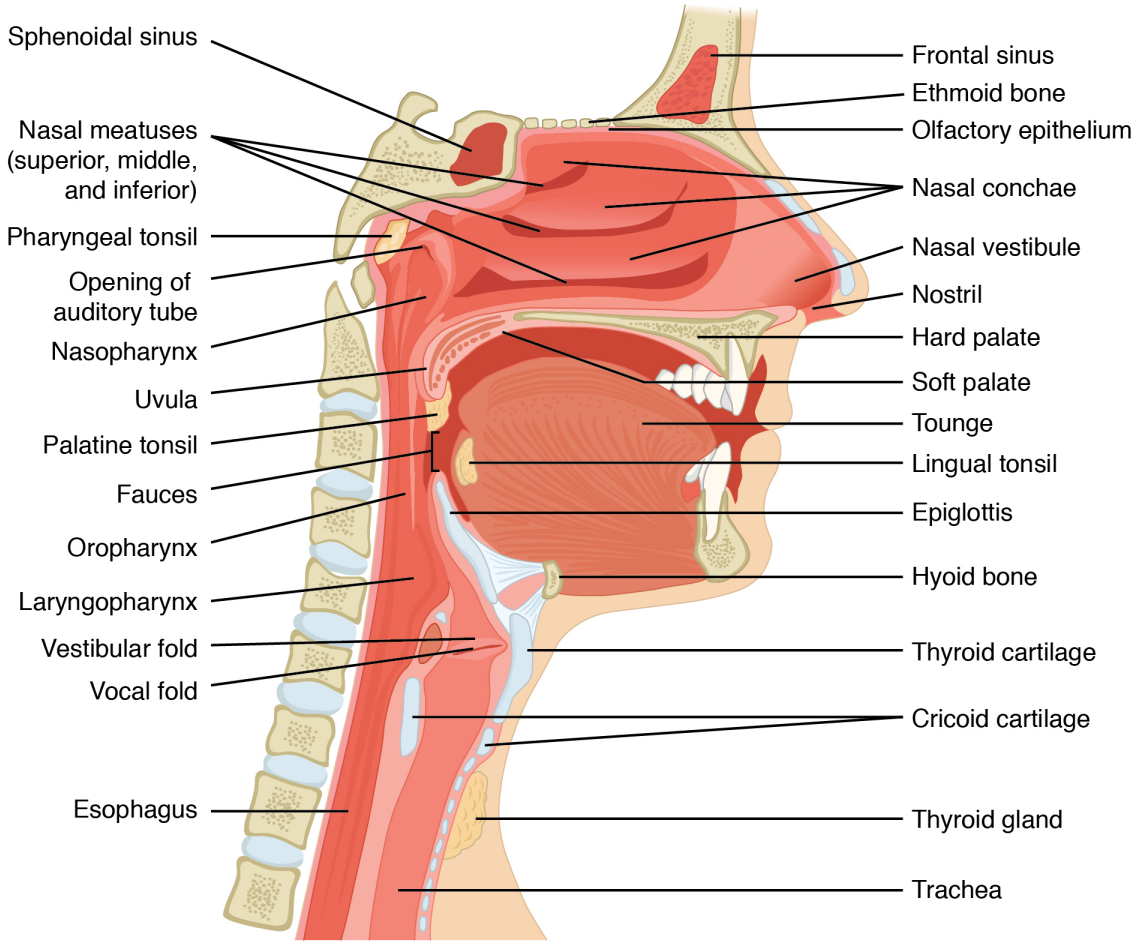


Figure 4: Upper airway (OpenStax College, 2013)

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 a nasal sound (then the velum goes up) or let the air through (then the velum goes down). 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. The French language does not feature pharyngeal sounds. 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.

2.2 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.

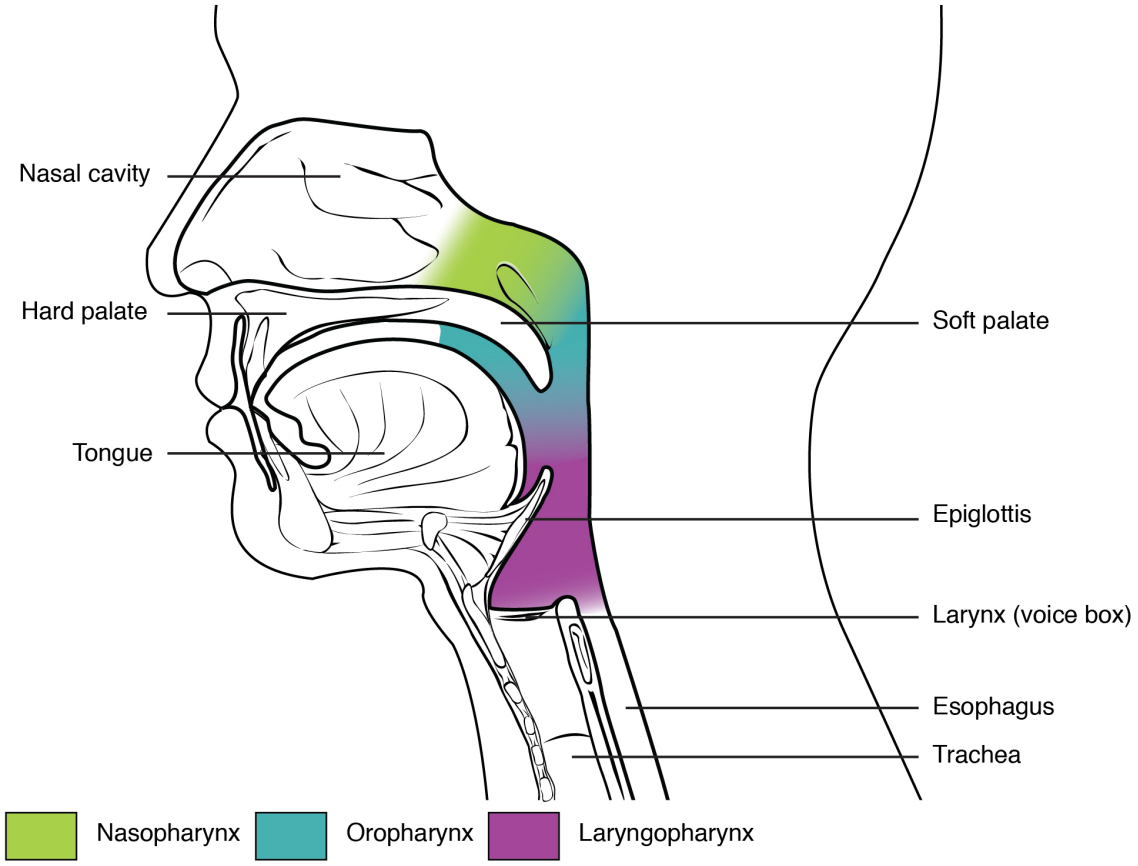


Figure 5: The division of the pharyngeal cavity into the nasopharynx, oropharynx, and laryngopharynx (OpenStax College, 2013)

The smallest distinctive unit of speech is called *phoneme*. It does not carry a meaning of its own, but can distinguish at least one word from another in a particular language. 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—International Phonetic Association (1999)—have introduced a special alphabet to note transcriptions.

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

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

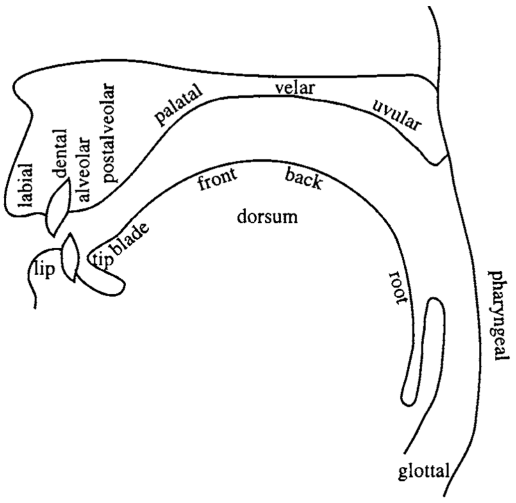


Figure 6: Mid-sagittal section of the vocal tract with labels for place of articulation (International Phonetic Association, 1999).

Vowels are articulated with an open vocal tract where the air flows 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 (Ladefoged and Johnson, 2014).

Vowels can be classified (International Phonetic Association, 1999):

- *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, which is not directly 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. 7.

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. Jones (1956) suggested a set of eight vowels, and Chomsky and Halle (1968) updated this notion. Three vowels in this set, /ɑ/, /i/, and /u/, have articulatory definitions 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.

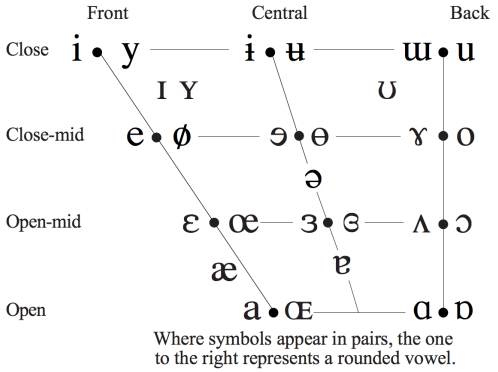


Figure 7: The vowels' chart based on their main features (International Phonetic Association, 1999).

Consonants can be classified (Ladefoged and Johnson, 2014; Laver, 1994; International Phonetic Association, 1999):

- *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. 6): 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 centre 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* (Ladefoged and Johnson, 2014; Laver, 1994; International Phonetic Association, 1999):

  - *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 pressure forces 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 behaviour 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*...

2.3 Phonetics of the French language

Articulation base of the French language, or the common settings such as position, tension, directedness of a speaker's organs of articulation when the speaker is ready to speak, is characterised by:

- High muscle activity (the articulatory organs are rarely lax, especially the lips; it is the lips' pronounced activity what gives French consonants a trace of /ə/ before a pause);
- High vigour and accuracy in sound formation;
- High degree of connectedness in the speech chain;
- Front lingual consonants tend to be produced against the teeth and front alveolar zone with the tongue blade rather than with the tongue tip; the tongue body accordingly favours an advanced, raised, and convex to the roof of the mouth position. O'Connor (1973) indicates that this increases the collision impact of the tongue and a hard articulator when pronouncing a front lingual consonant: again, the language is disposed to more tense tongue configurations.

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;
- There are oral and nasal vowels;
- There are short (regular) and long vowels, yet vowel quantity is not a contrastive feature;
- There are no diphthongs among stressed vowels. The vowels are articulated very clearly;
- Non-stressed vowels retain their properties fully and are not reduced;
- 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;
- 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 French speech connected;

- High degree of connectedness in speech also implies the high degree of presence of the vowel configuration in the preceding consonants. As much as physically possible, the articulatory organs will first acquire the position necessary to produce the vowel—for example, the lips will protrude to produce /u/—and only then will the constriction necessary to produce the consonant occur: /pu/, /tu/, /lu/. In contrast, /pi/, /ti/, /li/ will be pronounced with non-protruded lips, stretched out like in a smile;
- Certain patterns in accentuation and rhythmics.

2.3.1 Phoneme Inventory of the French Language

There are fifteen vowels in French (Lonchamp, 1984; Calliope, 1989b), eleven oral and four nasal ones:

1. /**i**/: "qu**i**", *Fr.* "who", /k**i**/—is a phoneme because it can be contrasted, for example, with "qu**e**", *Fr.* "that", /k**œ**/: close front unrounded vowel;
2. /**e**/[2]: "d**é**", *Fr.* "dice", /d**e**/—contrasting with "d**ais**", *Fr.* "canopy", /d**ɛ**/: close-mid front unrounded vowel;
3. /**ɛ**/[3]: "f**ai**t", *Fr.* "done", /f**ɛ**/—contrasting with "f**é**e", *Fr.* "fairy", /f**e**/: open-mid front unrounded vowel;
4. /**a**/: "t**a**", *Fr.* "your_*S_FEM*", /t**a**/—contrasting with "t**e**s", *Fr.* "your_*PL*", /t**e**/: open central unrounded vowel;
5. /**y**/: "t**u**", *Fr.* "you", /t**y**/—contrasting with "t**ou**s", *Fr.* "all", /t**u**/: close front rounded vowel;
6. /**ø**/: "p**eu**t", *Fr.* "can_*3P_SING*", /p**ø**/—contrasting with "p**u**", *Fr.* "could", /p**y**/: close-mid front rounded vowel;
7. /**œ**/: "s**œu**r", *Fr.* "sister", /s**œːʁ**/—contrasting with "s**û**r", *Fr.* "sure", /s**yːʁ**/: open-mid front rounded vowel;
8. /**u**/: "c**ou**rt", *Fr.* "short", /k**uːʁ**/—contrasting with "c**œu**r", *Fr.* "heart", /k**œːʁ**/: close back rounded vowel;
9. /**o**/: "p**ô**le", *Fr.* "pole", /p**oː**l/—contrasting with "P**au**l", *Fr.* name "Paul", /p**ɔ**l/: close-mid back rounded vowel;
10. /**ɔ**/: "p**o**mme", *Fr.* "apple", /p**ɔ**m/—contrasting with "p**au**me", *Fr.* "palm", /p**oː**m/: open-mid back rounded vowel;

[2] /e/ is often replaced by /ɛ/ in fluent speech.

[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/.

11. /**ɑ**/[4]: "p**â**te", *Fr.* "pasta", /p**ɑ**t/—contrasting with "p**a**tte", *Fr.* "paw", /p**a**t/: open back unrounded vowel;

12. /**ɑ̃**/: "**em**mener", *Fr.* "bring", /**ɑ̃**məne/—different from the word with an absent /ɑ̃/: "mener", *Fr.* "think", /məne/: nasal open back unrounded vowel;

13. /**ɔ̃**/: "m**on**t", *Fr.* "mount", /m**ɔ̃**/—contrasting with "m**o**t", *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**/: "f**in**", *Fr.* "end", /f**ɛ̃**/—contrasting with "f**i**ne", *Fr.* "fine", /f**i**n/: 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 (Anderson, 1982; Tranel, 1987), and its omission in words is elision.

To summarise the footnotes, there is a tendency towards neutralisation 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 in an unstressed position.

There are twenty consonants in the French language (Lonchamp, 1984; Calliope, 1989b), all of which are pulmonic egressive sounds. Twelve of them are obstruents (six plosives, six fricatives, all categorised into pairs of a voiced and an unvoiced constituent), and eight of them are sonorants (three nasals, two liquids, and three semivowels):

1. /**p**/: "ca**p**e", *Fr.* "cape", /ka**p**/: oral voiceless bilabial stop;

2. /**b**/: "cra**b**e", *Fr.* "crab", /kʁa**b**/ or /kʁɑ**b**/: oral voiced bilabial stop;

3. /**f**/: "con**f**irmer", *Fr.* "confirm", /kɔ̃**f**iʁme/: oral voiceless labiodental fricative;

4. /**v**/: "ca**v**e", *Fr.* "cellar", /ka**ːv**/: oral voiced labiodental fricative;

5. /**t**/: "**t**issu", *Fr.* "fabric", /**t**isy/: central oral voiceless laminal denti-alveolar[7] stop;

[4] /ɑ/ is often replaced by /a/, though /ɑ/ is preferred, for example, before /z/ or after /ʁw/ (Tranel, 1987).

[5] /œ̃/ is often replaced by /ɛ̃/ in fluent speech.

[6] Though it should be noted that it does play a role in listening comprehension (Fougeron and Steriade, 1997).

[7] So, realised as [t̪].

6. /**d**/: "ca**d**eau", *Fr.* "present", /ka**d**o/: central oral voiced laminal denti-alveolar[8] stop;

7. /**s**/: "**s**ymbole", *Fr.* "symbol", /**s**ɛ̃bɔl/: central oral voiceless laminal alveolar dentalised[9] sibilant fricative;

8. /**z**/: "ga**z**", *Fr.* "gas", /ga**ːz**/ or /gɑ**ːz**/: central oral voiced laminal alveolar dentalised[10] sibilant fricative;

9. /**ʃ**/: "**ch**ômage", *Fr.* "unemployment", /**ʃ**omaːʒ/: central oral voiceless palato-alveolar labialised[11] sibilant fricative;

10. /**ʒ**/: "â**g**e", *Fr.* "age", /ɑ**ːʒ**/: central oral voiced palato-alveolar labialised[12] sibilant fricative;

11. /**k**/: "o**cc**uper", *Fr.* "occupy", /ɔ**k**ype/: central oral voiceless velar stop;

12. /**g**/: "**g**lobal", *Fr.* "global", /**g**lɔbal/: central oral voiced velar stop;

13. /**m**/: "**m**unir", *Fr.* "provide", /**m**yniːʁ/: bilabial voiced nasal;

14. /**n**/: "**n**asal", *Fr.* "nasal", /**n**azal/: laminal denti-alveolar[13] voiced nasal;

15. /**l**/: "a**ll**umer", *Fr.* "light", /a**l**yme/: lateral oral apical alveolar[14] lateral voiced liquid approximant;

16. /**ʁ**/: "**r**ien", *Fr.* "nothing", /**ʁ**jɛ̃/: central oral voiced uvular liquid fricative[15];

17. /**ɲ**/[16]: "Bourgui**gn**on", *Fr.* "Burgundian", /buʁgi**ɲ**ɔ̃/: palatal voiced nasal;

18. /**w**/: "**ou**i", *Fr.* "yes", /**w**i/: the central oral labio-velar voiced approximant—semivowel glide, corresponding to the close vowel /u/;

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

20. /**j**/: "**y**eux", *Fr.* "eyes", /**j**ø/: 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 "parki**ng**" and can be replaced by /ŋg/ or /ɲ/ (Grevisse et al., 2011).

The acoustic properties of French can be found in the work by Lonchamp (1984).

[8] So, realised as [d̪].
[9] So, realised as [s̪].
[10] So, realised as [z̪].
[11] So, realised as [ʃʷ].
[12] So, realised as [ʒʷ].
[13] So, realised as[n̪].
[14] Realised 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/ (Gess et al., 2012).

2.3.2 Phoneme Duration in the French Language

The length of a phoneme depends on its neighbouring phonemes, on the position within the utterance, and on the syntactic and semantic structure of the utterance (Calliope, 1989a).

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 loses its third phase of production (the burst), and the second one loses its first phase of production (the catch). There is no pause between the consonants, but they belong to two syllables. The consonant from the syllable code'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.

Long *vowels* are related not only to their phonemic characteristics, but also to the stress: stressed vowels always last longer than the unstressed ones. Vowels can be short, semi-long, and long.

There are historically long vowels in French: /ɑ̃/, /ɔ̃/, /œ̃/, /ɛ̃/, /ɑ/, /ɔ/, and /ø/. When stressed, they are always longer.

Prolongation can also be driven by the context: any stressed vowel becomes long before any of the final consonants /ʁ/, /v/, /z/ or /ʒ/ or before the final combination /vʁ/.

There is evidence that duration is not a reliable cue in French vowel contrasts, 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 (Gottfried and Beddor, 1988).

2.3.3 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 (Schwartz, 1921), 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 (Grammont, 1950). 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 organisation, 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 favour 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 Fougeron and Delais-Roussarie (2004) 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 organise 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 organise the utterance into segmentable units and to highlight its logical centre. 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 (Fouché, 1959; Léon and Léon, 1970).

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 organisation 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, itemising, 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.

Jun and Fougeron (2000) modelled prosodic features of the French language in utterances of various communicative types based on four speakers. It is also argued by Fougeron (2001) 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.

2.4 Speech Synthesis

2.4.1 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 (Hardcastle, 1999).

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 (Brenner and Hall, 2007).

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 ionising

radiation, but nevertheless long-term biological and clinical safety for the subject's health remains to be proven (Knuuti et al., 2013).

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 (Gibbon and Nicolaidis, 1999).

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 behaviour 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.

2.4.2 Vocal Tract Modelling and Synthesising Speech

Present-day text-to-speech synthesis systems commonly include two components: at the front end, the system analyses the input text and converts it to a linguistic specification, and at the back end, the speech waveform is generated.

The speech synthesis research has come a long way from knowledge- and rule-based techniques that exploited phoneme-specific acoustic parameters to concatenative speech synthesis (the 1980s and 1990s). The synthesised speech was a concatenation of processed acoustic units taken from a vast pre-precorded speech database (Hunt and Black, 1996); within this approach, which is called "unit selection", it was possible to take into account not merely the acoustics, but also linguistic and pragmatic information such as stress, pitch, part-of-speech information, and position in the utterance.

Facing the impossibility to construct a database capturing enough variation in speaking styles, the researchers proceeded to statistical parametric speech synthesis, where acoustic parameters (a spectral envelope, information about the source, usually the fundamental frequency, and noise-like components for obstruents) are modelled as time series, stochastically generated as such from the linguistic specification (the sequence of phonemes to be produced, annotated with contextual information), and then used in a simple 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. The kind that is explored most is called HMM-based speech synthesis (Tokuda et al., 2013); there, the generative model for acoustic parameters uses hidden Markov models (HMMs), trained by force alignment from an annotated speech corpus.

The 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 centralise 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.

The disadvantage of these cutting-edge approaches is that they present a very technical solution that is in no way related to how human speech is actually produced.

The alternative approach, on the contrary, tries to stick to the natural speech mechanism as close as possible. The biomechanical approach solves the Navier-Stokes equations to estimate deformation of all 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 (Lloyd et al., 2012; Anderson et al., 2015).

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 synthesise speech of ample quality.

From the physical point of view, production of speech sounds is the result of the air flowing from the lungs, getting through an oscillator on its way and undergoing acoustic perturbations that are specific for each type of sound (see Sections 2.1–2.2). Modelling these processes is complex: to handle the aerodynamic and aeroacoustic phenomena, one has to build a joint model for the mechanic, fluid mechanic, and acoustic control of the system.

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 (Maeda, 1990a); it can take days to compute the flow dynamics at the vocal folds (Jansson et al., 2013). So, another approach is to develop models only for the most essential physical phenomena.

Pelorson et al. (1996) and Erath et al. (2011) provide theoretical foundations that allow us to define what phenomena are most important in phonation; they are supported by experiments of Ruty et al. (2007) and Scherer et al. (2001). 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 (Howe and McGowan, 2005; Krane, 2005).

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 (Fant, 1971b). Story (2013) generates area functions as a transition from one vocalic area function to another, with consonants superimposed as constric-

tions attained during this vowel-to-vowel transition. This approach, roughly based on Öhman's (1966) vowel substrates and consonantal perturbation, can be used to synthesise some fixed phrases but is difficult to generalise for use in text-to-speech (TTS) systems.

Another approach would be to construct the sagittal section of the vocal tract (the two-dimensional case, such as in the work by Mermelstein, 1973) or the vocal tract proper (the three-dimensional case, such as in the work by Birkholz and Jackel, 2003) 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 (Fant, 1960) to as sophisticated as hundreds-parameter ones (Gérard et al., 2006). One of the most famous models is the one by Maeda (1990b), 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. Maeda 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 Fig. 8).

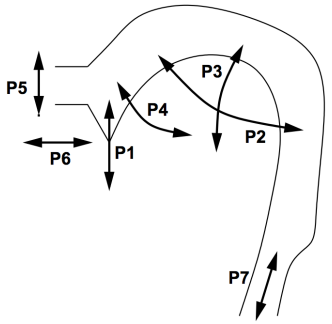


Figure 8: The seven parameters of Maeda's articulatory model: *P7* is the *larynx* height; *P2*, *P3*, and *P4* encode the *tongue*—its body, shape, and tip respectively; *P1* stores the *jaw* opening; *P6* and *P5* define the position of the *lips*—their protrusion and opening respectively.

First applied to vowels only, this model was followed by models to account for consonants too—in two dimensions (Laprie and Busset, 2011a; Laprie et al., 2014) as well as in three dimensions (Badin et al., 2002). (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 better 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. Serrurier and Badin (2008) 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. Laprie et al. (2015) 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 (Maeda and Laprie, 2013) 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:

$$A = \alpha d^{\beta} \tag{3}$$

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 Saltzman and Munhall (1989) (accompanied by the work on how to operate it: Nam et al., 2012), the gesture-based dynamic model by Browman and Goldstein (1992) and gestural dominance model by Birkholz et al. (2006), and the targets-based model by Birkholz (2007).

2.4.3 Coarticulation and Assimilation

One could expect the flow of speech to be a concatenation of isolated phonemes that occur one by one. In fact, this representation does not work because of a phenomenon called *coarticulation*.

Speech is a continuous, connected stream of sounds, a result of joint work of multiple organs with different tissue properties and variating speed of receiving and taking commands because of different nerve fibre lengths to carry these commands. Synchronisation of breathing with the vocal folds and the articulators is not as easy as it may seem, and the human ability for control over timing of articulatory movements is quite remarkable. Consonants can be produced very fast, and they require very accurate aiming; consider the French /d/, which is a voiced dental stop: to utter it, the speaker has to synchronise her tongue tip, vocal folds in the larynx, and possibly jaw movement in some cases, and it is no easy task because it involves activation of muscles that generally react at different speeds. Moreover, not only do some signals reach the articulators slower or faster, there is also a problem of varying weights and degrees of mobility in different articulators. While the whole tongue mass is rather large, the tip of the tongue is capable of darting to a rigid articulator thanks to its light weight. Vocal folds and lips are very mobile too. On the other hand, the tongue body, as already mentioned, and the velum cannot be moved as quickly; and when they are in motion, it is harder for a speaker to end this motion abruptly because of the inertia. This explains why the tongue body makes especially economical movements (e.g. see Fig. 9).

So, as she produces the segments in speech that are presumably stored in an abstract form as separate and independent from each other, the speaker has to factor in a lot of variables from functioning of the brain and the central nervous system to the reaction rate of different muscles moving the articulators along with the articulators' weight that needs to be moved. Moreover, unlike keys in the keyboard, there are no separate vocal tracts to produce each sound and there has to be a smooth transition between them; if articulation were too immaculate, the speech rate would drop dramatically to the point when there actually will be no speech any longer.

In the face of the strict requirement for efficiency of articulatory movements and as long as other people still understand the speaker, since it turns out that there is no semantic difference between, say, /s/ with protruded lips and /s/ with lips in a form of a smile and listeners perceive it as the same /s/, the speaker can save time and position her lips for the vowel in advance, before this vowel even occurs. Then, after /s/, the transition to the vocal tract configuration that will produce the rounded or unrounded vowel will be momentary, because all articulators are almost where they are needed. This is how speech segments occurring near each other become influenced by and more like each other (Kühnert and Nolan, 1999). This notion is called *coarticulation*—though it should be remarked that the aforementioned physical effects make a great contribution to the observed coarticulatory effects in connected speech, but are not enough to explain all of them, primarily because of the discrepancy between speech phenomena in different languages.

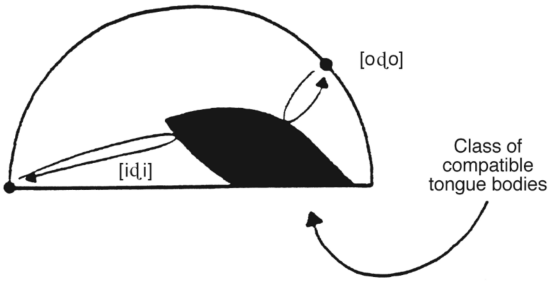


Figure 9: The minimal tongue body displacement observed while producing retroflex alveolars surronded by /i/ and /o/ on both sides. The black area depicts all the tongue body configurations that allow for constriction between the tip of the tongue and the alveolar ridge. The vowel-to-consonant and back consonant-to-vowel transitions are shown by arrows (Lindblom, 1983).

The influence of the preceding speech sounds on the current ones is called *carryover*, *retentive*, *perseverative*, or *left-to-right coarticulation*. The influence of the coming speech sounds on the current ones in called *anticipatory* or *right-to-left coarticulation*.

Another notion that involves sounds becoming phonetically similar to neighbouring sounds too is called *assimilation*. The traditional difference is that coarticulation deals with similitude within the range of allophonic variation—that is, a /t/ is still a /t/ whether it is before /a/ or /before /i/—and assimilation is concerned with changing phonemes altogether. For example, the English ending "*-s*" that marks the plural number in nouns can be pronounced as /s/ or /z/, depending on the preceding sound: if it is voiced, the voiced variant /z/ is chosen, and if it is voiceless, "*-s*" is read as /s/, cf. "cats" /kæ**ts**/ and "dogs" /do**gz**/: /g/ is voiced, and /t/ is not.

The assimilation analogues of left-to-right and right-to-left coarticulation are *progressive* and *regressive* assimilation respectively. Assimilation is usually classified into assimilation of voice like in the "dogs" vs. "cats" example above, assimilation of place—when it is the place of articulation which gets shifted towards the one of a coming speech sound,— and assimilation of manner—applied to the manner of articulation.

Both in coarticulation and assimilation, the influence of the preceding sound is less apparent and also less important: the impact of the past phonemes is attributed

to passive inertia (Kühnert and Nolan, 1999), and since attention both of the speaker and of the listener is directed to the future rather than to the past, it is quite natural that these inertia effects do not tend to expand into progressive assimilation. Progressive assimilation is very rare in the French language; its usual example is the word "subsister", *Fr.* "continue to exist", /sybziste/, where the vocal folds start vibrating at the phoneme /b/ and do not stop yet at /s/, transforming it into /z/. On the other hand, anticipatory coarticulation and regressive assimilation are much more common.

In the French language, there are two types of regressive assimilation: in consonants and in vowels. Consonants with no vowel between them in the stream of speech become assimilated by the feature of voice but do not change their tensity: when assimilated, a voiced consonant does not entirely become its voiceless counterpart and vice versa. Unstressed vowels get partially assimilated by the degree of openness, but it is considered to be a feature of solely informal style.

Because of their different role and extent, the usual approach in text-to-speech (TTS) synthesisers is to treat coarticulation and assimilation differently: by the level of transcription, assimilation is already taken into account, but coarticulation is not.

Speech production is often described in terms of *gestures*—complex strategic movements of articulators that realise the articulation. Every gesture has three time phases: the initial one—the approach—when the articulators come from the previous gesture or from the silent position to the vocal tract configuration associated with the current gesture; the hold phase when the target position or its approximation is held; and the final phase, the release, which brings the vocal tract to the state of a new gesture or to the state of rest.

But, again, fluent speech is, in fact, not a sequence of consecutive articulatory gestures. To make the flow of speech continuous and smooth, the gestures always overlap heavily. For this reason it is rather difficult to manually label a dynamic recording of a vocal tract with active gestures—there simply is no unique answer, since we can only see the movement resulting from the coordination of all gestures that are active at the moment (Nam et al., 2012).

Consequently, one of the viable approaches would be to concentrate solely on *articulatory targets*, as they are strongly associated with individual phonemes. These targets are the ones that are attempted to be attained during the hold phase of a gesture. A target can be *actual* or *virtual* depending on whether it can be reached or not. Even if it is reachable, it still may be not achieved: from the very beginning of research on coarticulation there have been findings, such as the one by Daniloff and Moll (1968), that show this. The progression from one target to another is a compromise between the intent to reach the coming target and the spatial and temporal constraints imposed by the need to produce the neighbouring phonemes as well (Lindblom, 1963b, 1990), and it is possible to estimate which articulators are *critical* for phoneme production and therefore should necessarily reach their target, and which ones are not bound (Singampalli and Jackson, 2007). Vowel-to-vowel transitions are to progress slowly and can be sampled at a lower frequency; consonants, on the other hand, require a high rate of samples per second and occur much faster.

These targets will have to be adapted according to the context to account for

coarticulation (Maeda, 1996), which brings us to *coarticulation models*. The aspects to be modelled are (Farnetani, 1999):

- The temporal domain of coarticulation;
- The outcome of the gestural conflict: what the coarticulated data will look like.

Some works involving coarticulation have already been mentioned above when we discussed articulatory speech synthesisers. The notable examples of coarticulation models include the *look-ahead model* (Henke, 1967; Benguerel and Cowan, 1974), the *articulatory syllable model* (Kozhevnikov and Chistovich, 1965), the *time-locked*, or *coproduction model* (Bell-Berti and Harris, 1982), the *window model* (Keating, 1990). *Articulatory phonology*, proposed by Browman and Goldstein (1992), deals with articulatory gestures as distinctive articulatory units that the speaker makes to produce a constriction in the vocal tract, multiple articulators being involved at every constriction. Thanks to the work by Nam et al. (2012), this approach seems to be a good candidate for further research.

But the most famous and influential coarticulation model is the one by Öhman (1966, 1967). It presents all vowels as linear combinations of the three "extreme" shapes: /i/, /a/, and /u/, and models transitions in vowel-consonant-vowel (V_1CV_2) utterances as transitions between weighted expressions of these extreme shapes superposed with consonantal gestures. Among all the implementations and elaborations we should note a recent work by Birkholz (2013) which uses a specially constructed static MRI database of vowels and consonant-vowel (CV) syllables, aims for high quality both in the articulatory and auditory spaces, and demonstrates the potential for highly intelligible articulatory speech synthesis. The CV syllables are modelled as a smooth movement from the initial vocal tract position, specific for the consonant, to the target shape of the vowel, without any direction change. The target shape of the vowel is assumed independent of the consonant before it, and any variation is attributed to the vowel undershoot (as per Lindblom, 1963a).

3 Using the Model in Speech Synthesis

Since no coarticulation can be modelled without handling *articulation*, the work has developed into building a small-scale coarticulation-aware speech synthesis system that unites existing tools. It is designed to facilitate further refinements and not only to meet the needs of the current implementation.

3.1 Motivational Example

Since the very beginning of the speech synthesis era, it was clear that one cannot do with context-independent data only: we cannot record the realisation of the phonemes /h/, /ə/, /l/, /ʊ/, /w/, and /d/ separately and expect to produce a natural-sounding /hələʊ wəːld/ from it.

For acoustic speech synthesis systems, the easiest and most efficient solution is to record a vast database of speech, expecting it to cover a substantial portion of the language phonological variety. The representation of an utterance there would be a collection of diphones or larger units from the speech corpus: /hə/ (a recording of the transition from the phone realising /h/ to the one realising /ə/), /əl/, /lə/, /əʊ/, and so on. This collection will need adjustments to make it sound as if it were pronounced together (in the raw format the prosody will be inconsistent, and it will sound very unnatural). But in articulatory speech synthesis, the data are usually much sparser both because it is more difficult and expensive to collect them and because their processing is relatively more resource-consuming as well. Due to recent advances in image sequencing speed, growing availability of dynamic magnetic resonance imaging (MRI) machines, and improving computer image analysis techniques, we do not expect it to be a perpetual obstacle; however, data scarcity may become the driving force for creation of efficient algorithms and does not necessarily have to be a disadvantage.

Our dataset construction is inspired by the work of Birkholz (2013) and is following exactly this line of thought: instead of starting from raw speech signals scattered all over the dataset with undershoots among them, it may be beneficial to attempt capturing clean configurations of the vocal tract that represent articulatory phenomena for uttering particular phonemes in a particular context.

Let us take an example how it can work. Let the utterance be "*Il y en a beaucoup*" (Fr. "*There is a lot*", /iljanabokuː/).

If we imagine its component phonemes to be associated with certain articulatory vector targets, these targets will be influenced by context, but they will still retain the status of a target. It means that we could imagine working with the vocal tract positioning associated with /i/ anticipating /ljanabokuː/, /l/ after /i/ and anticipating /janabokuː/, and so on, and proceed by simply navigating between these positions, whether they are achieved or not. Considering the fact that speech makes us articulate well enough to be understood by listeners (i.e. unfailingly come at least close to the intended targets) in an effort-sparing way, if the targets are estimated correctly, the resulting trajectory should not diverge from a real one greatly.

The first point to mention here is that this small difference may be perceptually

important or may be not. The second one is that both kinds of differences are worth studying. Coarticulatory effects are tightly linked to theories of speech production, the way how our brain processes some unknown abstract notion of an utterance to exercise the needed muscular motion. Even when it is difficult to see a universal pattern valid for all utterances in all languages and make conclusions about human speech production, any information that is there to be reported is not negligible for this research.

Coming back to the example, if we are operating with articulatory targets and consider what influences they are under, we can factor out the previous phonemes. Most authors take interest in anticipatory, or right-to-left, coarticulation and attribute the effects of retentive, or left-to-right, coarticulation to passive inertia (Kühnert and Nolan, 1999). So this leaves us with /i/ anticipating /ljanaboku:/, /l/ anticipating /janaboku:/, /j/ anticipating /anaboku:/, and so on.

Then we can reduce the size of the stack. It may be a daunting task to give a particular figure for what the limits of the human mind are for taking the context into account, neither in terms of time of speech production nor in terms of the number of phonemes or wider speech segments. The most studied point is labial coarticulation because lips are very easily accessible, and while numerous studies have reported that lips rounding can last much longer than is strictly needed in an utterance, even spanning over syllable and word boundaries (e.g. in "*structural*", Fr. "*structural*", /stʁyktyʁal/, lips usually round from the closure phase of the first consonant in the sequence, /s/, until /a/), there still are a lot of debates and controversy. The research has positively shown coarticulatory effects that span over several hundreds of milliseconds and up to four–five consonants (Daniloff and Moll, 1968). The notion of limitations in speech has brought about the idea of a certain temporal and spatial window to be modelled, such as in the work by Keating (1990).

It seems reasonable to consider the coming vowel to be the most important factor in anticipatory coarticulation under two conditions: it is not too much ahead in time and there is no conflicting movement planned between the vowel and the current phoneme in production. Then we are dealing with /i/, /l/ anticipating a semi-vowel /j/, /j/ as before /a/, /a/ as followed by another /a/, and so on.

Even though speech is performed in a not entirely conscious way, we can trust a subject to imagine they were about to say a phoneme in a particular context, just on the verge of emitting the sound, and ask to position their articulators correspondingly. Then we could capture this position to subsequently join all the collected configurations and estimate the vocal tract evolution for the whole utterance: /iljanaboku:/. The formed sequence of vocal tract configurations will be able to be used to synthesise speech.

3.2 Data

The data used for the present model was collected in the year 2014 with a magnetic resonance imaging (MRI) machine General Electric Signa 3T. The gradient echo spin technique was used, with the echo time of 1.0840 ms and the repetition time of 3.1200 ms. The flip angle was 10°. The space between slices was 1.016 mm (due to

a not exactly centred position of the subject), and slice thickness was 2 mm. The resulting image was rectangular, 256×256 pixels.

The subject was a male native speaker of the French language. When capturing a vowel, he was instructed to silently position his articulators as if he was in the process of pronouncing it, at the moment when the vowel sounds the most clearly. As for consonants, they were put into consonant-vowel syllables, and the speaker had to make a constriction as if he was about to say a particular syllable just after the capture.

Figure 10: An example of images in the dataset (/ʃi/). One can see it is /ʃ/ or /ʒ/ because the tongue only approaches the palate; the tongue dorsum is raised for /i/.

We have extracted the midsagittal sections from the data, enhanced brightness in the obtained images, and filtered out adjustment captures as well as images containing errors (e.g. instead of the first /ki/ and /kɛ/ the speaker pronounced /ti/ and /tɛ/ by mistake, and there were several syllables for /ʁ/ that were captured more than once or distorted due to movement of the speaker). Due to movement, the image in /li/ displayed two tongue contours, but it was possible to see which contour is correct, so this syllable was not removed.

The resulting dataset contained 94 images like the one in Fig. 10. Tables 1, 2, and 3 list all phonemes that are recognised as belonging to the standard French dialect of the French language in order to indicate both present and omitted items:

Vowels					
Vowel	Example	Image №	Vowel	Example	Image №
i	p**i**s	S7	**y**	p**u**	S15
e	**é**pine	S8	**ø**	d**eu**x	S16
ɛ	p**è**re	S9	**œ**	p**eu**r	S17
a	p**a**s	S10	**ə**	r**e**p**e**ser	N/A[17]
ɑ	p**â**te	N/A[18]	**ɑ̃**	p**an**	S18
o	p**eau**	S13	**ɔ̃**	p**on**t	S19
ɔ	p**o**rt	S12	**ɛ̃**	p**ein**t	S20
u	p**ou**	S14	**œ̃**	br**un**	N/A[19]

Table 1: Vowels of the French language of France, both present in the dataset and not.

Semivowels			
Semivowel	Example	Syllable	Image №
j	**y**olle	ji	S80
w	**ou**ate	wa	S81
ɥ	h**ui**t		N/A[20]

Table 2: Semivowel-vowel syllables of the French language of France, both present in the dataset and not.

[17] /ə/ was not included in the experiment because its realisation can be considered identical to the one of /œ/.

[18] /ɑ/ was excluded due to an error in the image.

[19] /œ̃/ was not included in the experiment because in the French language of France it is often replaced by /ɛ̃/; some speakers even do not distinguish them.

[20] /ɥ/ was not treated.

Consonants															
Consonant	Example	Syllable	Image №	Consonant	Example	Syllable	Image №	Consonant	Example	Syllable	Image №	Consonant	Example	Syllable	Image №
p, **b**	**p**istache,	/pi/	S35	**k**, **g**		/kɑ̃/	S52	**m**	**m**irage	/mi/	S72	**f**, **v**		/fy/	S93
	beau	/pɛ/	S89	(cont.)		/kɛ̃/	S53			/mɛ/	S95	(cont.)		/fø/	S94
		/pa/	S36			/kɔ̃/	S54			/ma/	S73	**s**, **z**	**s**aison,	/si/	S62
		/po/	S90	**l**	**l**entilles	/li/	S23			/mo/	S96		ga**z**elle	/sɛ/	S63
		/pu/	S37			/le/	S82			/mu/	S74			/sa/	S64
		/py/	S38			/lɛ/	S27			/my/	S97			/so/	S65
t, **d**	**t**emps,	/ti/	S39			/la/	S24			/mɔ̃/	S75			/su/	S66
	doux	/tɛ/	S40			/lo/	S83	**n**	**n**uit	/ni/	S76			/sy/	S67
		/ta/	S41			/lu/	S25			/nɛ/	S98			/sø/	S68
		/to/	S42			/ly/	S26			/na/	S77	**ʃ**, **ʒ**	**ch**ou,	/ʃi/	S55
		/tu/	S43	**ʁ**[21]	**r**ouge	/ʁi/	S102			/no/	S99		g**é**nial	/ʃɛ/	S56
		/ty/	S44			/ʁɛ/	S85			/nu/	S78			/ʃa/	S57
k, **g**	**c**ourage,	/ki/	S45			/ʁa/	S103			/ny/	S100			/ʃo/	S58
	gagner	/kɛ/	S46			/ʁo/	S86	**ɲ**	indi**gn**é		N/A[22]			/ʃu/	S59
		/ka/	S47			/ʁu/	S104	**f**, **v**	**f**ête,	/fi/	S69			/ʃy/	S60
		/ko/	S58			/ʁy/	S33		**v**eau	/fɛ/	S91			/ʃø/	S61
		/ku/	S49			/ʁɑ̃/	S87			/fa/	S70				
		/ky/	S50			/ʁɛ̃/	S34			/fo/	S92				
		/kø/	S51			/ʁɔ̃/	S88			/fu/	S71				

Table 3: Consonant-vowel (C-V) syllables for consonants of the French language, both present and missing in the dataset.

[21]Instead of the standard [ʁ], some regional dialects of hexagonal French feature [χ] or [ɾ].

[22]/ɲ/ was not treated.

We have manually annotated the images in the dataset to determine the shapes of the following articulators:

- Larynx;
- Epiglottis;
- PharynxWall (the back vocal tract wall from laryngopharynx to oropharynx) and VelumWall (from oropharynx to nasopharynx);
- Tongue;
- Velum (the soft palate and the uvula);
- Palate (hard palate);

As well as shapes that are necessary for orientation:

- Headregistration (outline of the nose, from the nose bridge to the tip, and markers for the back of the head);
- Mandibula (used for determining the jaw position and sublingual cavity; the visible part in the MRI captures is its marrow).

An example of the resulting contours can be seen in Fig. 11.

The hard palate contour has been extended to estimate the shape of the front teeth which cannot be seen in the conventional MRI because of the high mineral content in teeth (50% of a tooth's dentin volume and 90% of its enamel is minerals, and the rest is water and proteins: Pasteris et al., 2008).

3.2.1 Applying the Articulatory Vector Model

The resulting set is the main one that we have used in the synthesis. We apply a principal-component-analysis-based (PCA-based) articulatory model (Laprie and Busset, 2011b; Laprie et al., 2014, 2015) on the data to obtain articulatory vectors of a variable length representing the vocal tract configurations: the model encodes the jaw position in 3 parameters, the tongue in 12, the lips in 2, the epiglottis in 2, the larynx in 2, the velum in 5—in total 26 articulatory parameters. The number of parameters describing the velum shape can be reduced to 2 in the subsequent implementations. The nasal cavity is not handled by the model.

Since the model is numeric and does not have any information on what areas of an image are of high priority because of their dominating acoustic impact and therefore what errors in modelling should necessarily be corrected, the contours estimated by the model had to be readjusted. The tongue contour has been corrected in 13 images, the lips in 30, the epiglottis in 71, the larynx in 59, the velum in 19. See Table 4 for further details.

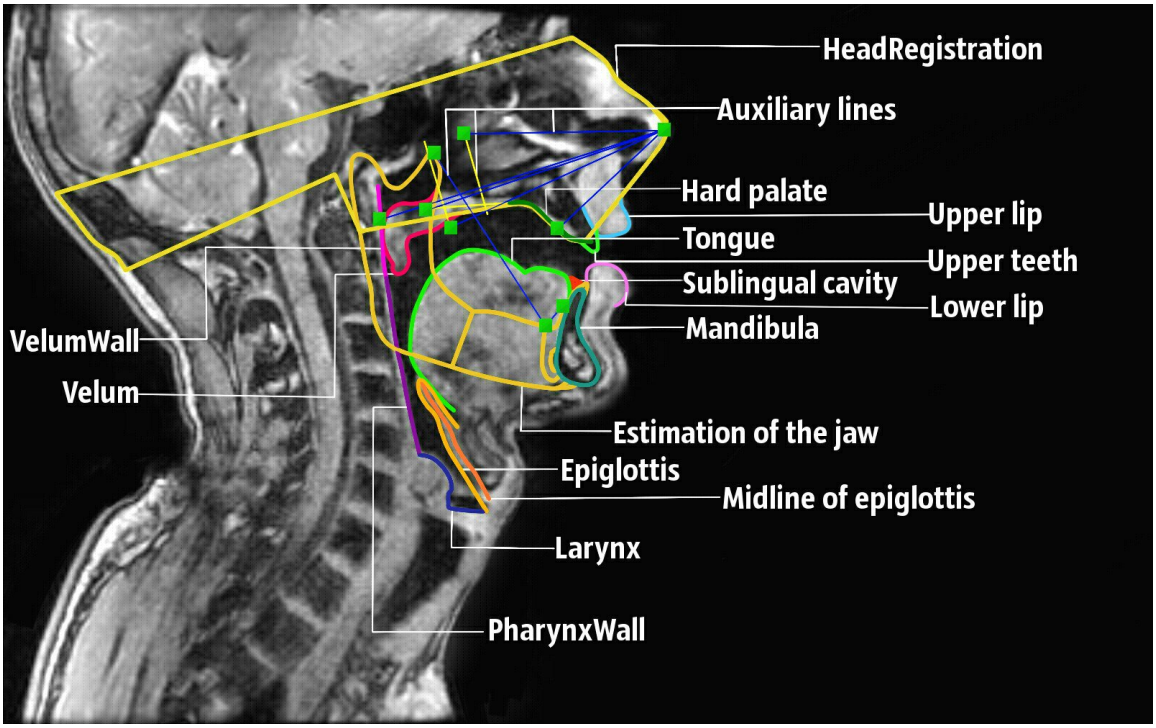


Figure 11: An example of the processed data. This configuration corresponds to /a/: it can be seen that it is an open oral vowel because there is no constriction and the velum is up; moreover, the lips are not rounded (not protruded) and the tongue is in a central position.

Organ	Ex.	Case
Tongue	/ɔ/	The back of the tongue is estimated to be too close to the pharynx wall—closer than in the image. While geometrically insignificant, since this error is at the point of the main constriction, it will have an acoustic impact.
	/y/	The tongue shape does not replicate closely the narrowing between the tongue and the palate because the displayed tongue shape considerably differs from the other ones in the dataset.
	/ɔ̃/	If the airway is blocked, the accuracy in matching the following articulators is not important anymore, whereas before the blockage it is crucial.
	/sø/	The given tongue shape is apparently impossible to replicate accurately neither manually nor by the model, but the correction should yield better acoustic results because we can carefully choose the place of the erroneous delineation.
Lips	/ø/	One lip did not protrude as much as the other one. Because of that and since the model's two parameters are the protrusion and opening rather than the indicators of the upper and lower lips respectively, the modelled protrusion is less pronounced too.

Table 4: *Continued on next page*

Organ	Ex.	Case
	/fa/	There was a dilemma in multiple images: either put the narrowing between the lips at the correct place at the expense of shortening the overall vocal tract length, or make the closure occur further than in the real vocal tract, matching the length to the one in the image. In such cases, no action was taken.
Epiglottis	/ʁɛ̃/	The model mismatched the epiglottis shape in many images for no apparent reason. Sometimes the likely problem was about the compromise between the length of the epiglottis and its angle. This behaviour has been filed as a possible bug.
Larynx	/kɑ̃/	The model failed to replicate the larynx shape in many images for no apparent reason. One could also investigate if there is any inconsistency in marking the larynx contours throughout the dataset. This behaviour has been filed as a possible bug.
Velum	/ʁɛ̃/	Velum is a critical articulator for the phoneme, and having no information that the velum's contact area with the tongue is the most important, the model treated it wrongly to be able to provide a close match in other areas of the articulator.
	/pi/	The velum curvature is unusual in comparison with the one of other captures, and the model tried to replicate it in great detail and failed. We can simplify the velum shape, making it occupy the same area, but in a less curved position.
	/i/	As the dataset was collected in supine position and the subject had to maintain the articulation position for approximately 15 s with no phonation, the velum was often more open that it should be. But the policy was to not correct the data: only data modelling. So, no action was taken in such cases.

Table 4: Manual corrections in the vocal tract configurations that were estimated by the articulatory model.

We have also started investigating modelling of the process of articulators' collision. When the articulators are not in contact, their smooth transitions are much easier to model than when their shapes become drastically deformed. Without the need to depict deformations of the tongue when it hits the hard palate or teeth, the model could require much fewer parameters to encode the tongue configurations. One could try to solve this problem by introducing virtual targets, imagining that collisions occur when the speaker tries to reach an impossible position with their articulator.

Then the resulting trajectory will become much easier to predict, and we would only have to model a reasonable cutting algorithm to remove the imaginary part—or develop a more sophisticated collision algorithm that would be non-specific to any articulator. So we have prepared a supplementary set of contours where, if the tongue hits the palate and gets deformed, its tip is extended further than physically possible, restoring the natural tongue form (see Fig. 12). The same applies for other

soft articulators: the velum can cross the tongue; and as for the lips, considering they are modelled essentially from two points in space whose position is interlinked, the approach of virtual targets could considerably improve modelling the area of their contact, which is an essential piece of information for sound production. To be more specific, within the current implementation, the contact between the lips as in Fig. 13a and in Fig. 13b will lead to configurations that will be modelled as identical; Fig. 13c shows how virtual targets can come in handy in describing tightly drawn lips.

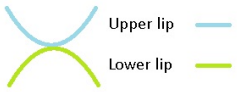


Figure 13a: A schematic mid-sagittal section of the upper and lower lips when in contact.

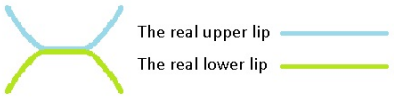


Figure 13b: A schematic mid-sagittal section of the upper and lower lips when they are tightly drawn.

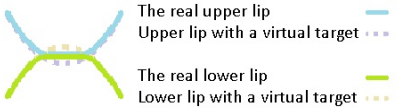


Figure 13c: A schematic mid-sagittal section of the upper and lower lips when they are tightly drawn: restoring the unattainable natural lips' shape with two virtual targets.

But this set of contours with virtual targets was not used in the present work. However, after the preparations, it will be possible to revisit this approach and see which results it yields.

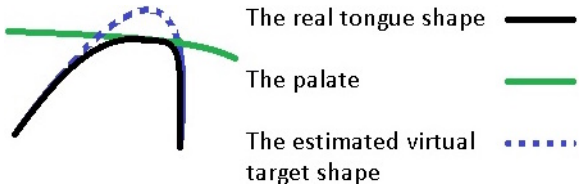


Figure 12: A schematic mid-sagittal section of the tongue and the palate when in collision: restoring the unattainable natural tongue shape with a virtual target.

Instead, after the captured vocal tract positions had been translated into articulatory vectors of 26 parameters each, these vectors were organised in a set of text files with the extension **.dat*. Any line can have a comment at its end, preceded by a combination of two slash signs: "//". Otherwise, the syntax of each file is as follows, line by line:

[1] Name of the phoneme (e.g. *ch* for /ʃ/)

Can be a single-character phoneme name or a multi-character one. The allowed character class: [A-Za-z~] (i.e. Latin letters and the tilde sign).

[2] The phonetic description: class, mode of articulation, place of articulation, and voicing, separated by the plus sign

(e.g. *Consonant + Fricative + Dental + Voiced* or *Vowel + Open + Back + Unrounded + Nasalized*—vowels do not need markers for "Voiced" because all of them are voiced).

Supported input in this line:

- *Vowel*, *Consonant*, *Silence*[23]

  NB: The class of semivowels has to be entered as "*Consonants*".

- *Voiced*, *Voiceless*;

- In consonants: *Stop*, *Flap*, *Nasal*, *Affricate*, *Fricative*, *Sibilant*, *Liquid*, *Semivowel*, *Approximant*, *Trill*;

- In vowels: *Rounded*, *Unrounded*; *Open*, *Open-mid*, *Close-mid*, *Close*;

- In consonants: *Bilabial*, *Labialized*, *Labio-dental*, *Labio-velar*, *Dental*, *Palato-alveolar*, *Palatal*, *Velar*, *Uvular*;

- In vowels: *Front*, *Central*, *Back*.

[3] Critical articulators, separated by the plus sign.

The articulator(s) that is (are) critical for production of the phoneme. For example, /k/ requires the tongue dorsum and the velum.

Allowed articulator names: *Jaw* or simply *J*, *Tongue* or *T*, *Lips* or *Li* along with their separate components *LipsAperture* or *LiA* and *LipsProtrusion* or *LiPr*, *Epiglottis* or *E*, *Larynx* or *La*, *Velum* or *Vel*.

[4] Average duration of the phoneme, in milliseconds (Calliope, 1989a). E.g. "*65 ms*". This duration will be used if the phoneme in utterance is short. If the line is empty, we assign the duration of a vowel to be 90 ms, and for a consonant 50 ms.

[5] A comment line for traceability in the dataset, describing the number of parameters per articulator:

// #jaw = 3, #tg = 12, #lip = 2, #epi = 2, #lar = 2, #velum = 5.

In the following part of the document the contents of the file depend on the phoneme class.

[6] If it is a consonant, the line will contain the names of the consonant and of the anticipated vowel, separated by the plus sign, a colon, the articulatory parameters encoding the vocal tract configuration captured for this syllable, and, finally, a comment for where this articulatory vector comes from, e.g.

[23]The special "Silence" class serves as a vocal tract configuration for the beginnings and ends of utterances, to avoid starting from an open mouth or ending before its closure. During silence in the middle of an utterance the articulators are normally mobile and therefore, in that case, silence should be treated along with the utterance phonemes; this special initial and final "silence" was not included in the dataset and was estimated from /ma/ because it is a very natural mouth position—to the extent that many languages have been constantly creating words very close to "mama" to mean "mother" (Jakobson, 1962).

z + epsilon : 1.04 0.47 0.15 0.72 -2.59 4.53 -5.60 -2.72 -1.92 1.07 -1.94 -1.55 1.46 0.15 -0.33 0.00 -0.16 -4.22 -2.76 -3.33 -0.26 4.15 -1.84 -0.76 -1.62 -0.99 // from the voiceless s + epsilon, S63

In case of a consonant, all subsequent lines will be of the same format.

If the phoneme is a vowel, vowelName : articulatory parameters // comment, e.g.

a~ : 0.74 -0.59 -0.14 -19.59 10.10 -7.02 -1.40 0.47 -4.20 0.03 -0.78 -1.17 -0.43 0.24 0.23 0.38 -0.03 3.28 1.11 0.65 2.11 -0.03 12.28 -0.76 1.36 -0.42 // S18

[7] In case of a vowel, this is the last line. It contains two float numbers and a comment which will be explained in Section 3.2.2, e.g.:

-0.287951, 1.24392 // Projections: a~ represented as u + s (i - u) + q (a - u)

3.2.2 Expanding the Dataset

As we showed in Table 3, it was not all possible consonant-vowel combinations that were explored. We follow the same approach as Birkholz (2013) and expand the dataset by estimation of the missing samples from the present ones.

This operation involves the notion of *cardinal*, or *corner*, *vowels*. Within this assumption, /a/, /i/, and /u/ (or, specifically for the French language, /y/) are articulated at the most extreme vowel positions (Fig. 14), and all the others should be spatially located between them (come back to Section 2.2 for more details).

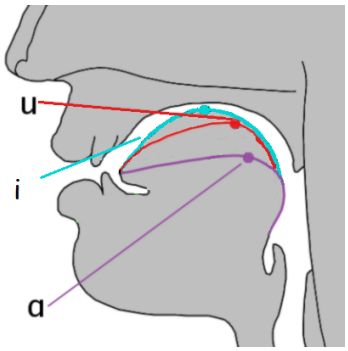


Figure 14: Corner vowels: /a/, /i/, and /u/.

So, considering the fact that the articulatory vector space is 26-dimensional in our case, if the used metric (Eq. 5) is derived from the Euclidean norm (Eq. 4):

$$\|z\| := \sum_{k \in \mathbb{N} \cap [1,26]} z_k^2 \tag{4}$$

$$\mathrm{dist}(x, y) := \|x - y\| = \sum_{k \in \mathbb{N} \cap [1,26]} (x_k - y_k)^2, \tag{5}$$

and if we define a function dist proj:

$$\mathrm{dist\,proj}(x, \vec{a}, \vec{i}, \vec{u}, s, q) := \mathrm{dist}(\vec{a} + s \times (\vec{i} - \vec{a}) + q \times (\vec{u} - \vec{a}), x), \tag{6}$$

then the "projection" of an arbitrary vowel onto the corner vowels is the pair (s_{opt}, q_{opt}) which, as defined by Eq. 7, is exactly the one that was mentioned earlier in Section 3.2.1 when explaining the syntax of the **.dat* files (line [7]):

$$(s_{opt}, q_{opt}) = \underset{(s,q) \in \mathbb{R} \times \mathbb{R}}{\arg\min}\, \mathrm{dist\,proj}(vowel, \vec{a}, \vec{i}, \vec{u}, s, q), \tag{7}$$

The pair defined in the Eq. 7 gives us the closest point in the two-dimensional subspace containing the corner vowel vectors to the arbitrary vowel in question. The distance between this vowel and the closest point to it gives us the value of the error for such a projection—how far away in terms of the Euclidean distance of the parameter space the estimation falls:

$$err = \min_{(s,q)\in\mathbb{R}\times\mathbb{R}} \operatorname{dist}\operatorname{proj}(vowel, \vec{a}, \vec{i}, \vec{u}, s, q) \tag{8}$$

It means that if the corner vowels assumption for the modelled vectors holds and all vowels are found close to the convex hull of the /a/, /i/, and /u/ vectors, then

$$/vowel/ \approx /a/ + s_{opt} \times (/i/ - /a/) + q_{opt} \times (/u/ - /a/), \quad s^2_{\text{opt}} + q^2_{\text{opt}} \leq 1 + \epsilon, \tag{9}$$

where ϵ is a non-negative real number that is negligibly small.

Both this work and its predecessor, the one by Birkholz (2013), suffer from inaccuracies either in vowel shapes or in their rendering, propagated further into the model. Birkholz solved this by introducing modifications into the data so that there was a closer match in the acoustic space, and only after that did he apply the vowel projection method. Nevertheless we concentrated on correcting the modelled data solely from the articulatory point of view, as described in Table 4.

When applied to the missing samples, Eq. 9 means that the vector for a consonant-vowel (CV) syllable configuration that was not treated during the experiment, such as /tɛ̃/, can be estimated from the present entries—/ta/, /ti/m and /tu/ in the case of /tɛ̃/:

$$/CV''/ \approx /Ca/ + s^{(\text{V})}_{\text{opt}} \times (/Ci/ - /Ca/) + q^{(\text{V})}_{\text{opt}} \times (/Cu/ - /Ca/), \tag{10}$$

where $(s^{(\text{V})}_{\text{opt}}, q^{(\text{V})}_{\text{opt}})$ is the pair obtained from applying Eq. 7 to the vowel V in question.

The same was done for anticipation of the semivowels /j/ and /w/, even though their representation in the dataset was within a syllable rather than on their own.

All newly estimated vectors are appended to the **.dat* files in the same format as for the existing entries. The extension for expanded phoneme data is **.artv*.

3.3 Utterance Processing

The ultimate goal is to proceed as follows when synthesising an utterance given its transcription:

- Break the utterance into groups that serve as speech blocks in which coarticulatory effects are confined.
- Break these groups into syllables.
- Determine what phonemes constitute each syllable.

- Go to the level of separate articulators. For each phoneme and each articulator, determine the span of coarticulatory phenomena and what is the most significant anticipatory contribution to each phoneme.
- Form a sequence of articulatory vectors to describe the vocal tract evolution that will yield the required utterance.
- Estimate the evolution of the supraglottal pressure, the fundamental frequency (intonation), and vocal folds operation.
- Translate the articulatory vectors in the estimated sequence into vocal tract configurations.
- Use the collected estimations in an acoustic simulation system and obtain a recording of the synthesised sound.

In the following sections 3.3.1–3.3.6 we will cover each procedure step by step. Refer to Section 3.3.8 for a summary of the program design, and Appendix for further details.

3.3.1 User Input Parsing

To process an utterance, we consider the available phonemes in the dataset and construct a required utterance from them. The user can learn what phonemes are available by calling the following lines of code:

```
consonants, vowels, silence = scan_db()
```

or

```
phonemes = scan_db(sort=False)
```

from *main.py*—depending on whether it is important what phonemes belong to which classes (the first option) or not (the second one). Then the user can use the command

```
vowelnames = [v.name for v in vowels]
```

or similar to retrieve all names of the available phonemes.

We start from parsing the extended transcription entered by the user. The syntax conventions are as follows:

- The user can mark rhythmic groups by the vertical bar sign, surrounded by blank spaces on both sides: " | ". During the synthesis, there will be a pause of 40 milliseconds in these places.
- The user may choose to mark or not to mark the syllable boundary. If they do, the symbol to be used is the hyphen sign: "-". Otherwise the algorithm of syllable segmentation by Bigi et al. (2010) will be applied (see the rules in Table 5).
- Within a syllable, the user can mark the intonation contour with the following symbols:

  - "/" for voice rising;
  - "\" falling[24];
  - "_" for an even intonation contour.

  The intonation instructions should be placed at the beginning of a syllable: "_/ba:".

- The stress sign "ˈ" can be put either in the beginning of an explicitly marked syllable or immediately before the syllable-forming vowel otherwise. The stress sign is expected to be *after* the intonation schema if there is any: "_/ˈdy".

- Any multi-character phoneme name should be enclosed in curly brackets, e.g. "{epsilon}" (for /ɛ/).

- A long phoneme should be marked by the ":" sign immediately after it.

So, if `transcription` is the whole character sequence that makes an utterance (with phonemes, their duration marks if needed, stress signs, syllable and syntagm boundaries where it applies), then the user can use *main.py* to create a new **Utterance** object:

```
myutterance = Utterance(transcription)
```

Examples:

- `Utterance("\\l{epsilon}\\za\\ba/_{zh}ˈu:r | s{o~}t{o~}m\\bˈe")`
  (standing for /lɛzabaˈʒuːʁ sɔ̃tɔ̃mbe/: "Les abat-jours sont tombé", "The lampshades have fallen down")

- `Utterance("ila\\pa\\ma:l")`
  (standing for "Il a pas mal", "He feels no pain")

- `Utterance("ka//mˈy")` (standing for "Camus", as in *Albert Camus*, or "pug-nosed")

- `Utterance("{zh}{oe}k//rwa")` (standing for /ʒøkʁwa/: "Je crois", "I believe")

The user may specify non-default values in this constructor: where to look for phoneme files and what extension they must have.

Block	Syllables
VV	V-V
VCV	V-CV
VCCV	VC-CV
if not	V-CGV
or	V-FLV
or	V-SLV
VCCCV	VC-CCV
if not	V-FLGV
or	V-SLGV
or	VSL-SV
VCCCCV	VC-CCCV
VCCCCCV	VCC-CCCV

Table 5: The syllable segmentation rules (Bigi et al., 2010). *V* stands for "vowels", *C* for "consonants", *G* for "glides", *L* for "liquids", *F* for "fricatives", and *S* for "stops".

[24]Actually, "\\" because "\" is a special character that has to be escaped.

3.3.2 Coarticulation: Going from a Sequence of Phonemes to a Sequence of Articulatory Vectors

Within our assumptions, there are two options for a phoneme:

- To anticipate the coming vowel;
- To merely be linked to the next phoneme without shifting one's articulatory target towards it, as if the current phoneme were on its own.

The expanded dataset provides almost all vowel and semivowel contexts as well as pure vowel configurations. The information that we do not have is:

- The vocal tract configurations for consonants on their own, without the effect of anticipation of any vowel[25];
- Vowels when in anticipation of other phonemes;
- Anticipating multiple phonemes.

The first point occurs very frequently. One case simply is where there is no vowel coming after a particular consonant: it is the end of an utterance. The other case is consonant clusters: there may be a consonant ahead that contradicts the gesture for the vowel after it and, therefore, prevents the speaker from anticipating the vowel until past the consonant.

We needed to estimate these pure consonant configurations from the existing data. The idea was that anticipating /a/ is rather neutral. The articulatory targets for /Ca/, /Ci/, /Cu/, and /C/ before any other vowel are almost the same at the place of constriction, but usually differ in the tongue body position. So, in order to compensate for the pre-vowel tongue body shifting, for an articulatory vector /C/, corresponding to a consonant that does not anticipate any vowel, we introduce the following expression:

$$/C/ \coloneqq \frac{1}{3} \times /Ca/ + \frac{1}{3} \times /Ci/ + \frac{1}{3} \times /Cu/ \tag{11}$$

The other two aspects, the fact that within our dataset vowels cannot anticipate other vowels and the impossibility to anticipate several phonemes, can be neglected for reasons that will be discussed later in this section. The main idea is that the articulators that are critical for production of the current phoneme will have to reach their target position to make the emitted sound be perceived correctly, and it will be other, "free" articulators which will be able to anticipate the coming vowel or several phonemes.

To determine what articulatory vector to choose as the target for a particular phoneme in context—the estimated solitary position or an entry from the dataset—we exploit the following algorithm realised in `Syntagm.__init__(...)`:

[25] If it is a stop consonant at the end of an utterance, for technical reasons, the speaker will add a vocalic sound—which is often /ə/ in French—after it. However, there is no anticipation involved.

- *Vowel*

  If the phoneme is a vowel, within our assumptions it always has the context-free articulatory position.

- *Consonant*

  This phoneme target may shift from its isolated position to the one when anticipating the next syllable-forming vowel, whether from its own syllable or not. There are several points to consider:

  – Time window: the limit for anticipation is set 200 ms. If the vowel is scheduled later than in 200 ms, the system prevents its anticipation.

  – Stack limit: 5 phonemes. If the coming vowel is farther, it is not anticipated yet.

  – Engaged articulators and involved places of articulation: we have organised all places of articulation from the back to the front, counting double articulation by the more front constituent (see Table 6). If the syllable requires movement in conflicting directions, e.g. the vowel that is a candidate for anticipation requires articulators' moving to the front, but before then there is a backward movement to pronounce another consonant, the vowel may be blocked for anticipation until the conflict ahead resolves.

    The algorithm loops over phonemes in the group stored in `Syntagm` in the reverse order, adding either the last encountered vowel name or a special mark "Solo" to the `Syntagm.anticipations` list to mean "anticipate the vowel" and "assume the isolated form" respectively. If the algorithm has reached a consonant that does not anticipate a particular vowel, no consonant that comes later (or, chronologically speaking, before) will be able to anticipate this vowel (i.e. it is impossible for the speaker to start anticipating a vowel and then reconsider).

    In the unreversed order of speaking, if the current consonant is C_{curr}, the anticipated consonant is $C_{anticip}$, the syllable-forming vowel is V_{govern}, and μ is a numeric representation of a phoneme's place of articulation, in our case obtained as per Table 6), we can define the gaps between their places of articulation:

$$\Delta_{C_{anticip}C_{curr}} \coloneqq \mu_{C_{anticip}} - \mu_{C_{curr}} \tag{12}$$

$$\Delta_{V_{govern}C_{curr}} \coloneqq \mu_{V_{govern}} - \mu_{C_{curr}} \tag{13}$$

  Then we will be able to establish the spatial conditions under which the modelled coarticulatory effects can happen between two phonemes: the movement direction (Eq. 12) from the current consonant to the next one should either

Phoneme class	№	Phoneme class	№
Unrounded nasal vowel, central or back Uvular consonant	1	Palato-alveolar consonant	5
		Dental consonant	6
Velar consonant	2	Rounded vowel: open, open-mid, or close-mid Labialized consonant Labio-dental consonant Labio-velar consonant	7
Unrounded oral vowel, central or back	3		
Front vowel Palatal consonant	4		
		Close rounded vowel Bilabial consonant	8

Table 6: Assigning order numbers for the places of articulation present in the dataset.

coincide (Eq. 14) with the direction from the current consonant to the syllable-forming vowel (Eq. 13), or the absolute value of the difference between the gaps should not be greater than 5 (Eq. 15):

$$\Delta_{C_{anticip}C_{curr}} \times \Delta_{V_{govern}C_{curr}} \geq 0 \tag{14}$$

$$\mid \Delta_{V_{govern}C_{curr}} - \Delta_{C_{anticip}C_{curr}} \mid \leq 5 \tag{15}$$

Such a technical solution allows us to engrain that, for example, both in a syllable /bla/ (such as in "blaser", /bla.se/—*Fr.* to sate, to bore) and in a sequence /lba/ (e.g. "albatros", /al.ba.tʁos/—*Fr.* albatross) the articulators perform two almost simultaneous gestures for /b/ and /l/, and both of the consonants anticipate the vowel /a/, while in /lka/ (found in "alcaloïde", /al.ka.lɔ.id/—*Fr.* alcaloid) there is a conflict and /l/ does not seem to anticipate /a/.

– Syllable boundaries: the algorithm allows for coarticulation over syllable boundaries, but the condition for it is stricter. So, for a consonant from the syllable coda to anticipate the next vowel, it must satisfy all the in-syllable conditions in addition to Eq. 16:

$$\mid \Delta_{V_{govern}C_{curr}} - \Delta_{C_{anticip}C_{curr}} \mid \leq 3 \tag{16}$$

It should be noted that this case is less common in French because the French language tends to open syllabication. The general principle is that open vowels occur in closed syllables, and, vice versa, close vowels appear in open syllables (Schwartz, 1921).

Then we added a simulation of the effect of vowel undershoot. The vowel target is scheduled to be reached in 150 ms after the vowel onset. However, if the vowel is shorter, at 20 ms before the vowel end the speaker begins turning to the next target, since there is not enough time to reach the target of the current vowel. We could also imagine another approach: when there is not enough time to reach the vowel target, to change the target at some point, but keep it at the same time point when the vowel target used to be; in such a way the tongue speed will be preserved. Another solution that is often used in research is a second-order filter (Kröger et al., 1995).

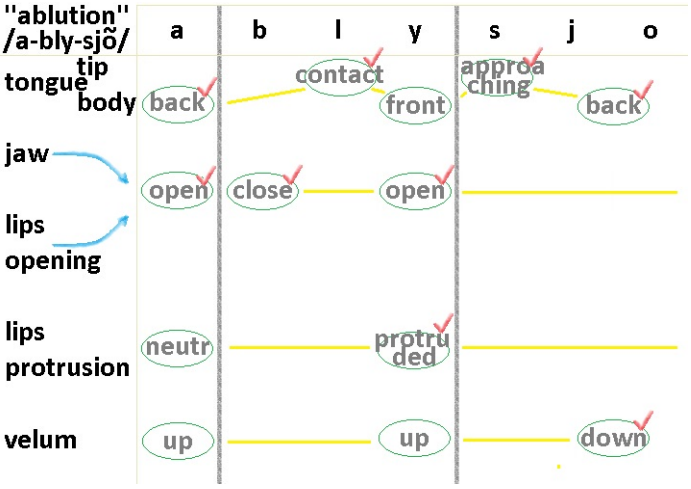


Figure 15: A sequence of critical and optional targets to hit for separate articulators for the word "ablution" (/ablysjɔ̃/). The conditions of articulators are given in the green ovals, and the critical ones are marked by red check marks. The lips opening and the position of the jaw have been united because the studies show that the jaw can barely move in some subjects, especially the ones of an elderly age (Zemlin, 2010).

As for consonants, if the consonant is a stop, a nasal, or a fricative, despite the different nature of these classes, their temporal management is implemented in the same way: if the consonant is not long, its whole duration t_C is divided into consecutive periods of $0.045 \times t_C$, $0.9 \times t_C$, and $0.045 \times t_C$. If, indeed, the consonant is long (because it is geminated, drawled to convey some expressive meaning or for any other reason), its temporal organisation is prolonged at the expense of the middle part—the hold phase. The initial (the catch in case of a stop) and the final ones (the burst for a stop) stay the same: if $coef_L$ is the coefficient to derive the long phoneme duration from the average one (in our implementation $coef_L$ is always 1.7), then the periods are $\frac{0.045}{coef_L} \times t_C$, $\frac{coef_L - 0.09}{coef_L} \times t_C$, and $\frac{0.045}{coef_L} \times t_C$.

Having the utterance represented as a sequence of target articulatory vectors, we can proceed to the ways of joining them.

The simplest way would be to perform linear or cosine interpolation between the target vectors.

We have also prepared the instrumental base for a more refined articulatory operation. It is possible to go to the level of articulatory subspaces that correspond to separate articulatory organs, single out the critical articulators for each phoneme, and in such a way define the points in particular subspaces that necessarily have to be achieved, starting the transition as soon as possible, i.e. as soon as the articulator is not bound. Then the other targets in free subspaces could be optional, being

passed with the least effort possible, which can be modelled by a smoothing spline (see Fig. 15 for a schema).

3.3.3 Area Functions

Having synthesised a sequence of articulatory vectors to represent the utterance and retrieved the encoded sequence of the corresponding vocal tract mid-sagittal sections, we have to give volume to the two-dimensional shapes. Since we use a one-dimensional synthesiser, we translate the shapes into area functions. An area function describes how the cross-sectional area, $A(k,t)$, varies along the vocal tract, from the glottis to the lips, progressing by lengths $x(k,t)$. This way, the vocal tract configuration (such as in Fig. 16) is approximated by the information on a fixed number of cylindrical tubes—in our case 40 (see Fig. 17a for a visualisation of area functions and Fig. 17b for the respective estimated speech signal spectrum).

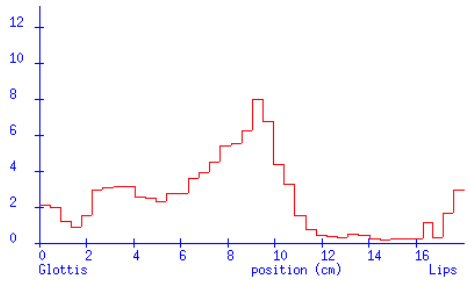


Figure 17a: The area function built for the vocal tract position in Fig. 16 for the vowel /i/.

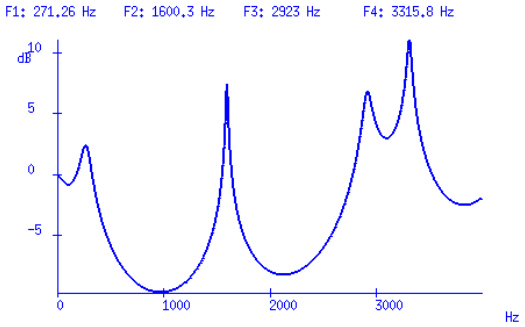


Figure 17b: Speech signal spectrum as estimated by XARTICUL for the area function in Fig. 17a corresponding to the vowel /i/.

The area functions are obtained from XARTICUL[26] by the algorithm of Heinz and Stevens (1965) with coefficients adapted by Shinji Maeda. Yves Laprie has also attempted to improve the results by using the α and β coefficients (see Eq. 3) reported in the works by Soquet et al. (2002) and McGowan et al. (2012), who worked on static MRI too. On our dataset, the coefficients by Soquet et al. (2002) showed worse results than the traditional ones from the work by Heinz and Stevens (1965)—considering a great discrepancy between the coefficients for a female's and a male's vocal tract, it could have been due to the fact that these results were not robust enough. As for the work by McGowan et al. (2012), it exploits a locally linearised solution and expects to compute the α and β coefficients that would be well-suited to the current vocal tract shape; however, the benefit is negligible.

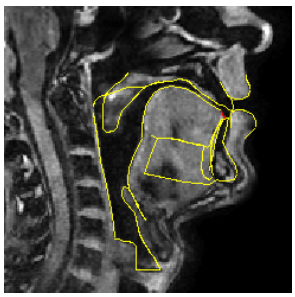


Figure 16: Vocal tract positioning to produce the vowel /i/.

[26] `http://www.loria.fr/~laprie/xarticul.htm`

Since the velum modelling is still in the phase of testing, XARTICUL provided the areas of the vocal tract with the uvula cut away and the area of the uvula separately, in **.xa* and **.xav* files. To proceed with the acoustic propagation, we subtract the corresponding areas and obtain the area with the uvula taken into account and transform the result from the format of XARTICUL to the format of the acoustic simulation system by Elie and Laprie (2015, 2016).

At the moment, the feature of nasality has to be added to the area functions manually.

3.3.4 Modelling Subglottal and Supraglottal Pressure

For modelling the intraoral pressure, we divide the phonemes into two classes: obstruents, which require raising the pressure, and sonorants, which do not.

In stops and fricatives, the pressure rises at the hold phase: when there is a complete blockage of the vocal tract airway in case of stops, accumulating air masses before the point of constriction, and when the articulators approach each other for a fricative consonant, resulting in a turbulent motion.

The pressure keeps being high until the hold phase is over. When the constriction is lifted, the drawn air quickly goes out, and the pressure evens out to the level of the atmospheric pressure.

We did not model the implications of the supraglottal pressure elevation on articulation, voice pitch and intensity, though it is attainable in the future implementations: it is known that articulation influences transglottal pressure differential, and the air displacement necessary for voicing will make the supraglottal volumes grow (Perkell, 1969; Kent and Moll, 1969), which should be taken into account when managing a dataset of MRI captures taken only for voiceless consonants (in the current implementation there is no difference at all between uttering a voiceless consonant or its voiced counterpart); within the program's tool box it is also possible to make a decrease in pitch or intensity as the intraoral pressure rises (Zemlin, 2010), along with the influence of high and low vowels (the ones that require the jaw to be at a relatively high and at a low position during articulation respectively) as reported by Koenig et al. (2011).

The glottal opening—and hence supraglottal pressure—instructions are stored in files with the extension **.ag0*.

As for subglottal pressure, there are relatively few studies to report on its values. It is known that during simple utterances such as pronouncing a neutral vowel with different lung volumes subglottal pressure's behaviour alters greatly from the relaxation curve typical for the pressure in lungs: subglottal pressure appears to be sustained at the same level. So we keep subglottal pressure constant, 1200 Pa, throughout the whole utterance.

3.3.5 Modelling Vocal Folds Oscillations

Generally speaking, if the phoneme is voiced, the vocal folds vibrate throughout all its duration, and if it is voiceless, no phonation occurs: the vocal folds are abducted.

The exception is voiced stops, where the vocal folds activate only in the middle of the hold phase. In our implementation, if t_{catch} is the moment when the hold phase is initiated and t_{burst} is when it is released, then voicing starts at

$$t_{voiceonset} \coloneqq \frac{1}{2} \times (t_{catch} + t_{burst}) \tag{17}$$

As for voiceless consonants, if they are followed by a voiced phoneme, the vocal folds are adducted after the burst—a little earlier than the next phoneme begins. Other cases, such as negative voice onset time for voiced plosives, attested, for example, by Pépiot (2013), were not modelled.

The vocal folds instructions are stored in files with the extension **.agp*.

3.3.6 Modelling Intonation

We have introduced a base pitch of 140 Hz for the synthesised speech since the speaker is male. The user can modify the pitch by encoding it graphically directly inside the transcription with the "/", "\", and "_" symbols to mean raising the voice, lowering it, and keeping it even.

Since modelling intonation was not among the goals of the present work, if the user opts to leave the pitch information out, we assume a very simple linear contour: from 140 Hz to reach 160 Hz at 60% of the utterance duration, and 120 Hz by its end.

The pitch information is encoded in files with the extension **.f0*.

3.3.7 Obtaining Speech Sounds

Finally, we gather all the generated instructions—the area functions, the vocal folds oscillations, the pitch, and subglottal and supraglottal pressure—to use these files in the acoustic simulation system implemented by Elie and Laprie (2015, 2016), which produces sounds in the Waveform Audio File Format (WAV).

3.3.8 Program Design

The program's source code in Python 2.7 is available at `https://github.com/Anastasiia-Tsukanova/Articulatory-Speech-Synthesis`. Refer to the Appendix to learn the specifics of the structure of the program.

Its structure is more complicated that is strictly needed for the implementation for two reasons:

1. Since we plan to continue working on the system, it was decided to support two variants of articulator naming: short names, e.g. "Li", and full names, "Lips". Short names should become handy for quick manual testing of approaches that are candidates for implementation.

2. At this moment, only the lips are a multi-component articulator: one part is their aperture, and the other is their protrusion. But in further implementations, we could imagine singling out the tongue tip and dorsum or any other part

of an articulator. To make subarticulators not hard-wired, the structure had to become more opaque: in some cases, we need all articulators *and* their parts; in other cases, only the higher level; and in other cases, a mixture of higher-level articulators and the more fine-grained parts when they are available.

The program is also accompanied with a translation of a part of XARTICUL's code into Python 2.7 to make visualisations.

3.4 Results

To synthesise speech, we have joined several components:

- A dataset of still MRI captures;
- An articulatory model;
- Knowledge of phonetics allowing us to expand the dataset;
- An algorithm for controlling the articulators, the vocal folds, the voice pitch, subglottal and supraglottal pressure;
- Area functions construction;
- An acoustic simulation system.

The main idea was to bring forward the uttermost benefit of articulatory speech synthesis: clear and full control over the articulators. For that, this work tries to embed coarticulation in the base of the system as deep as possible instead of generating coarticulatory effects while being directed by noisy data.

The dynamic control over articulation in the system is purely rule-based, which can be seen as a disadvantage because no set of rules and exceptions will be enough to cover the vast variety of speech phenomena. On the other hand, it is a very reasonable starting point, indispensable before progressing to more advanced models. So it is very important to learn where the current implementation performs well and what its limitations are. The next section is dedicated exactly to that.

4 Evaluation and Critique

This chapter is dedicated to analysis of the results.

4.1 Dataset

4.1.1 The Nature of the Data

The data collection was conducted with an MRI machine, which produces a relatively high-quality 3D image describing the positions of almost all articulators we may need.

While the expected, most normal speech is uttered in an upright position, our subject was lying down, which changed his articulation and probably affected the obtained results. Vorperian et al. (2015) reports that for vowels the supine position induces a significant increase in pitch and the third formant (F3), and a change in the vowel's configuration in the space defined by corner vowels; though Flory (2015) indicates that some speakers are able to speak in such a way that the difference between positions would be negligible, and even if the difference is obvious in the acoustic space, listeners are not capable of catching that.

The speaker was instructed to attain articulation as if he were on the verge of uttering a given consonant-vowel (CV) syllable. However, since the utterance was not taken from the natural process of fluent speech and the task was so specific, the speaker was likely to overarticulate. Moreover, it is a demanding and sometimes impossible task to sustain articulatory positions for dozens of seconds.

Because of this difficulty, the speaker's velum often went down more than it should, resulting in a vocal tract shape corresponding to nasalisation. Moreover, of course the position of the velum changes along the sequence of captures, and when we use these images in subsequent implementations, these excessive movements of an artificial origin will appear in the synthesis results.

Nevertheless we suppose that the collected articulatory targets are valid for the following reasons:

1. Overarticulation is not a problem if we have a model for the undershoot (Lindblom, 1990).

2. In general, the transitions of the vocal tract are governed by vowels. Consonants are, in a way, superimposed on them (which was actually modelled by Story, 2009), so CV-syllables should be a viable point for starting a dataset (i.e. VC-syllables would have been much less fruitful). Its small size of two phonemes has the benefit of having much fewer variants than, say, VCV-syllables.

3. Speakers of French do not reduce unstressed vowels, which has let us completely discard the aspect of stress when creating the dataset.

During construction of the dataset we selected the sagittal section, obviously disregarding the lateral /l/, which is produced by a constriction along the centre line of the vocal tract, with the air rounding the obstacle at both sides; another class of

phonemes that loses much information when going two-dimensional is sibilants—or, rather, all phonemes that need the tongue to form a groove.

Regrettably, the articulatory model (Laprie and Busset, 2011b; Laprie et al., 2014), applied on these mid-sagittal contours to encode the vocal tract positions in an efficient way, has its own flaws that were propagated further in the system:

- The larynx and epiglottis were often treated wrongly (see Table 4).
- Basically, the model discards the information on the area of contact between the lips, which is a relevant articulatory feature, for example, to produce /u/. Moreover, the lips are so bound that it is impossible to find the coefficients to make the model correctly replicate the labiodental phonemes /f/ and /v/.
- The subject exhibited an especially flexible velum, capable of a magnitude of complex curly shapes. Five parameters that were assigned for the velum model were too few to replicate the shape but enough to let it try. From the acoustical point of view, it could have been better to limit the velum model to two parameters so that it made a crude estimate instead of sophisticated fitting.

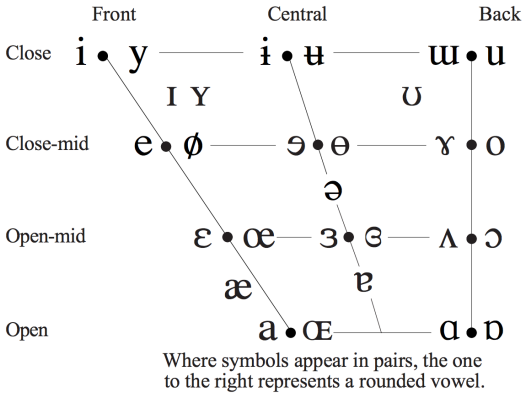
Manual data annotation, which was used to mark the contours of the articulators, is a labour-intensive solution, but with a benefit of a much more accurate result than with any kind of automatic image processing. On the other hand, while automatic tools can be misguided, they are at least always consistent, so if we attribute the problems with capturing pharyngeal articulators (see Section 3.2.1) to some inconsistency in outlining the larynx, automatic tools could be of help.

4.1.2 Dataset Expansion

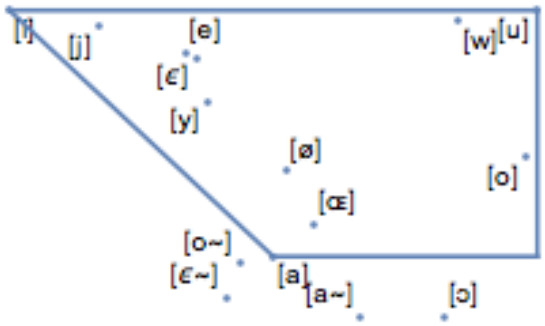
The dataset is expanded on the basis of the cardinal vowels assumption, which is actually formulated for the acoustic space, and not for the $\mathbb{R}^{26}$—the parametric space of the articulatory model. Moreover, if we find the relation between, say, /ɛ/ and the /i/-/a/-/u/ set, it does not mean that the same relation would be valid for /bɛ/ and /bi/-/ba/-/bu/. This is why we needed to test the validity of projections.

The organisation inside the projected vowel space is displayed in Fig. 18b and can be compared to the "real" one in Fig. 18a. One can notice the misplacement of the nasal vowels that naturally could not be estimated from the convex hull of articulatory vectors of oral vowels. Other than that, the vowels are organised rather well: the front vs. back as well as the close vs. open divisions are identifiable, the semivowels appear very close to their vowel counterparts.

We have selected all vowels but the corner ones, all semivowels, and all qualifying CV-syllables in the dataset—those which could be replicated by the projection of the vowel (V) onto the convex hull of corner vowels and the vocal tract configurations in the context of corner vowels. Then we compared the selection with the entries that were actually recorded, both visually and quantitatively in the articulatory parameter space (Eq. 8).



(a) The vowel space as defined by International Phonetic Association (IPA); the representation is grounded on acoustic (and articulatory) features.



(b) The space of vowels projected onto the convex hull of the corner vowels; the representation is grounded on the Euclidean distance between articulatory vectors.

Figure 18: Comparison of the IPA vowel space and the space of vowel projections

As can be seen in Fig. 19, the articulators that contribute most to the total error in the parametric space are the velum (nasal vowels, since all corner vowels are oral, and /e/, /ɛ/, and /o/), the tongue (which could be attributed simply to the fact that the tongue is encoded through most of the parameters), and the epiglottis. Except for absence of lowered velum within the corner vowels' convex hull which resulted in poor projection results for the nasal vowels, there does not seem to be any other pattern for which vowels turned out to be closer to the convex hull of the corner vowels (/œ/, /e/, /ø/, /ɔ/, and the semivowel /w/) and which were far (/ɛ/, /o/, /y/)—we can find exceptions to any definitive conclusion.

However, if we look for tendencies, we can note that the lips usually contributed to the error value in rounded vowels rather than in unrounded ones. But it should not be attributed to lips protrusion: except for /e/, the lips aperture always takes up a larger percentage of the projection error value.

One can also see that the jaw component was most present in semivowels, which can be explained by the fact that they are even closer that the close corner vowels /i/ and /u/.

As for consonant-vowel (CV) syllables, the distance between the original dataset sample and the estimated combination based on the projections seems to be related to the consonant rather than to the vowel in question (see Table 7). The main articulator to make the estimate go astray from the original vector is always either the tongue or the velum.

Visually there are several points to report on:

- In projections, the velum position would not allow to produce a nasal sound.

[27] /py/ is an exception: the main contributor here is the tongue.
[28] /ko/ is an exception: the main contributor to the distance is the velum.
[29] /ʁɑ̃/ is an exception: here, the main contributor to the distance is the velum.

C	e	ɛ	o	ø	y	ɑ̃	ɔ̃	ɛ̃	M	σ	Organ
p	N/A	**31.7**	**22.44**	N/A	**9.21**[27]	N/A	N/A	N/A	**21.12**	**9.23**	Velum
t	N/A	**6.71**	**7.26**	N/A	**10.69**	N/A	N/A	N/A	**8.22**	**1.76**	Tongue
k	N/A	**9.68**	**15.98**[28]	**9.31**	**15.87**	**17.35**	**17.12**	**16.61**	**14.56**	**3.24**	Tongue
l	**11.85**	**8.74**	**11.01**	N/A	**9.36**	N/A	N/A	**10.66**	**10.32**	**1.13**	Tongue
ʁ	N/A	**11.25**	**23.79**	N/A	**13.27**	**26.97**[29]	**15.64**	**17.49**	**18.07**	**5.59**	Tongue
m	N/A	**18.66**	**19.41**	N/A	**26.35**	N/A	**21.97**	N/A	**21.60**	**3.01**	Velum
n	N/A	**13.82**	**18.83**	N/A	**18.3**	N/A	**20.52**	N/A	**17.87**	**2.48**	Tongue
f	N/A	**20.84**	**24.25**	**21.49**	**22.45**	N/A	N/A	N/A	**22.26**	**1.29**	Velum
s	N/A	**8.61**	**7.38**	N/A	**7.55**	N/A	N/A	N/A	**7.85**	**0.54**	T&Vel
ʃ	N/A	**5.67**	**8.5**	**8.64**	**9.17**	N/A	N/A	N/A	**8.00**	**1.36**	Tongue
M	**11.85**	**13.57**	**15.89**	**13.15**	**14.22**	**22.16**	**18.81**	**14.92**			
σ	**0.00**	**7.65**	**6.49**	**5.91**	**6.05**	**4.81**	**2.54**	**3.03**			

Table 7: The Euclidean distance between the estimated and real syllable vectors. "**C**" stands for the consonant in the syllable that is followed by vowels from the columns; "M" stands for the mean value, "σ" for the standard deviation, and "**Organ**" for the main contributor to the distance value.

Some velum shapes are physically impossible.

- In many cases, though the tongue is roughly of the required shape, the constriction gets too wide or too narrow.
- Because of errors in the position of articulators, the estimate can contain a new point of constriction.
- The lips are often too open.

4.2 Evaluation Set

To get a comprehensive and balanced representation of possible phoneme combinations, we have selected:

- all vowels—12 samples;
- a set of vowel-to-vowel transitions, balanced in terms of height, backness, roundedness, and nasality (i.e. there necessarily is an open-open transition, an open-close-mid, open-open-mid...)—15 samples: /ɑ̃a/, /ae/, /aɔ/, /ɛ̃ɑ̃/, /iu/, /oø/, /ai/, /œi/, /yɔ̃/, /uɔ/, /ɔɛ̃/, /øy/, /ɔ̃e/, /uo/, /eɑ̃/, /oɛ/;
- for every consonant or a semivowel C, an utterance /Ca aCa/—18 samples;
- a set of VCV transitions, based on non-covered VV combinations—25 samples: /apu/, /ybi/, /ɛ̃te/, /ody/, /ika/, /œkɑ̃/, /ɔ̃ke/, /øgɛ/, /ygɔ̃/, /ɑ̃ga/, /ify/, /uvɛ̃/, /ɔ̃se/, /azi/, /uʃɔ̃/, /ɛʒa/, /ale/, /ily/, /oʁi/, /uʁɑ̃/, /ami/, /œmɔ̃/, /one/, /ɛny/, /awi/;

- a set of phrases, both with entries covered in the dataset and those that had to be estimated—7 samples:
  - "Bonjour" ("Hello") /bɔ̃ʒuʁ/,
  - "Camus" ("Camus"—last name—or "pug-nosed") /kamy/,
  - "Les abat-jours sont tombés" ("The lampshades have fallen down") /lɛzabaʒuʁ sɔ̃tɔ̃be/,
  - "Il a pas mal" ("He does not feel any pain") /ilapamal/,
  - "Je crois" ("I believe") /ʒœkʁwa/,
  - "Je sais que c'est banal mais le problème est là" ("I know it is trivial, but the problem is there") /ʒœsɛkœsɛbanal mɛː lœpʁoblɛmɛla/,
  - "Est-ce que tu peux... prendre un rendez-vous ?" ("Can you—take an appointment?") /eskœtypøː pʁɑ̃drəɛ̃ʁɑ̃devu/.

Not all of them will be informative, since the feature of nasality is not supported and has to be added manually.

4.3 Results

The overall intelligibility is rather low, which impeded carrying out perception tests. Some utterances were synthesised with a very low energy level. The main reasons for noisy results are the following:

- Errors made at earlier levels—such as involuntary changes of the vocal tract configuration by the speaker during the experiment or limitations of the articulatory model—become propagated.
- Estimation of the 3D volume by a 2D shape is a vulnerable point, and the acoustic simulation unit itself is so too. Sometimes the pictured vocal tract evolution is correct, but the sound synthesised from it is not.
- A limitation inherent in our usage of the articulatory model: we do not control the presence or absence of the constriction and its width.
- There need to be more experiments to learn to avoid artefacts at obstruents, which diminish the intelligibility drastically.
- We do not have precise measurements of the nasal cavities, and the area functions built on the data provided by Serrurier (2006) seem to be unadapted to the subject—when we add nasality, there is too much energy radiated by the nostrils in nasal sounds.

So the evaluation criteria will be as follows:

- Visually: evaluate the trajectories of the articulators;

- Acoustically: compare the formants of synthesised signals to the real ones;
- Perceptually: look for auditory cues.

For the vowels, we compare the formants[30] with those reported for male speakers by Lonchamp (1984). All formant frequencies are indicated in Hz.

/**a**/: F1 is 632 Hz, F2 is 1325 Hz, and F3 is 2349 Hz. Reported values: 684 Hz, 1256 Hz, and 2503 Hz. The sound is perceived as /a/, but not clearly.

/**e**/: 547, 1666, and 2547 Hz vs. 365, 1961, and 2644 Hz, which is not very close, but the vowel still can be recognised.

/**ɛ**/: 619, 1595, 2661 Hz vs. 530, 1718, and 2558 Hz. The vowel can be recognised correctly.

/**i**/: 380, 1904, 2928 Hz vs. the reported 308, 2064, and 2976 Hz. It sounds like /i/.

/**o**/: 481, 933, 2198 Hz vs. 383, 793, and 2283 Hz. The sound is good, but the velum shape experiences an extreme thinning on the way from silence to /o/ and back.

/**ɔ**/: 715, 1048, 2609 Hz vs. 531, 998, and 2399 Hz. The formants of the synthesised speech signal have a better match with our recordings of a human speaker: 852, 1104, and 2789 Hz. The sound can be recognised.

/**œ**/: 452, 1805, 2916 Hz vs. 517, 1391, and 2379 Hz, though the signal comes with a very low energy level and clipped.

/**ø**/: 452, 1285, 2033 Hz vs. 381, 1417, and 2235 Hz. The sound can be recognised.

/**u**/: 527, 1090, 2195 Hz vs. 315, 764, and 2017 Hz, though the signal comes with a very low energy level and clipped due to the fact that the articulatory model cannot interpret the lips' position correctly and makes a complete obstruction of the vocal tract. This also lead us to be unable to evaluate the syllables and phrases which included /u/.

/**y**/: 301, 1596, 3765 Hz vs. 300, 1750, and 2120 Hz, which means a completely wrong value of F3. But the perceived sound is not bad thanks to the F1 and F2 that are close to the real values.

For the nasal vowels and nasal consonants, the area of the nasal tract has to be added separately, so they are not included here.

For other consonants, we can conclude that:

Stops, in the same way as Birkholz (2013) concluded, are recognised well because of their common perception cue. For example, the essential feature to produce /k/ is the velar pinch, when F2 and F3 come very close to each other just before the burst, and it is handled by the model very well (see Fig. 20). Among all plosives, just like in the work by Birkholz, /g/ seems to be the best recognisable plosive. But other stops can be confused, especially because of the acoustic artefacts. To enhance the results, more refined rules are due.

[30]The formants are estimated with WinSnoori: `http://www.loria.fr/~laprie/WinSnoori/index.html`.

Approximants and **semivowels** require a more glide-like articulatory position control, because their target is rather a gesture than a fixed point in space. There is no fluidity in the synthesised motions of the articulators.

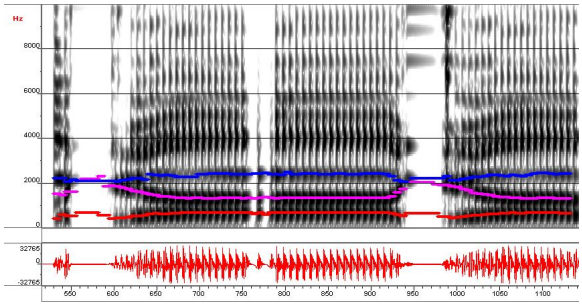


Figure 20: Spectrogram of the synthesised /ka aka/ speech signal. One can observe the velar pinch—the essential feature for producing the phoneme /k/.

Fricatives: as discussed above, the model makes it impossible to correctly synthesise labiodental fricatives, and, additionally, the projection-based estimation of missing samples often makes the constriction not pronounced enough. The sound could not be evaluated because of the artefacts brought by the acoustic simulation unit. /ʁ/ is a special case: there is no artefact, but the area functions do not capture the contact between the tongue and the velum, resulting in an approximant-like sound.

In comparison to human speech, the dynamics of the formants in the synthesis results (so, what the coarticulation model is actually responsible for) is rather good. There seems to be a need for CV-transitions to occur faster, though (see Fig. 21).

The phrases comprised consonant clusters, where model had to establish whether the vowel is already anticipated or not yet. If we denote anticipation with parentheses *( )* and syllable segmentation with "–", then the model's conclusions are as follows:

- /b(ɔ̃) ɔ̃ – ʒ(u) u ʁ/;
- /k(a) a – m(y) y/;
- /l(ɛ) ɛ – z(a) a – b(a) a – ʒ(u) u ʁ – s(ɔ̃) ɔ̃ – t(ɔ̃) ɔ̃ – b(e) e/;
- /i – l(a) a – p(a) a – m(a) a l/;
- /ʒ(œ) œ – k ʁ(w) w a/;
- /ʒ(œ) œ – s(ɛ) ɛ – k(œ) œ – s(ɛ) ɛ – b(a) a – n(a) a l – m(ɛ) ɛː – l(œ) œ – p(o) ʁ(o) o – b(ɛ) l(ɛ) ɛ – m(ɛ) ɛ – l(a) a/;
- /e s(œ) – k(œ) œ – t(y) y – p(ø) øː – p(ɑ̃) ʁ(ɑ̃) ɑ̃ – d(ɛ) r(ɛ) ɛ – ɛ̃ – ʁ(ɑ̃) ɑ̃ – d(e) e – v(u) u/[31].

Note the cross-syllable coarticulation in /e s(œ) – k(œ) œ –. . . / and different treatment of consonant clusters in /k ʁ(w) w a/ and /p(ɑ̃) ʁ(ɑ̃) ɑ̃/.

The trajectories of the articulators seem to be well-handled by the approach of articulatory targets: the implemented coarticulation model makes the movements

[31]Here is the problem with the absence of /ə/: it should be /p(ɑ̃) ʁ(ɑ̃) ɑ̃ – d(ə) r(ə) ə ɛ̃ . . . / or /p(ɑ̃) ʁ(ɑ̃) ɑ̃ – d(ə) r(ə) ɛ̃ . . . /

ergonomic and natural. The only concern is absence of the undershoot effect, but it is going to be dealt with in the future implementations.

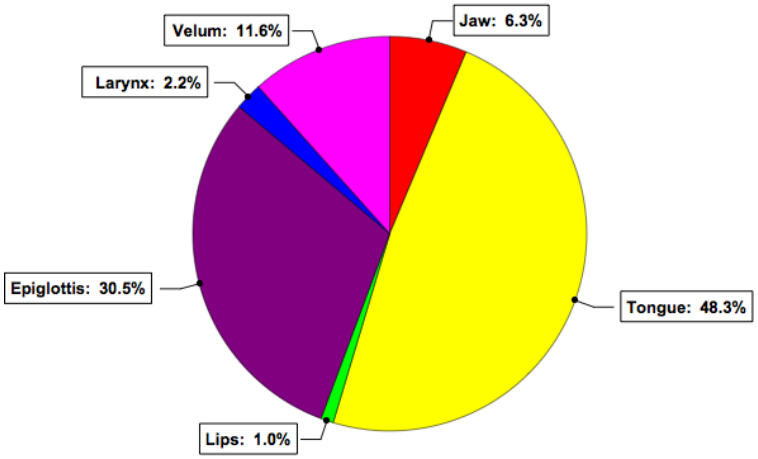
The intelligibility of phrases was considerably hindered because of the acoustic artefacts brought by the errors at the earlier junctures of the system and some incongruence in acoustic simulation control. We chose the example without artefacts, /kamy/, to correct it manually. After making the /my/ transition faster and adding nasality, the word became easily recognisable (see Fig. 22 for the spectrograms).

Another comparison was done on /ilapamal/, but without any corrections: see Fig. 23. Again, one can observe the necessity of timing adjustments.

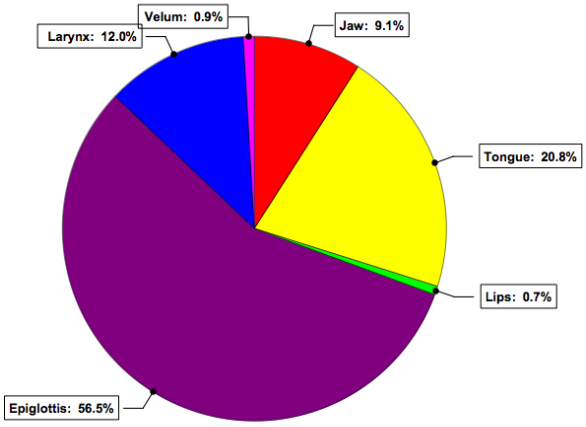
Just like in the work of Birkholz, generalisation of any example-driven corrections will be possible only with more sophisticated rules. As the /kamy/ example shows, most of the work should be dedicated to the timing strategies, for they play a crucial role.

When the acoustics of isolated sounds are more correct, we will be able to perform more substantial evaluation:

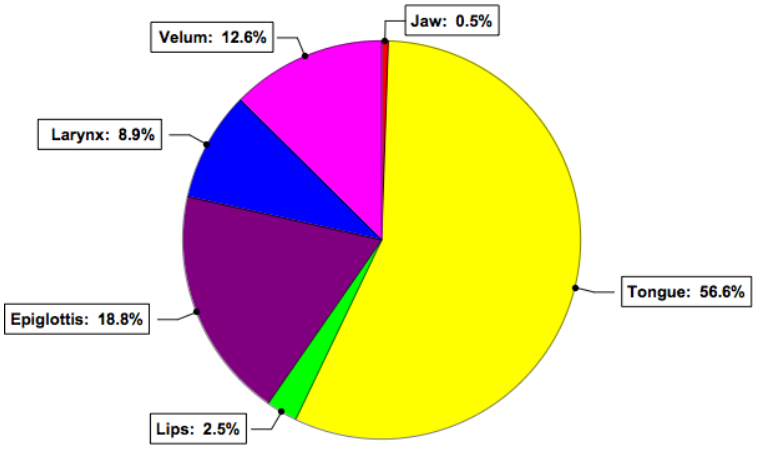
- Statistically compare different coarticulation strategies by the difference in listeners' recognition rate of the same utterance synthesised with the investigated strategies;
- Compare the formants of the real and synthesised sounds by computing the correlation between their trajectories, their spectra, and their mel-frequency cepstral coefficients.
- Compute the correlation between real and synthetic formant trajectories.



(a) /j/: the error of **16.6**

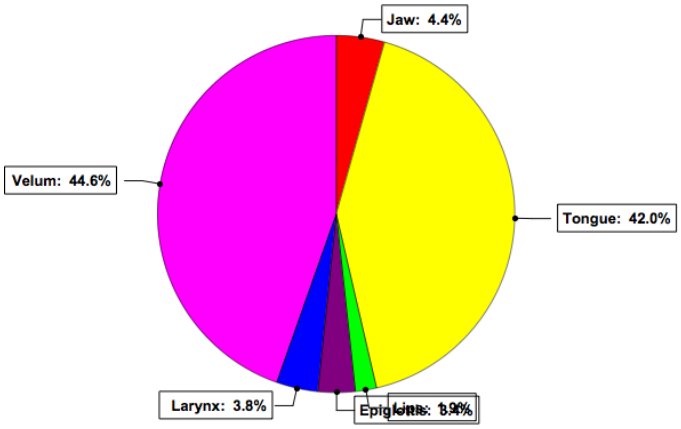


(b) /w/: the error of **11.57**

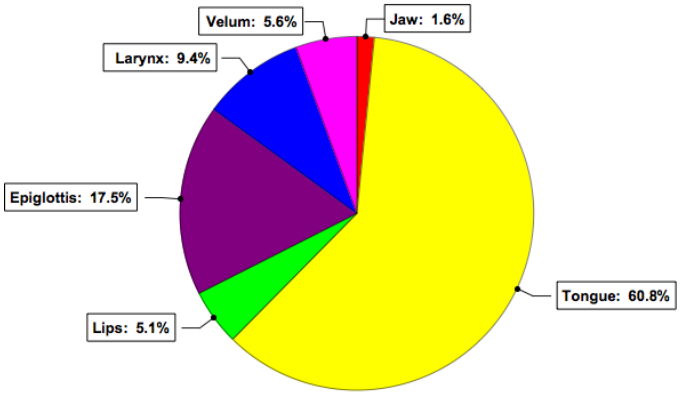


(c) /y/: the error of **19.02**

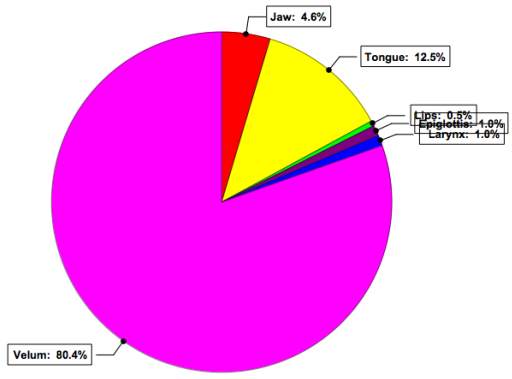
Figure 19: Projection onto the space of corner vowels: contributions to the global distance between the actual vector and its projection, by component articulators.
Continued on next page



(d) /e/: the error of **9.02**

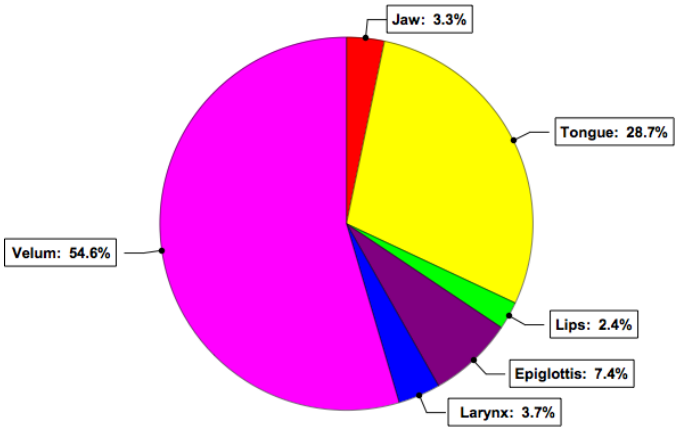


(e) /ø/: the error of **12.28**

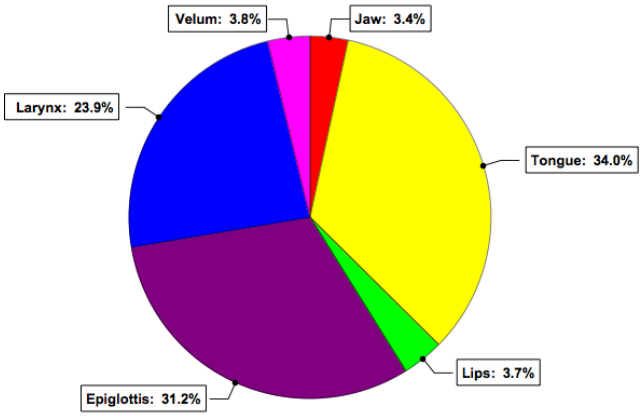


(f) /ɛ/: the error of **21.14**

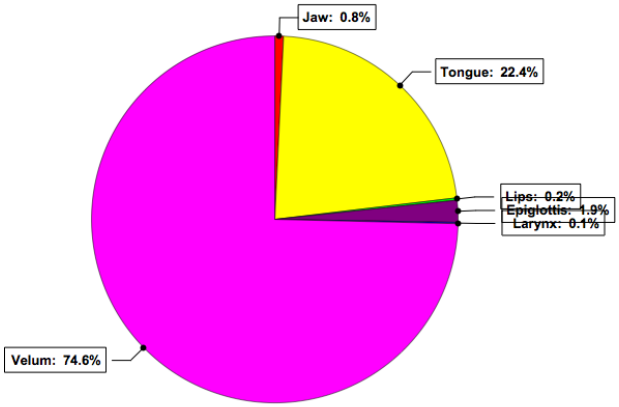
Figure 19: *Cont.*: Projection onto the space of corner vowels: contributions to the global distance between the actual vector and its projection, by component articulators. *Continued on next page*



(g) /ɛ̃/: the error of **15.08**

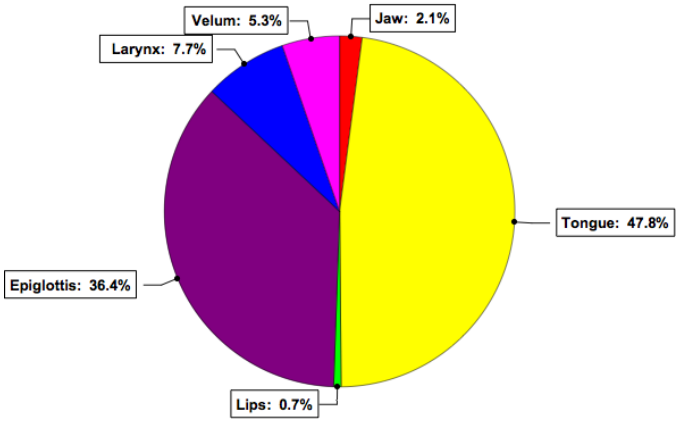


(h) /œ/: the error of **7.59**

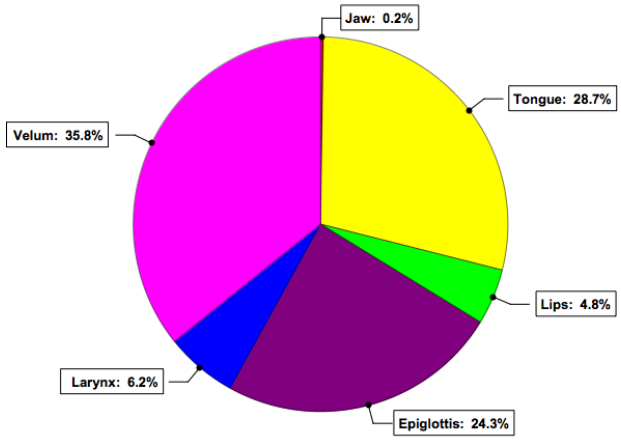


(i) /o/: the error of **21.11**

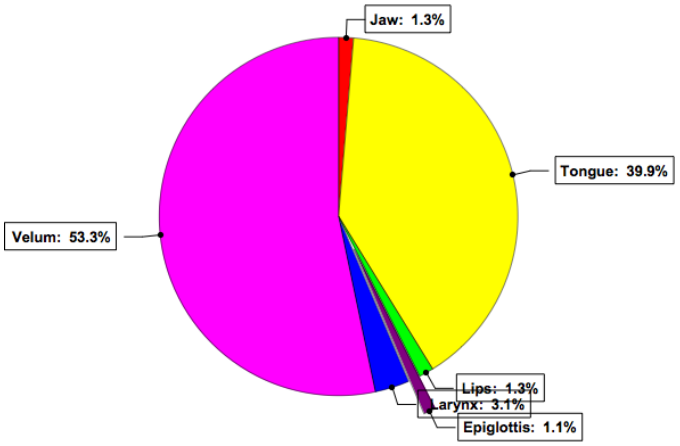
Figure 19: *Cont.*: Projection onto the space of corner vowels: contributions to the global distance between the actual vector and its projection, by component articulators. *Continued on next page*



(j) /ɔ/: the error of **13.98**

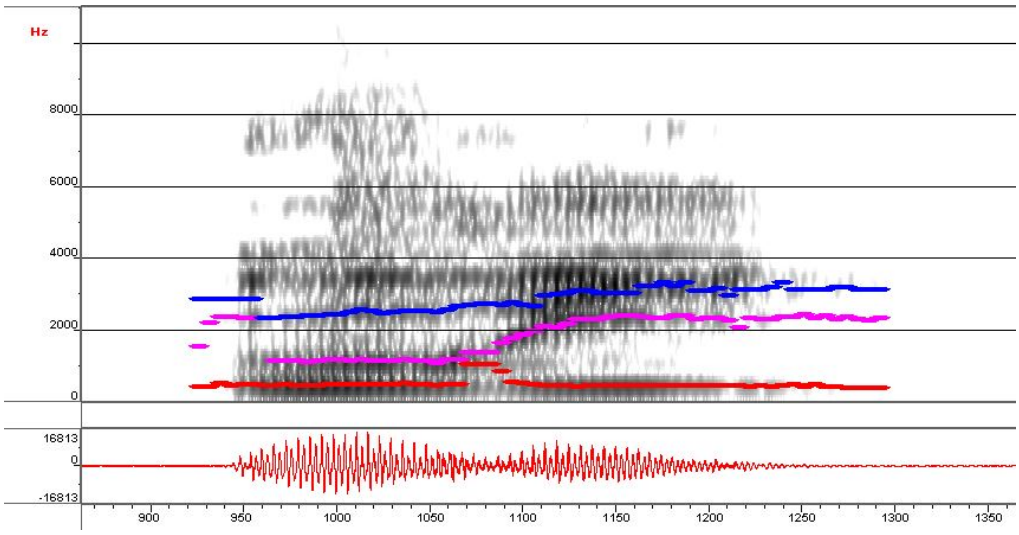


(k) /ɔ̃/: the error of **18.31**

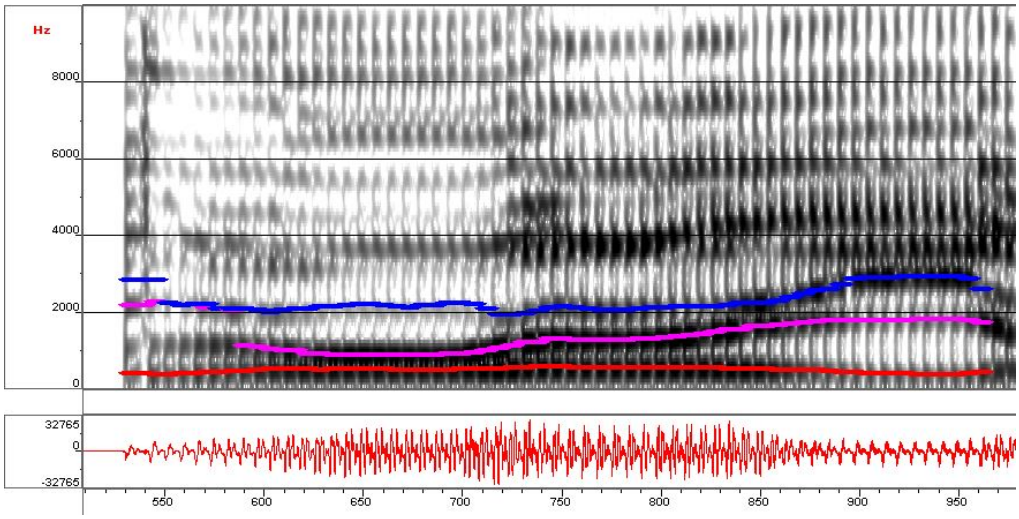


(l) /ɑ̃/: the error of **19.94**

Figure 19: *Cont.*: Projection onto the space of corner vowels: contributions to the global distance between the actual vector and its projection, by component articulators.

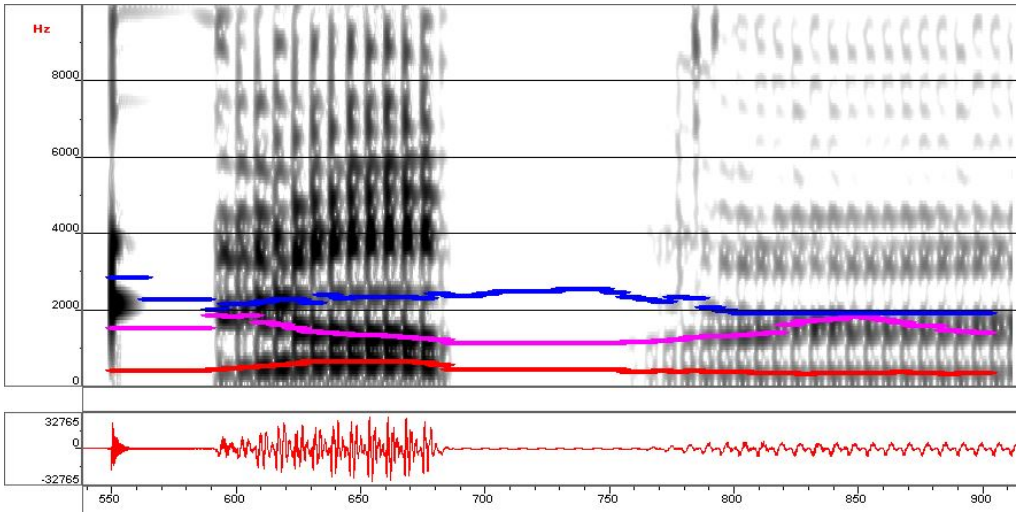


(a) /оʁі/ uttered by human.

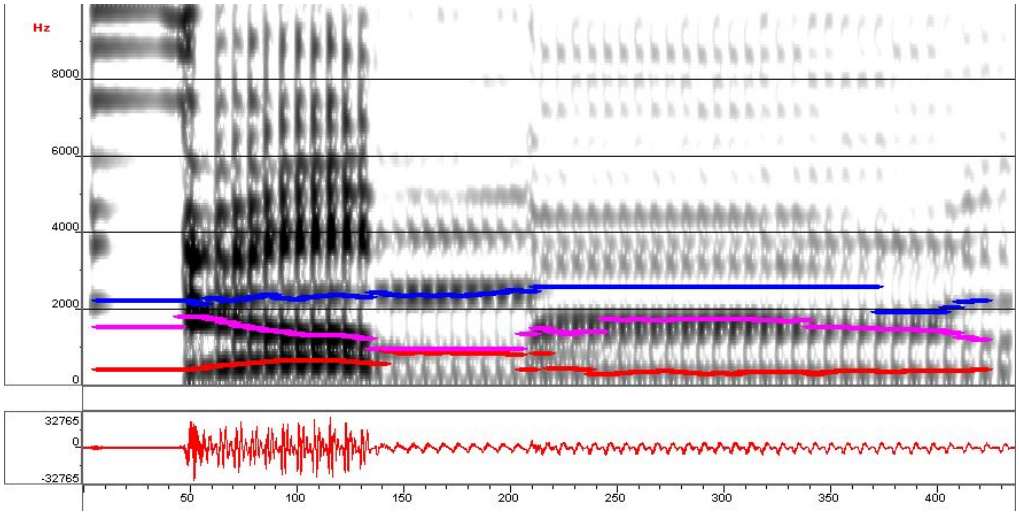


(b) /оʁі/ synthesised by the system.

Figure 21: Spectrograms for the same utterance being uttered by human and synthesised by the system. F1 is highlighted with the red colour; F2 is pink, and F3 is blue.

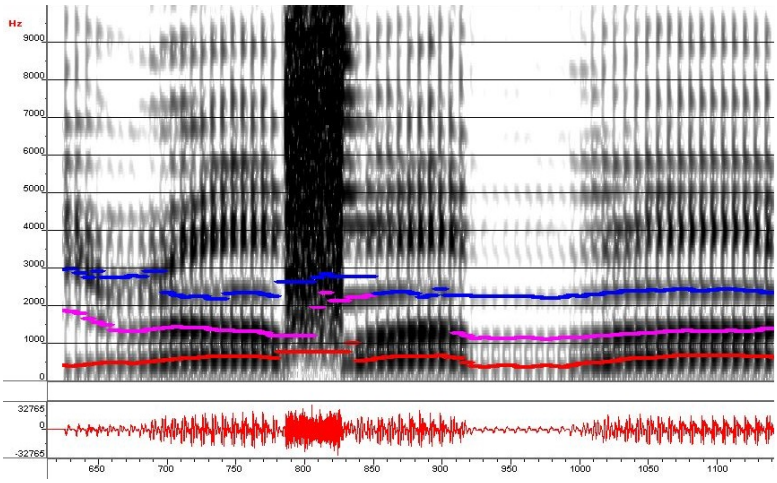


(a) /kamy/ as derived by the model.

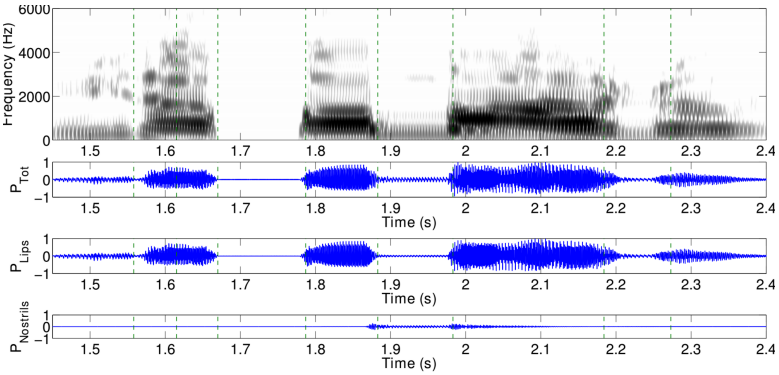


(b) /kamy/ after corrections.

Figure 22: Spectrograms for the same utterance before and after adjusting the temporal control and adding nasalisation. F1 is highlighted with the red colour; F2 is pink, and F3 is blue.

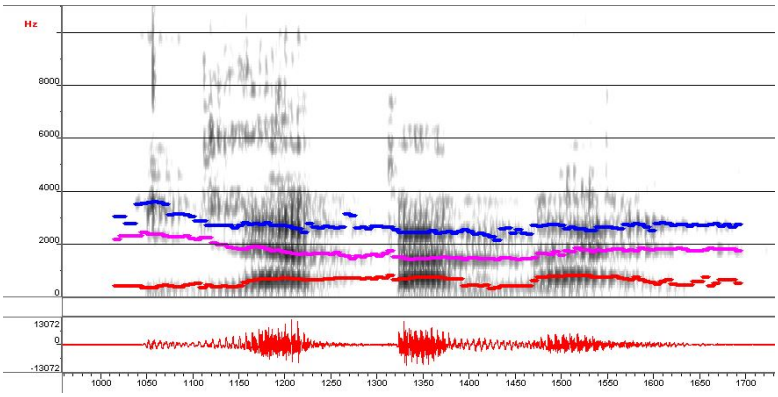


(a) /ilapamal/ as derived by the model.



/i/ /l/ /a/ /p/ /a/ /m/ /a/ /l/ /ə/

(b) /ilapamal/ from Laprie et al. (2015): spectrogram, total acoustic pressure signal—P_{Tot}, acoustic pressure radiated at the lips—P_{Lips}, and acoustic pressure radiated at the nostrils—$P_{Nostrils}$.



(c) /ilapamal/ uttered by human.

Figure 23: Spectrograms for the same utterance pronounced by a human speaker and synthesised in the present system and in the study by Laprie et al. (2015).

5 Conclusions

The work addresses a wide range of issues in articulatory speech synthesis, unlike most other studies in the field, aiming for a full-scale coverage of various speech phenomena rather than attending solely to specific points.

It is inspired by the work of Birkholz (2013) and Öhman (1967), adapting it for the specifics of the French language. The system joins an existing dataset of static pictures, an articulatory model by Laprie and Busset (2011b); Laprie et al. (2014), a new coarticulation model, an area function estimation system developed in Xarticul, and the acoustic simulation by Elie and Laprie (2015, 2016).

Since the presented original part of the work heavily depends on other units and their interplay, the evaluation was performed jointly. The evaluation set consisted of vowels, vowel-to-vowel and vowel-consonant-vowel transitions, and phrases. Our conclusions reached in Section 4.3 are based on:

- Visual analysis of the generated animation for the articulators' transition;
- Comparison of the generated signals to that of human speakers in terms of the F1, F2, and F3 formant frequencies;
- Comparison of the F1, F2, and F3 formant dynamics in a real and synthesised signal (evaluating the temporal control of perceptually important features, e.g. how quickly a consonant-vowel transition is made);
- Analysis of the acoustic signal from the perceptual point of view.

5.1 Proposed Amendments and Future Work

On a small scale, there are specific points that may or have to be addressed in the next implementation:

- The model could use a set of cardinal vowels extended by /y/, which is a highly relevant vowel for the French language because of its extreme lips protrusion.
- The s and q coefficients calculated for each vowel (Eq. 7) can be optimised in the acoustic space rather than in the space of articulatory vectors, as was done by Birkholz.
- The timing for the CV-transitions has to be adjusted: they should occur faster.
- The control over articulatory vectors should become more fine-grained, singling out separate articulators and differentiating between the high- and low-priority areas.
- Getting the timing control correct is crucial.
- More refined intraoral and subglottal pressure handling is due (see Section 3.3.4).

- Production of the fricative phonemes is being improved in the acoustic simulation system by Elie and Laprie; hopefully, we will be able to avoid producing artefacts.
- Nasality should be supported at least for the chosen speaker.
- Quantifiable tests would allow us a more comprehensive comparison with the results of Birkholz's work.
- The artefacts could be minimised by means of a better interplay between the source and constriction.

On a larger scale, it could be a good idea to avoid cumulating errors at every juncture of the system and try to interweave the operations of separate units, which would also have a benefit of faster computation.

Besides, the articulatory control unit is completely rule-based, which gives a solid baseline for what a rule-based, static-data-driven approach is capable of. It is a good starting point for further development of the articulatory control model, which could benefit greatly—in terms of the spatial and temporal control over the articulators—from some kind of dynamic input. For instance, the present work could be revised with real-time MRI which has proven its ability to bring in new information for the speech processing community (Narayanan et al., 2004; Byrd et al., 2009; Ventura et al., 2009; Ramanarayanan et al., 2010; Echternach et al., 2010; Teixeira et al., 2012; Burdumy et al., 2015).

All in all, we hope that the present work will be of benefit to the speech synthesis research. Since articulatory speech synthesis allows us to have the full control over the articulators, it can be the answer to many problems arising in the state-of-the-art unit selection method such as prosody control and adaptability for multiple speakers; and finally, it is of scientific interest for the field of speech production theories.

Appendix: Detailed Program Structure

- The `VocalTract` class

  A class for describing the used vocal tract model.

  - `__init__`(self)

    Configures the parameters of the vocal tract model.

    Attributes `self.nbjaw` (3), `self.nbtongue` (12), `self.nblips` (2), `self.nbepiglottis` (2), `self.nblarynx` (2), and `self.nbvelum` (5) (all integers) set the number of parameters per each articulator.

    Articulator's name is set at the corresponding position in `list paramnames` (list of strings: ["Jaw", "Tongue", "Lips", "Epiglottis", "Larynx", "Velum"], from where it is transferred to all attributes where it is needed (see below).

    The user can also use the short name which is to be set in `parampseuds`, a list of strings (["J", "T", "Li", "E", "La", "Vel"]), from where it is transferred to all attributes where it is needed (see below).

    Any articulator can be split into named subarticulators in `self.complexarticulators` with the following structure: *key*: name of the articulator to be split (string: "Lips"); *value*: list of tuples, where each tuple is of three values: full name of the part of the articulator (string: "LipsAperture" or "LipsProtrusion"); its short name (string: "LiA" for lips aperture, and "LiPr" for lips protrusion); how many parameters are assigned to this part (1, 1). Take note that the parts are ordered (i.e. the fact that lips aperture is the parameter №16 and lips protrusion is the parameter №17 is determined from the order in which these subarticulators are entered).

    `self.shortnames` and `self.fullnames` are dictionaries for translating short names into full names and vice versa.

    `self.articulators` is a dictionary whose keys are strings that are full or short names of articulators or their parts, and values are lists of integers: the parameter numbers in an articulatory vector that corresponds to the particular articulator or its part.

    `self.fullnarticulators` is a copy of `self.articulators`, but only with fully named keys.

    `self.superficialorder` contains the correct order of superarticulators (list of strings, the elements are full names of articulators).

    `self.lowerorder` (list of strings) contains the correct order of the lowest layer of articulators: either simple ones or their subparts instead of superarticulators.

    `self.complexorder` is a list whose elements are strings (for the case of a simple articulator) or lists of strings (for a complex articulator).

    `self.totalparams`: integer - how many articulatory parameters encode a vocal tract configuration.

`self.nbarticulators`: integer - how many articulators are distinguished in a vocal tract.

`self.which` is a list of length `self.totalparams`; each element in it tells which articulator is responsible for this position, e.g.: ["Jaw", "Jaw", "Jaw", "Tongue", ...]. Complex articulators are handled as lists: ["Jaw", ..., "Tongue", ["Lips", "LipsProtrusion"], ["Lips", "LipsAperture"], "Epiglottis", ...]

- `fetch_missing_articulators(self, articulators)`

  For one articulator or a set of articulators, this method returns the complementary set to cover the full vocal tract, handling the short and full articulator's names and complex articulators.

  Input:

  `articulators`: string or list of strings: name of an articulator or of a part of it or a list of articulators' names

  Output:

  `missingarts`: list of strings: the complementary list of names of articulators or their parts (e.g. "Tongue", "LipsProtrusion").

- `reorder_articulators(self, articulators, vocal=True)`

  Returns a correctly reordered list of articulators.

  Input:

  `articulators`: list of strings (names of articulators or their subparts, either full or short).

  `vocal`: boolean whether to print warnings about encountered ambiguous orderings. By default, True (yes, to print). *Output*: reordered input list.

- `parameter_numbers(self, articulators)`

  For a given set of articulators' names, returns the set of articulatory parameters they are responsible for.

  Input:

  `articulators`: a string or list of strings: names of articulators, in their full or short forms.

  Output:

  list of integers: which parameters they are responsible for.

- The `Phoneme` class

  A class for relating the phoneme dataset files to the phonemes as handled in speech synthesis.

  - `__init__(self, name, folderpath="Data/Speech-Synthesis-Database/DB-Full/", extension=".artv")`

    Parses the file with a given phoneme.

    Input:

`name`: string, the name of the phoneme;

`folderpath`: string, the location of the dataset of phonemes. By default, `folderpath` is set to "Data/Speech-Synthesis-Database/DB-Full/".

`extension`: string, the extension of the phoneme files. By default, `extension` is set to ".artv" which is the format of the database after its expansion.

Output: self.

- `slice(self, articulators, key="Solo")`

  Returns a list of the parameters for a particular articulator in an articulatory vector or a list of articulatory vectors.

  Input:

  `articulators`: string or list of strings matching the `VocalTract` class names.

  `key`: string. It can be the name of the phoneme that is anticipated, or "Solo" to get the pure vowel or consonant configuration. By default, it is set to "Solo".

  Input: list of floats.

- The `Syllable` class

  A class for managing necessary phonemes in a syllable.

  - `__init__(self, sylltext, folderpath= "Data/Speech-Synthesis-Database/DB-Full/", extension=".artv")`

    Input:

    `sylltext`: string, the syllable to be processed; multicharacter phonemes are to be put in curly brackets: "{...}". For other special characters, see below.

    `folderpath`: string, the location of the database with the phonemes. By default, `folderpath` is set to "Data/Speech-Synthesis-Database/DB-Full/".

    `extension`: string, the extension of the phoneme files. By default, `extension` is set to ".artv" which is the format of the database after its expansion.

    Obtains a sequence of articulatory targets for the given input syllable and puts them into the `targets` attribute:

    * `self.constinuents` = list of `Phonemes`
    * `self.anticipated` = list of keys for the `artv` attribute of `Phonemes` in `self.constituents`. An example for a CCVC-syllable: `self.targets = ["a", "a", "Solo", "Solo"]` This means that the first articulatory vector that can be used as the target is `/TheFirstConsonant/.artv["a"]`, and then come `/TheSecondConsonant/.artv["a"]`, `/Vowel/.artv["Solo"]` (i.e. the pure vowel position), and, finally, `/Consonant/.artv["Solo"]`. These target vectors will be operated on by the coarticulation model.

There are special markers allowed in the sequence. The parsing algorithm assumes the following order:

1. Stress-related symbols (if excluded, the syllable is not stressed);
2. Melody-related symbols (if excluded, the intonation contour is even);
3. Phonemes in the syllable;
4. Phoneme that is intended long should be immediately followed by the prolongation-related symbol (if there is none, the phoneme is of regular duration).

Among all the types of stress (rhythmic, syntagmatic, secondary, emphatic...), degrees of phoneme prolongation, and melodic patterns, this implementation makes use of the bare minimum:

* ' The normal rhythmic stress which falls onto the last syllable of a rhythmic group;
* / Rising intonation;
* _ Even intonation contour;
* \ Falling intonation (since it is a special character, it has to be escaped: \\);
* : Long phoneme (vowel or consonant; can be used for geminated stops too).

The corresponding attributes store this information:

* `self.contour`: string. Contains graphics for intonation, e.g. _/, _/, \. If no information has been given, `self.contour` = "_" (even).
* `self.stress`: boolean. Whether the current syllable is stressed.
* `self.lengths`: list of strings. The elements are in one-to-one relation with the phonemes in the syllable. "Reg" means regular phoneme duration, and "Long" means long.

If the user marks the boundary between syllables in the user input, then the syllable may begin by a sequence for the intonation contour, then there may be an accent sign, and then should come the phoneme names, e.g.: "/_'{zh}uːr".

If the user relies on the syllable segmentation algorithm, the stress mark has to be immediately preceding the vowel it is on.

Output: `self`.

– `express(self, vocal=True, anticipations=list())`

Prints the constructed syllable in terms of its constituents.

Input: `vocal`: boolean - whether to print the output. By default, yes (True).

`anticipations`: list of keys for the artv attributes in `Phoneme`. When empty, the usual transcription will be returned. By default, it is empty.

Output: string, the textual representation of the syllable.

- `is_polyvocalic(self)`

  Determines if the syllable is actually a syllable or a merging of phonemes (when the user relies on the implemented syllable segmentation).

  Output: boolean: True for the case "it is actually several syllables", False for "it is one syllable".

- `split(self, syntext, folderpath="Data/Speech-Synthesis-Database/DB-Full/", extension=".artv")`

  Splits a syllable that did not have the user input information into more syllables as per the algorithm described by Bigi et al. (2010).

  Input:

  `syntext`: string, text with markings on the intonation and stress, but without separation into syllables.

  `folderpath` = "Data/Speech-Synthesis-Database/DB-Full/",

  `extension` = ".artv"

  For example, entry "\\l{epsilon}-\\za-\\ba-/_'{zh}u:r", "s{o~}-t{o~}m-\\'be" may become `syntext` if there is no information on syllable segmentation: "\\l{epsilon}\\za\\ba/_{zh}'u:r", "s{o~}t{o~}m\\b'e".

- The `Syntagm` class

  A class for processing whole groups of phonemes in an utterance and constructing articulatory targets according to the assumptions of the implemented coarticulation model and taking note of coarticulatory effects that span over syllable boundaries. The name "syntagm" was chosen as a convenience: actually it is closer to "rhythmic groups".

  - `__init__(self, syntext, folderpath= "Data/Speech-Synthesis-Database/DB-Full/", extension=".artv")`

    Input:

    `syntext`: string, the rhythmic group being processed. Boundaries between syllables are to be marked by hyphen, "-". The user can also rely on the syllable segmentation algorithm implemented within the `Syllable` class (`see Syllable().split()`) and not mark any boundaries between syllables at all, or mark it only at a place where the algorithm makes a mistake.

    There are syntax rules for syllables: see documentation of the `Syllable` class, applying to the case of syllable segmentation being marked by user or by the program.

    `folderpath`: string, the location of the database with the phonemes. By default, `folderpath` is set to "Data/Speech-Synthesis-Database/DB-Full/".

    `extension`: string, the extension of the phoneme files. By default, `extension` is set to ".artv" which is the format of the database after its expansion.

    Output: `self`.

- `express(self, vocal=True, showcoart=False)`

  Prints the constructed syntagm in terms of syllables it consists of.

  Input:

  `vocal`: boolean - whether to print the output. By default, yes (True).

  `showcoart`: boolean, whether to include which are the most prominently anticipated vowels and semivowels in the transcription. By default, not (False).

  Output: string, the textual representation of the group of phonemes.

- The `Utterance` class

  A class for breaking the utterance into rhythmic groups and syllables, forming articulatory targets within the assumptions of the implemented coarticulation model, and writing all the files that are necessary to synthesise the given utterance.

  - `__init__(self, uttertext, folderpath= "Data/Speech-Synthesis-Database/DB-Full/", extension=".artv")`

    Input:

    `uttertext`: string, the utterance being processed. Boundaries between rhythmic groups / syntagms (the current implementation does not differentiate between various pauses) are to be marked by " | ", i.e. vertical line | enclosed in spaces on both sides. There are syntax rules for syntagms and syllables: see documentation of the `Syntagm` and `Syllable` classes respectively.

    `folderpath`: string, the location of the database with the phonemes. By default, `folderpath` is set to "Data/Speech-Synthesis-Database/DB-Full/".

    `extension`: string, the extension of the phoneme files. By default, extension is set to ".artv" which is the format of the database after its expansion.

    Output: `self`.

  - `express(self, vocal=True, showcoart=False, onlytextual=False, limit=None, protectcurlybraces=False)`

    Prints a constructed utterance in terms of syntagms it comprises.

    Input:

    `vocal`: boolean, whether to print the output. By default, yes (True).

    `showcoart`: boolean, whether to include which are the most prominently anticipated vowels and semivowels in the transcription. By default, not (False).

    `onlytextual`: boolean, whether to strip the resulting string of all special characters. By default, not (False).

    `limit`: integer: what is the maximally allowed length of the result. If None, no limit is forced. By default, None.

`protectcurlybraces`: boolean: whether to protect the "{}" signs with the purpose of further usage of string.format(). By default, no.

Output: string, the textual representation of the utterance.

– `how_long(self, t0=2000, pause=40, iterstep=10, coartmode="COS", speechrate="Normal")`

Calculates how long the utterance is going to take.

Input:

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

integer: the total duration of the utterance, in ms.

– `temporal_grid(self, t0=2000, pause=40, iterstep=10, coartmode="COS",speechrate="Normal")`

Determines what are the moments in time when, according to given iteration step, coarticulation mode, and speech rate, vocal tract configuration samples are going to be generated.

Input:

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`grid`: list of integer moments in time, in ms, when we set an articulatory configuration to occur.

`griddecipher`: list of strings, where all elements correspond to the elements in `grid` and explain what is happening at this moment (see the table below).

`gridvoicing`: list of boolean values: whether or not vocal folds vibrate at a particular moment (which one, is determined from the corresponding element in grid). It is necessary to store it separately, not take it from the phoneme itself, because phonemes are not necessarily pronounced with voice during all their production.

`gridaddresses`: list of tuples (q, n, k), where all elements are integers and indicate the phoneme in production in the following way:

```
self.synts[q].sylls[n].constituents[k]
```

If currently nothing is being produced, the tuple is replaced by "#". So, `gridaddresses` may look like `["#, "#", (0,0,0), (0,0,1), (0,0,2), (0,1,0), "#", "#"...]`

Elements of `griddecipher`:

* "#" Silence
* "o"+... The vocal tract is open. Producing an open vowel.
* "mo"+... The vocal tract is open-mid. Producing an open-mid vowel.
* "mc"+... The vocal tract is close-mid. Producing a close-mid vowel.
* "c"+... Producing a closed vowel.
* "v"+... The vocal tract is open, but it is not specified how much.
* "A"+... Articulators are positioned for an approximant or for a semivowel.
* "F"+... Articulators are positioned for a fricative.
* "S"+... Articulators are positioned for a stop.
* "L"+... Articulators are positioned for a liquid consonant.
* "N"+... Articulators are positioned for a nasal consonant.
* "C"+... Articulators are positioned for a consonant, but it is not specified of which kind.

Additionally, there are signs for different stages of phoneme production:

* "->" Reaching for the phoneme target position if positioned before the phoneme class sign, and transitioning from there if it comes after it:

        - "->o": transitioning to the target for an open vowel;
        - "mc->": transitioning from the target position;
        - "->F": articulators are approaching each other to make constriction required for a fricative;
        - "S->": after the burst phase of a stop.
    * "*" The burst phase of a stop: "S*".
    * "!" Being in the target position, e.g.:
        - "c!": production of a closed vowel when the target configuration has been reached;
        - "S!": the hold phase of a stop.
    * "~!" Being near the target position but not necessarily in it. It is expected to be used in liquids, semivowels, approximants:
        - "A~!" — the resulting sequence for an approximant may be "->A", "->A", "A~!", "A~!", "A!", "A~!", "A->".

– `how_many_samples(self, t0=2000, pause=40, iterstep=10, coartmode="COS", speechrate="Normal")`

  Calculates how many iterations in vocal tract configuration it will take to synthesize the given utterance.

  Input:

  `t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

  `pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

  `iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

  `coartmode`: string, the mode of the coarticulation model. Expected values:

  * "LIN": linear transition between target vectors;
  * "COS": cosine (smoother) transition between target vectors;
  * "COMPLEX": cosine transition with finer operation of articulators.

  By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

  `speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

  Output:

  integer: the total number of samples for the utterance, in ms.

– `record_art_vectors(self, artvectfile="Syntheses/{}_{}_data/{}_{}.xxx", t0=2000, pause=40, iterstep=10, coartmode="COS", speechrate="Normal")`

  Records the articulatory vector sequence for the utterance.

  Input:

`artvectfile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.xxx".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

– `record_af_list(self, affile="Syntheses/{}_{}_data/{}_{}.af", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Records the area function file sequence for the utterance.

Input:

`affile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.af".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None. Writes an *.af file containing the sequence of area functions involved.

– `record_glottal_pressure(self, glpressfile="Syntheses/{}_{}_data/{}_{}.ag0", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Records the glottal pressure changes for the utterance.

Input:

`glpressfile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.ag0".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an *.ag0 file that controls glottal pressure.

– `record_vocal_folds(self, vocfoldsfile="Syntheses/{}_{}_data/{}_{}.agp", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Records a file regulating vocal folds oscillations in the utterance.

Input:

`vocfoldsfile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.agp".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an *.agp file that controls vocal fold oscillations.

– `record_intonation(self, intonfile="Syntheses/{}_{}_data/{}_{}.f0", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Records the fundamental frequency for the utterance synthesis.

Input:

`intonfile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.f0".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an *.f0 file that controls voice pitch.

– `record_phonetic_description(self, phonfile="Syntheses/{}_{}_data/{}_{}.utt", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Records a file explaining the temporal grid.

Input:

`phonfile`: string, the path for the output file. By default, it is set as "Syntheses/{}_{}_data/{}_{}.utt".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an *.utt file that explains the temporal grid.

– `record_xart_script(self, xartfile="Syntheses/generateAFs_{}_{}.xart", artvfile="Syntheses/{}_{}_data/{}_{}.xxx", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal")`

Makes a script for Xarticul to create area function files.

Input:

`xartfile`: string, the path for the output file. By default, it is set as "Syntheses/generateAFs_{}_{}.xart".

`artvfile`: the path to the file with articulatory vectors. By default, it is set as "Syntheses/{}{}_data/{}{}.xxx".

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an xart script for translating *.xxx-articulatory vector files into the area functions.

– `record(self, outputloc="Syntheses/", t0=2000, pause=40, iterstep=10, vocal=True, coartmode="COS", speechrate="Normal", xarticul=True, visualise=True)`

Records the constructed utterance according to the chosen coarticulation mode coartmode.

Input:

`outputloc`: string, the path to a folder where to put the results. In `outputloc`, a folder called "{utterance}_{mode}_data/" is going to be created (if it does not exist).

`t0`: integer, the moment when utterance production should begin, in ms. By default, `t0` is 2000 [ms].

`pause`: integer, number of milliseconds for a pause between syntagms. By default, `pause` is 40 [ms].

`iterstep`: integer. A new articulatory vector is formed at least every `iterstep` ms. By default, `iterstep` is 10 [ms].

`vocal`: boolean. Whether to print the `status` of the function when it finishes. By default, yes (True).

`coartmode`: string, the mode of the coarticulation model. Expected values:

* "LIN": linear transition between target vectors;
* "COS": cosine (smoother) transition between target vectors;
* "COMPLEX": cosine transition with finer operation of articulators.

By default, it is "COS". Currently it is a dummy argument: the explored time moments are the same for any coarticulation mode.

`speechrate`: string, a dummy argument for regulating speech rate. Currently, all speech is synthesised at a normal rate.

`xarticul`: boolean, whether the file names should obey Xarticul's conventions and whether to create xarticul scripts for the utterance. By default, yes (True).

Output:

`status`: string describing what has been done if vocal is False; otherwise None.

Writes an xart script for translating *.xxx-articulatory vector files into the area functions.

Inside the utterance-specific folder in `outputloc`, we create:

* "{utterance}_{mode}.xxx": list of articulatory vectors corresponding to the utterance. A single target vector can be obtained as `ph.artv[context]`, where `ph` in `syll.constituents`, where `syll` in `synt.sylls`, and `context` in `contexts`, and `contexts` in `synt.anticipations`, and `synt` in `self.synts`. These target articulatory vectors are to be manipulated according to coartmode, resulting in a sequence of vectors to be stored in an xxx-file.
* "{utterance}_{mode}.af": list of the corresponding area function files. Currently, the area function files, called AF00...{ID}, are produced by Xarticul.
* "{utterance}_{mode}.ag0": operating glottal pressure over time for producing the utterance.
* "{utterance}_{mode}.agp": operating vocal folds oscillations over time for producing the utterance.
* "{utterance}_{mode}.f0": operating the intonation contour.

- Function `fetch_phoneme_by_name(phonemelist, name)`

  From a given list of `Phonemes`, fetches the one (the first one) with the given name.

  Input:

  `phonemelist`: list of `Phonemes`,

  `name`: string.

Output:

`Phoneme`.

- Function `scan_directory_for_phonemes(folderpath= "Data/Speech-Synthesis-Database/DB-Full/", extension=".artv", sort=True)`

  Scans for *.`extension` files in `folderpath` and produces a list of `Phonemes` in the required form that is determined by sort.

  Input:

  `folderpath`: string, `extension`: string - the location of the files to look for and their expected type. They can be set to "Data/Speech-Synthesis-Database/DB-Full/" and ".artv".

  `sort`: boolean: whether to sort them into consonants, vowels and silence. By default, yes (True).

  Output:

  list of `Phonemes` if sort is False, three-element tuple of two lists of **Phoneme** and one `Phoneme` (or None in place of that) if sort is True.

- Function `slice(phonemes, keys, articulators)`

  Slices a set of articulatory vectors, returning only those parameters that are related to a particular articulator or a particular set of articulators.

  Input:

  `phonemes`: list of `Phonemes`.

  `keys`: list of names that the elements from `phonemes` anticipate.

  `phoneme.artv[key]` will result in an articulatory vector that the program will deal with.

  `articulators`: string or list of strings. Expected arguments: see the description of the constructor in the `VocalTract` class.

  Output:

  list of floats.

- Function `database_expand(inputfolderpath= "Data/Speech-Synthesis-Database/DB-Before-Expansion/", outputfolderpath="Data/Speech-Synthesis-Database/DB-Full/", inputextension=".dat", outputextension=".artv")`

  Adds extra samples to the database, estimating them from the corner vowels.

  Input:

  `inputfolderpath`: string is the name of the folder with the vectors. By default, "Data/Speech-Synthesis-Database/DB-Before-Expansion/".

`outputfolderpath`: string is the name of the folder where the system has to put the expanded version of the database. This folder does not have to exist before the start of the program. By default, "Data/Speech-Synthesis-Database/DB-Full/".

`inputextension`: string is used as the identifier of which files to use. See that this ending does not coincide with the ones of the other files that the user may keep in the same folder. By default, `inputextension` is set as ".dat"

`outputextension`: string is the extension for the output files containing the expanded database entries. By default, `outputextension` is set as ".artv" Such a file may be opened in a text editor.

Output:

None.

Creates `outputfolderpath` if it does not exist. Writes an expanded version of the database there, with the file extension `outputextension`.

- Function `compare_articulatory_vectors(real, estimated)`

  Compares two articulatory vectors. A function that is used to evaluate the similitude of the real sample with an estimated one.

  Input:

  `real`, `estimated`: lists of float

  Output:

  `totalerr`: float

  `articulerr`: list of floats that are the contribution of a particular articulator to the difference between the vectors, in agreement with the articulators as defined by class **VocalTract**, constructed in alphabetical order, over the full articulator names.

- Function `evaluate_projections(inputfolderpath="Data/Speech-Synthesis-Database/DB-Before-Expansion/", outputfolderpath="Data/Projections/Evaluation-Of-Corner-Vowels-Assumption/Visualisations & Reports/", inputextension=".dat", outputextension=".evpr")`

  Compares, when possible, the estimated samples from the ones that picture anticipation of corner vowels to the real ones.

  Input:

  `inputfolderpath`: string is the name of the folder with the vectors. By default, "Data/Speech-Synthesis-Database/DB-Before-Expansion/".

  `outputfolderpath`: string is the name of the folder where the system has to put the estimations. By default, "Data/Projections/Evaluation-Of-Corner-Vowels-Assumption/Visualisations & Reports/".

`inputextension`: string is used as the identifier of which files to use. See that this ending does not coincide with the ones of the other files that the user may keep in the same folder. By default, `inputextension` is set as ".dat".

`outputextension`: string is the extension for the output files containing evaluation of the corner vowel approach. By default, `outputextension` is set as ".evpr". Such a file may be opened in a text editor.

Output:

None

Creates `outputfolderpath` if it does not exist. Writes the evaluation files there, with the given file extension.

Bibliography

Anderson, P., Harandi, N. M., Moisik, S., Stavness, I., and Fels, S. 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.

Anderson, S. R. The analysis of French schwa: or, how to get something for nothing. *Language*, pages 534–573, 1982.

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., and Savariaux, C. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553, 2002.

Bell-Berti, F. and Harris, K. S. Temporal patterns of coarticulation: Lip rounding. *The Journal of the Acoustical Society of America*, 71(2):449–454, 1982.

Benguerel, A.-P. and Cowan, H. A. Coarticulation of upper lip protrusion in French. *Phonetica*, 30(1):41–55, 1974.

Bigi, B., Meunier, C., Nesterenko, I., and Bertrand, R. Automatic detection of syllable boundaries in spontaneous speech. In *LREC*, 2010.

Birkholz, P. and Jackel, D. 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.

Birkholz, P. Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In *INTERSPEECH*, pages 2865–2868, 2007.

Birkholz, P. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

Birkholz, P., Jackèl, D., and Kröger, B. J. 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.

Brenner, D. J. and Hall, E. J. Computed tomography—an increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007.

Browman, C. P. and Goldstein, L. Articulatory phonology: an overview. Status report on speech research, Haskins Laboratory, 1992.

Burdumy, M., Traser, L., Richter, B., Echternach, M., Korvink, J. G., Hennig, J., and Zaitsev, M. Acceleration of MRI of the vocal tract provides additional insight into articulator modifications. *Journal of Magnetic Resonance Imaging*, 42(4):925–935, 2015.

Byrd, D., Tobin, S., Bresch, E., and Narayanan, S. Timing effects of syllable structure and stress on nasals: a real-time MRI examination. *Journal of Phonetics*, 37(1): 97–110, 2009.

Calliope. *La parole et son traitement automatique*. Masson, Paris, 1989a.

Calliope. Description acoustique. In *La parole et son traitement automatique*, chapter 3. Masson, Paris, 1989b.

Chomsky, N. and Halle, M. *The sound pattern of English.* ERIC, 1968.

Daniloff, R. and Moll, K. Coarticulation of lip rounding. *Journal of Speech, Language, and Hearing Research*, 11(4):707–721, 1968. doi: 10.1044/jshr.1104.707. URL `http://dx.doi.org/10.1044/jshr.1104.707`.

Echternach, M., Sundberg, J., Arndt, S., Markl, M., Schumacher, M., and Richter, B. Vocal tract in female registers—a dynamic real-time MRI study. *Journal of Voice*, 24(2):133–139, 2010.

Elie, B. and Laprie, Y. Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of vocal folds with glottal chink. working paper or preprint, September 2015. URL `https://hal.archives-ouvertes.fr/hal-01199792`.

Elie, B. and Laprie, Y. A glottal chink model for the synthesis of voiced fricatives. In *Accepted to ICASSP 2016*, Shengai, China, March 2016.

Erath, B. D., Peterson, S. D., Zañartu, M., Wodicka, G. R., and Plesniak, M. W. 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.

Fant, G. *Acoustic Theory of Speech Production.* The Hague: Mouton & Co., 1960.

Fant, G. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971a.

Fant, G. The F-patterns of compound tube resonators and horns. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, 1971b.

Farnetani, E. Labial coarticulation. In Hardcastle, W. J. and Hewlett, N., editors, *In Coarticulation: Theory, data and techniques*, chapter 8. Cambridge university press, Cambridge, 1999.

Flanagan, J. L. *Speech analysis, synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

Flanagan, J. L. and Landgraf, L. L. Self-oscillating source for vocal-tract synthesizers. *Audio and Electroacoustics, IEEE Transactions on*, 16(1):57–64, 1968.

Flory, Y. *The impact of head and body postures on the acoustic speech signal.* PhD thesis, University of Cambridge, 2015.

Fouché, P. *Traité de prononciation française.* Klincksieck, Paris, 1959.

Fougeron, C. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of phonetics*, 29(2):109–135, 2001.

Fougeron, C. and Delais-Roussarie, E. Liaisons et enchaînements: «fais_en á fez_en parlant». *Actes des Journées d'Etudes sur la Parole*, pages 221–224, 2004.

Fougeron, C. and Steriade, D. Does deletion of French schwa lead to neutralization of lexical distinctions? In *EUROSPEECH*, 1997.

Gérard, J.-M., Wilhelms-Tricarico, R., Perrier, P., and Payan, Y. A 3D dynamical biomechanical tongue model to study speech motor control. *arXiv preprint physics/0606148*, 2006.

Gess, R., Lyche, C., and Meisenburg, T. *Phonological variation in French: Illustrations from three continents*, volume 11. John Benjamins Publishing, 2012.

Gibbon, F. and Nicolaidis, K. Palatography. *Coarticulation: Theory, data and techniques*, pages 229–244, 1999.

Gottfried, T. L. and Beddor, P. S. Perception of temporal and spectral information in French vowels. *Language and Speech*, 31(1):57–75, 1988.

Grammont, M. *Traité de phonétique.* Librairie Delagrave, 1950.

Grevisse, M., Goosse, A., Grevisse, M., and Grevisse, M. *Le bon usage: grammaire langue française.* De Boeck., 2011.

Hardcastle, W. J. *Physiology of speech production: an introduction for speech scientists.* Academic Press, 1976.

Hardcastle, W. Electromyography. *Coarticulation: theory, data and techniques. University Press, Cambridge*, pages 270–283, 1999.

Heinz, J. M. and Stevens, K. N. 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.

Henke, W. Preliminaries to speech synthesis based on an articulatory model. In *Proceedings on Speech Communication and Processing*, 1967.

Howe, M. and McGowan, R. 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.

Hunt, A. J. and Black, A. W. 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.

International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Ishizaka, K. and Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell system technical journal*, 51(6):1233–1268, 1972.

Jakobson, R. Why "mama" and "papa"? *Selected writings*, 1:538–545, 1962.

Jansson, J., Holmberg, A., de Abreu, R. V., Degirmenci, C., Hoffman, J., Karlsson, M., and Abom, M. 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.

Jones, D. *The pronunciation of English*, volume 369. Cambridge University Press, 1956.

Jun, S.-A. and Fougeron, C. A phonological model of French intonation. In *Intonation*, pages 209–242. Springer, 2000.

Keating, P. A. The window model of coarticulation: articulatory evidence. *Papers in laboratory phonology I*, 26:451–470, 1990.

Kent, R. D. and Moll, K. L. Vocal-tract characteristics of the stop cognates. *The Journal of the Acoustical Society of America*, 46(6B):1549–1555, 1969.

Knuuti, J., Saraste, A., Kallio, M., and Minn, H. Is cardiac magnetic resonance imaging causing DNA damage? *European heart journal*, 34(30):2337–2339, 2013.

Koenig, L. L., Fuchs, S., and Lucero, J. C. Effects of consonant manner and vowel height on intraoral pressure and articulatory contact at voicing offset and onset for voiceless obstruents. *The Journal of the Acoustical Society of America*, 129(5): 3233–3244, 2011.

Kozhevnikov, V. and Chistovich, L. *Speech: articulation and perception*. Nauka, 1965.

Krane, M. H. Aeroacoustic production of low-frequency unvoiced speech sounds. *The Journal of the Acoustical Society of America*, 118(1):410–427, 2005.

Kröger, B. J., Schröder, G., and Opgen-Rhein, C. A gesture-based dynamic model describing articulatory movement data. *The Journal of the Acoustical Society of America*, 98(4):1878–1889, 1995.

Kühnert, B. and Nolan, F. The origin of coarticulation. *Coarticulation: Theory, data and techniques*, pages 7–30, 1999.

Ladefoged, P. and Disner, S. F. *Vowels and consonants*. John Wiley & Sons, 2012.

Ladefoged, P. and Johnson, K. *A course in phonetics*. Cengage learning, 2014.

Ladefoged, P. and Maddieson, I. The sounds of the world's languages. *Language*, 74 (2):374–376, 1998.

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

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

Laprie, Y., Vaxelaire, B., and Cadot, M. 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`.

Laprie, Y., Elie, B., and Tsukanova, A. 2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes. In *International Congress of Phonetic Sciences*, 2015.

Laver, J. *Principles of phonetics*. Cambridge University Press, 1994.

Léon, P. R. and Léon, M. *Introduction á la phonétique corrective á l'usage des professeurs de français á l'étranger*. Hachette, 1970.

Lindblom, B. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.*, 35(11): 1773–1781, November 1963a.

Lindblom, B. *On vowel reduction*. PhD thesis, Uppsala University, 1963b.

Lindblom, B. *Economy of speech gestures*. Springer, 1983.

Lindblom, B. Explaining phonetic variation: Sketch of the H&H theory. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 403–439. Kluwer Academic Publisher, New York, 1990.

Lloyd, J. E., Stavness, I., and Fels, S. 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.

Lonchamp, F. Les sons du français — analyse acoustique descriptive. Cours de phonétique, Institut de Phonétique, Université de Nancy II, 1984.

Maeda, S. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990a.

Maeda, S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marschal, A., editors, *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990b.

Maeda, S. Phonemes as concatenable units: VCV synthesis using a vocal tract synthesizer. In ans M. Pötzold, A. P. S., editor, *Sound Patterns of Connected Speech: Description, Models and Explanation, Proceedings of the symposium held at Kiel University, Arbeitsberichte des Institut für Phonetik und digitale Spachverarbeitung der Universitaet Kiel:31*, pages 145–164, June 1996.

Maeda, S. and Laprie, Y. 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`.

Maragos, P. Fractal signal analysis using mathematical morphology. In Hawkes, P. and Kazan, B., editors, *Advances in Electronics and Electron Physics*, volume 88, chapter 4, pages 199–246. Academic Press, 1994.

McGowan, R. S. An aeroacoustic approach to phonation. *The Journal of the Acoustical Society of America*, 83(2):696–704, 1988.

McGowan, R. S., Jackson, M. T.-T., and Berger, M. A. 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.

Mermelstein, P. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.

Moore, P. and Von Leden, H. Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica*, 10(4):205–238, 1958.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., and Goldstein, L. A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America*, 132(6):3980–3989, 2012.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771–1776, 2004.

O'Connor, J. D. *Phonetics*. Penguin Books, 1973.

Öhman, S. Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 39(1):151–168, 1966.

Öhman, S. Numerical model of coarticulation. *J. Acoust. Soc. Am.*, 41:310–320, 1967.

OpenStax College. Organs and structures of the respiratory system. `http://cnx.org/contents/t2sgkCQ-@8/Organs-and-Structures-of-the-R`, 2013.

Pasteris, J. D., Wopenka, B., and Valsami-Jones, E. Bone and tooth mineralization: Why apatite? *Elements*, 4(2):97–104, 2008.

Pelorson, X., Vescovi, C., Castelli, E., Hirschberg, A., Wijnands, A., and Bailliet, H. 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.

Pépiot, E. Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. In *XVèmes Rencontres Jeunes Chercheurs de l'ED 268*, pages à–paraître, 2013.

Perkell, J. S. *Physiology of speech production: Results and implications of a quantitative cineradiographic study.* Number 53 in M.I.T. Press research monographs. MIT Press, 1969.

Ramanarayanan, V., Byrd, D., Goldstein, L., Narayanan, S. S., Kobayashi, T., Hirose, K., and Nakamura, S. Investigating articulatory setting-pauses, ready position, and rest-using real-time MRI. In *INTERSPEECH*, pages 1994–1997. Citeseer, 2010.

Ruty, N., Pelorson, X., Van Hirtum, A., Lopez-Arteaga, I., and Hirschberg, A. 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.

Saltzman, E. L. and Munhall, K. G. A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382, 1989.

Scherer, R. C., Shinwari, D., De Witt, K. J., Zhang, C., Kucinschi, B. R., and Afjeh, A. A. 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.

Schwartz, W. L. Syllabication in French and suggestions for accenting the letter E. *The Modern Language Journal*, 5(7):374–377, 1921.

Serrurier, A. *Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production des nasales.* PhD thesis, Institut National Polytechnique de Grenoble, 2006.

Serrurier, A. and Badin, P. 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.

Singampalli, V. D. and Jackson, P. J. Statistical identification of critical, dependent and redundant articulators. In *Interspeech 2007: 8th annual conference of the International Speech Association, vols.1–4*, pages 2736–2739, 2007.

Soquet, A., Lecuit, V., Metens, T., and Demolin, D. 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.

Steinberg, D. D., Nagata, H., and Aline, D. P. *Psycholinguistics: Language, mind and world*. Routledge, 2013.

Story, B. H. Vowel and consonant contributions to vocal tract shape. *The Journal of the Acoustical Society of America*, 126(2):825–836, 2009.

Story, B. H. Phrase-level speech simulation with an airway modulation model of speech production. *Computer speech & language*, 27(4):989–1010, 2013.

Teager, H. M. and Teager, S. M. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., and Shosted, R. Real-time MRI for Portuguese. In *Computational Processing of the Portuguese Language*, pages 306–317. Springer, 2012.

Thomas, T. A finite element model of fluid flow in the vocal tract. *Computer Speech & Language*, 1(2):131–151, 1986.

Titze, I. R. The human vocal cords: a mathematical model. *Phonetica*, 28(3-4): 129–170, 1973.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013. doi: 10.1109/JPROC.2013.2251852. URL `http://dx.doi.org/10.1109/JPROC.2013.2251852`.

Tranel, B. *The sounds of French: An introduction*. Cambridge university press, 1987.

Ventura, S. R., Freitas, D., and Tavares, J. M. R. Application of MRI and biomedical engineering in speech production study. *Computer methods in biomechanics and biomedical engineering*, 12(6):671–681, 2009.

Vorperian, H. K., Kurtzweil, S. L., Fourakis, M., Kent, R. D., Tillman, K. K., and Austin, D. Effect of body position on vocal tract acoustics: Acoustic pharyngometry and vowel formants. *The Journal of the Acoustical Society of America*, 138(2): 833–845, 2015.

Zemlin, W. R. *Speech and Hearing Science, Anatomy and Physiology*. Pearson Education (US), Fourth Edition edition, 2010.