AALBORG UNIVERSITY
DENMARK

Model-based Analysis and Synthesis of Aging Effects on Human Voice Production

Master Thesis

August 5, 2020

Author:
Christie Laurent

Supervisor:
Cumhur Erkut

Aalborg University

Department of Architecture, Design and Media Technology

Faculty of IT and Design

Abstract

This document deals with voice synthesis techniques based on a combination of physical modelling and analytical elements, and serving as support for the design of an ageing voice model. The side contribution of this project is an adapted and complementary multiple-fold analysis tool that was developed in parallel. The aim of the whole project is to create a parametric ageing voice model where age becomes a tunable parameter, and this paper is meant to be its theoretical ground. Stated as above, this project digs into a barely explored branch of the voice synthesis field, even though voice synthesis is currently on-trend: numerous applications indeed exist nowadays, but very few consider age as a tunable parameter [Story et al., 2018, Schotz, 2006]. In a closely related branch, several have been trying to manipulate an existing voice to make it sound older or younger [Skoog Waller and Eriksson, 2016, Rupal and Seth, 2017]. For both examples, a certain knowledge about voice ageing is necessary; unfortunately, this is a complex phenomenon not yet fully understood at the current time. This document therefore gathers knowledge, theories and applications about the voice: it presents its production process and its characterisation, including its evolution over a lifetime; it addresses the physics and the physiology necessary to explain the previous elements and the different computing techniques employed to model it before introducing the ageing model that was developed. The fixed-age voice model (FAM) and the ageing voice model are finally evaluated in terms of credibility and quality.

Key-words: voice, ageing, physical modelling, source-filter, age, sound processing, voice synthesis, voice analysis

Contents

Chapter 1

Introduction

The voice is a typical human feature and has therefore naturally raised interest among audio computing researchers, especially since modern imagery techniques have improved and computational capacities have increased. Voice synthesis is nowadays an active area, as much as for AI design, the film industry or pure research purposes.

More specifically here, the focus is put on the synthesis of an ageing voice, and this paper aims at guiding the reader in the process of conception of a parametric voice synthesis model, where age is a tunable parameter. Stated as above, we are investigating a barely explored branch of the voice synthesis field: indeed, very few models consider age as a hyper-parameter. Such a project could raise interest in the video-game and film industries: voices of fictional characters could be synthesised without requiring actual speakers; or in medicine: to predict voice evolution in patients or help in diagnosis without resorting to heavy and costly devices such as for Magnetic Resonance Imaging (MRI). By providing the theoretical ground of this multidisciplinary project that combines voice analysis, the voice production system morphology, basic physical modelling of voice, computing synthesis techniques and physiological evolution, this paper aims at answering the following question. To what extent can an ageing voice be modelled starting with a simple physical model and without in-depth knowledge about the ageing phenomenon? Can this ageing effect be extended to natural voices?

First in Ch. 2, voice generalities, work related to voice synthesis and the underlying theory are introduced; general knowledge about voice over time is also presented. Then in Ch. 3, the guiding principles about this project are disclosed. Following in Ch. 4, the designing process of the ageing model that was developed is presented. The evaluation tools are then outlined in Ch. 5 before the actual results are exposed in Ch. 6 and discussed in Ch. 7. This project is finally contemplated in terms of achievements, possible improvements and perspectives in Ch. 8.

Chapter 2

Background / Related Work

This chapter presents the various notions, theory and equations necessary to the conception of a voice synthesiser where age is an additional parameter.

2.1 Generalities about the Voice

A *voice* is generally defined as the sound produced in a person's larynx and uttered through the mouth, as speech or song [Mithen,]. It differs from a *sound* by the meaning it conveys; however, it inherits number of its characteristics. This section addresses the main knowledge related to the voice: its characteristics, creation process, different types, and its evolution in time.

2.1.1 Voice Production System

Voice is produced by the *voice production system* (VPS) that can be decomposed into three main parts: the pressure source, *glottis*, the *cavities*, and the *extremities*, referred to as 1, 2, 3 and 4 respectively on Fig. 2.1. Note that the glottis is also an extremity.

Lungs

The lungs provide the main source of energy: air. Their outgoing pressure averages 785 Pa, that would be equivalent to 152 dB in loudness. This pressure is altered through the action of the other VPS components until it reaches the lips, where it can be understood as the strength of the voice - or speech *loudness*. The lungs are therefore mainly responsible for speech loudness.

Glottis

It is the fluctuating space located between the *vocal folds* (VF) - casually and misleadingly called *vocal cords* - and the arytenoid cartilage of one side of the larynx, and those of the other side. The VF are two tissue folds, not uniform in their structure [Hirano et al., 1975]. From a histological point of view, they are made of five layers, at and near the edge: the epithelial layer, the superficial, intermediate and deep layers of the lamina propria; and the muscle layer. From a physical point of view, the VF have three layers, i.e., the *cover* which consists of the epithelium and the superficial layer of the lamina propria, the *transition* which consists of the intermediate and deep layers of the lamina propria, and the *body* which is the vocalis muscle. The elasticity is the greatest in the body and the smallest in the cover when no laryngeal muscles are activated, which explains its non-symmetrical closing and opening.
Under the combined action of glottis muscles and lungs air pressure and with the Bernoulli

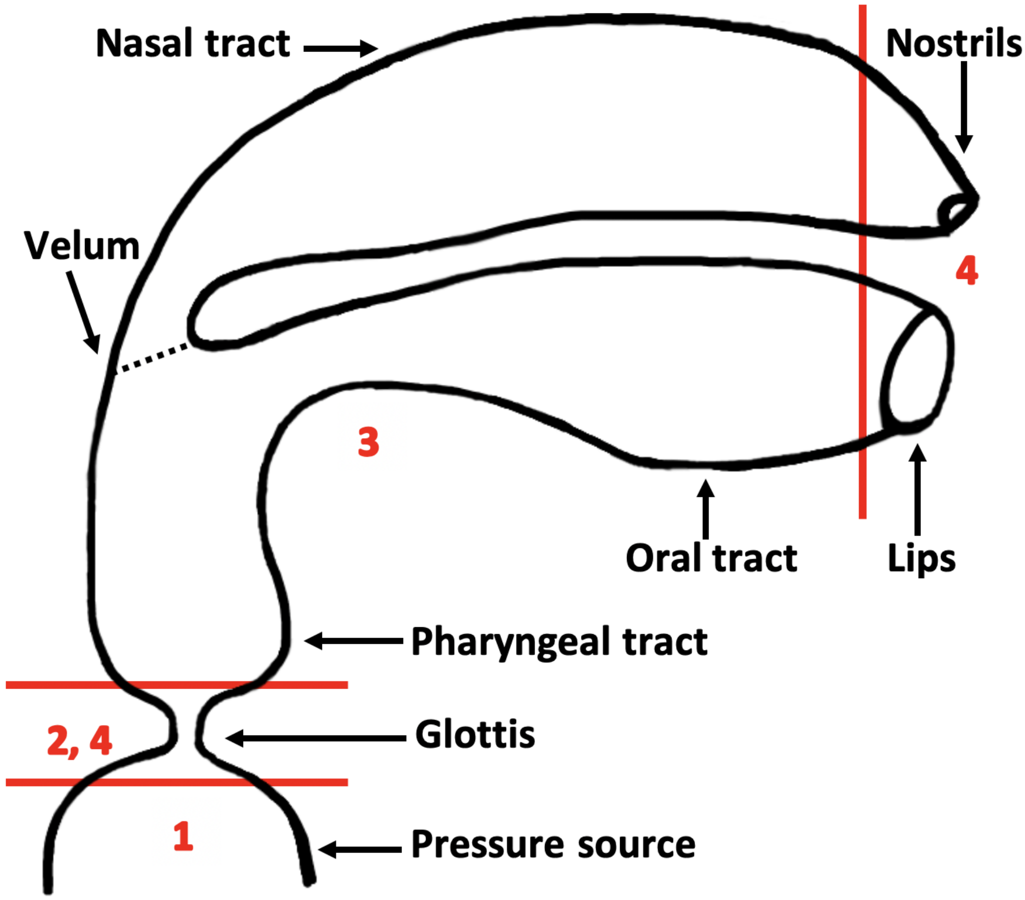

Figure 2.1: Simplified drawing of a vocal tract. **1**: power supply; **2**: excitation; **3**: cavities; **4**: system delimitation.

principle as ground theory, the VF cyclically abduct (due to posterior cricoarytenoid muscles) and adduct (due to lateral cricoarytenoid muscles), causing sudden air rushes and stops through the glottis. This phenomenon, called *phonation*, produces the *excitation signal*.
The glottis is mainly responsible for the fundamental frequency f_0, even though it is altered due to an existing (but passed over in silence here) coupling between the different parts of the VPS.

Cavities

The cavities are the areas in which the glottal signal propagates; they are designed altogether as the *resonator*. The resonator - when modelled - always includes the pharyngeal and oral tracts which are the largest and most impacting cavities on the direct air path from the glottis to the lips (see 2.1); they define the *vocal tract* (VT). Additionally, the resonator may also comprise the nasal tract and/or the trachea - the *trachea* is the part of the air tract below the glottis that goes down to the lungs.
Due to their varying sizes and tissue properties, the cavities possess their own resonant frequencies and induce some progressive decrease in intensity up to the extremities. These transformations relate to the individual's morphology, culture and health, and are therefore unique.

Extremities

This term encompasses the morphological parts of the VPS that communicate with the outside, namely the *lips*, the *nostrils*, and the glottis. Their configurations, open or closed, interact with the sound coming from the resonator and therefore alter the sound that is emitted. They relate to the *boundary conditions*.

Utterances

The sounds produced by a speaker are called *utterances* and can be categorised into two main categories: *vowels* and *consonants*. Utterances - whatever their type - are mostly characterised by the vocal tract shape (VTS), that is, by the shape determined by the cavities and extremities. On the one hand, vowels are determined by the whole profile of the VT. On the other hand, consonants are produced when the speaker locally constricts or totally occludes his VTS, producing different types of sounds depending on the place, shape and size of this bottleneck.

2.1.2 Voice Characterisation

This part addresses the voice under an analytical perspective with the focus on vowels.

Loudness

As every existing sound, the voice is physically characterised by its *sound pressure*, that fluctuates over time and depends on a number of parameters (see Sec. 2.1.1). It is measured by the sound pressure level, notated L and expressed in dB SPL, which is the logarithmic ratio of the effective pressure to the reference value $P_0 = 20\,\mu\text{Pa}$:

$$L = 20\,log_{10}(P/P_0). \tag{2.1}$$

The human ear covers the range [0-120] dB SPL while human normal *speech* level is usually contained in the interval [40-60] dB SPL [Stebbins, 1983, p.6].
Perception-based measures of the sound strength are not considered here; therefore, when we use the term *loudness* further, it is to be understood as a synonym of sound pressure.
A voice is a succession of stationary[1] sounds. Hence, some frequential analysis are particularly relevant and reveals peaks in the spectrum (see Fig.2.2).

Pitch

The lowest peak in frequency - also the highest peak in amplitude in Fig.2.2 - is the fundamental frequency f_0 of the emitted sound. It is highly correlated to the perceived sound *height*, or *pitch*. This cue is typically the main one used to distinguish a man from a woman, or a child from an adult. For instance for French speakers, a male's average f_0 is 133 Hz versus 234 Hz for a female [Pépiot, 2014]. When it is not sufficient to identify the speaker's gender - e.g. because male and female voice frequency ranges can overlap - the timbre comes as a discriminating cue.

Timbre

Letting aside the peak at f_0, all the other peaks in the spectrum participate in the general quality and shape of the sound more specifically.
In the case of a voice, and contrarily to manufactured musical instruments where they are rather thin, these peaks are agglomerated into *formants*, defined as localised concentrations of spectral energy centred on some *frequencies*, with a certain *bandwidth* and *amplitude*. The three aforementioned parameters vary with every individual: they characterise both a person's voice timbre and the nature of the vowel uttered. In the frequency domain, due to their relative harmonicity, vowels are characterised by pronounced formants. Back to our gender

[1] *stationary*: whom statistical properties do not or slowly change over a period of time

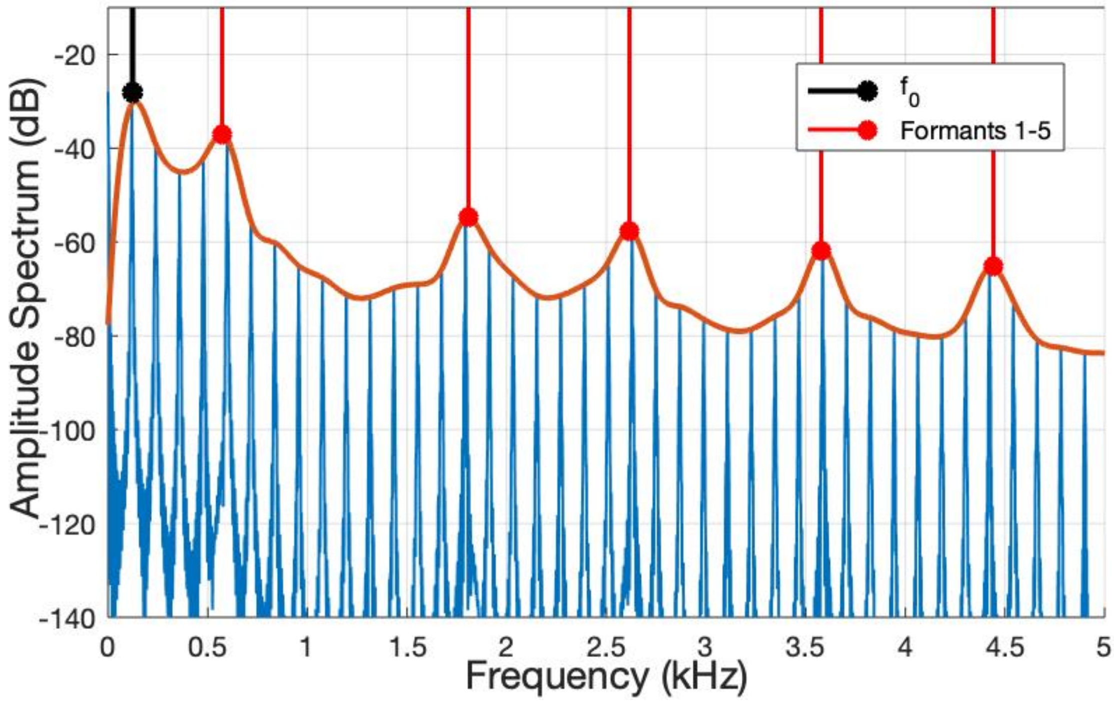

Figure 2.2: Log-spectrum of the utterance /ε/ for $f_0 = 120$ Hz, featuring f_0 and the five first formants.

discrimination problem, a female voice can be identified thanks to the relative higher 2^{nd} and 3^{rd} formants locations in comparison with a male of same f_0 [Peterson and Barney, 1952]; the sexual dimorphism in voice is indeed one of the largest observed in physical measurements of humans [Kent and Vorperian, 2018]. According to [Schotz, 2006], this differentiation between genders is particularly marked for voice intensity and speech rate.

2.2 Voice Modelling - Review

The VPS has been modelled with several techniques since it was born, among which analysis-synthesis methods and physical modelling methods.

2.2.1 Definitions

The source-filter theory [Fant, 1960] illustrated on Fig.2.3 consists in breaking the whole process of voicing into two elements: the excitation (source) and the resonator (filter). Modelling physically means to consider a source and several filters that are geometrically defined. A physical-inspired model is a hybrid model that combines physically modelled elements and parts modelled with other methods, e.g. analysis-synthesis.

2.2.2 Source

A first aspect of research in this field is the modelling of the voice source, i.e. the excitation signal at the glottis scaled by the pressure released by the lungs. The amplitude of a speech signal varies with time. Especially at the scale of a respiration, the temporal envelope of a speech signal corresponds to the energy progression in time from silence to speech them silence again. As such, it relates to the lungs pressure - guided by the intention of the speaker. An

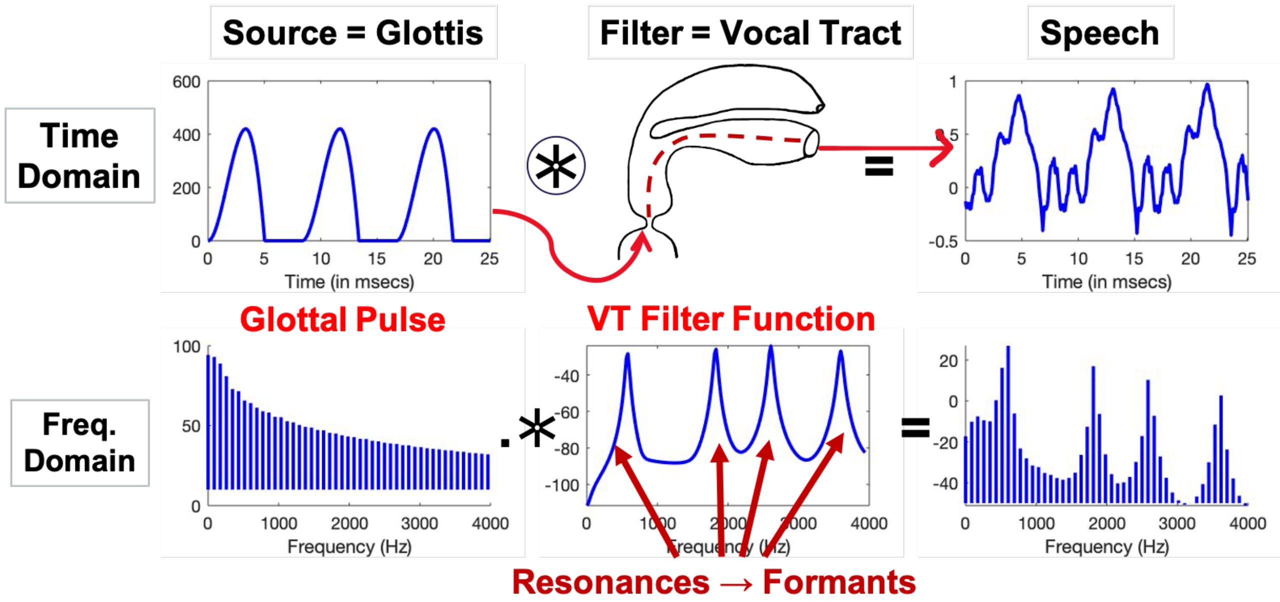

Figure 2.3: Source-filter theory: the voice (signal on the right) can be modelled through the decoupling of the excitation (left) and the resonator (middle). The upper and lower rows represent the time- and frequency-domain processes.

envelope in general can be divided into three parts: a stationary section flanked by the two transitory parts, the fading-in and the fading-out parts.

Originally for excitation modelling, broadband noise was used - this is the principle of the Vocoder [Dudley, 1964]. However, the excitation signal was made more complex in order to represent better the reality. From a physical model point of view, the double string-mass system was developed by [Ishizaka and Flanagan, 1972]; it was refined with a third mass in [Story and Titze, 1995].
From a sound processing point of view, several analytical functions modelling the glottal pulse (GP in text, g in equations) have been created since 1971 [Fant et al., 1985]. More precisely, these expressions define one *glottal pulse cycle* (GPC). The GP can be obtained digitally by concatenating multiple of these GPC or by looping on one of them. The function illustrated on Fig. 2.4 is the GPC defined by Fant (1979), referred to as (r.t.a.) *gp-fant79*. Other

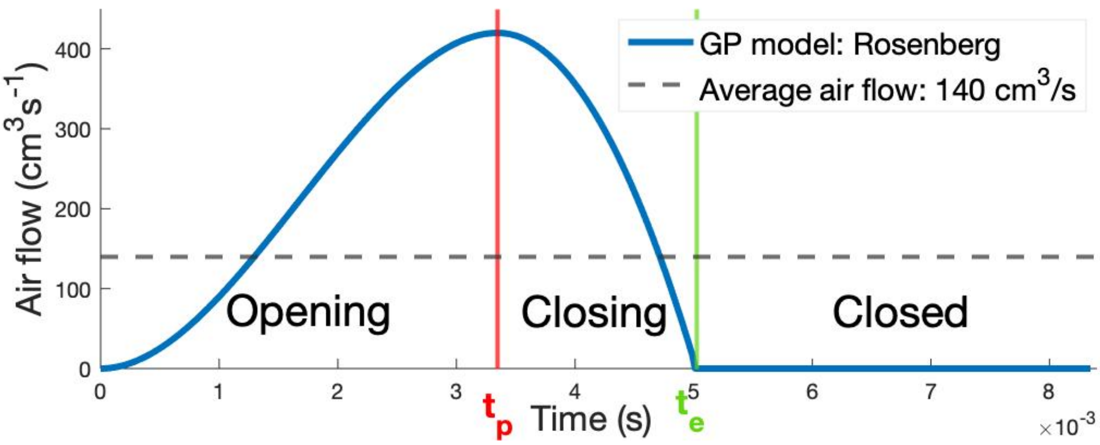

Figure 2.4: Glottal pulse over one period T for $f_0 = 136$ Hz, $f_s = 44100$ Hz, $tp = t_{max(gp)} = .40\,T$ (in red), $te = t_{gp=0,t>tp} = 0.44\,T$ (in green) - Fant model (1979)

models exist, such as the Rosenberg-B model (Rosenberg, 1971, r.t.a. *gp-rosenB*) and the

KLGLOTT88 model (Klatt and Klatt, 1990, r.t.a. *gp-klglott88*).These models differ from the slope of the opening and closing periods, as well as from the *opened-to-closed ratio* (OCR) that quantifies the proportion of time the glottis is open against the duration it is not. In order to model better the voice, noise and turbulences can be introduced at the glottis [Story, 2013].
Another method consists in modelling physically the excitation as a N-mass-spring model with N in $\{2, 3\}$, as in [Story, 2013]. In this case, is the GP area which is modelled rather than the glottal flow - of course both are interdependent.
Finally, some authors try to model the GP as precisely as possible in using inverse techniques: given a signal recorded at the lips e.g., apply the inverse filter of the VT to obtain the "true" GP - e.g. [Alku et al., 2006].

Whereas the previously exposed method is analytic, the N-mass model calls to physical modelling. It models the glottis as a self-oscillating source composed of N stiffness-coupled masses. N is equal to 2 or 3, in order to roughly model the conic shape of the closing-opening glottis - the large part being on the sub-glottis half [Ishizaka and Flanagan, 1972]. This conic shape which causes the velocity to be greater at that level. This is called the *vertical phase difference*. When N=3, the third mass is used to simulate the *body component's effect* [Story and Titze, 1995] in the glottis body-cover structure defined by [Hirano, 1974, Hirano et al., 1975]. This is meant to represent the transverse movement of the VF.
In 1982, the model is made more complex with the integration of the wave equation using the rectangular method in space, and the trapezoidal method in time [Maeda, 1982]. In 1985, Lijiencrantz experiments on an undersampled acoustic tube model in order to modify the VTS with ensuring energy conservation [Fant et al., 1985]. In 1992, Rene Carré derives a model from sensitivity analysis, based on distinctive regions: he had noted that movements in particular regions of the vocal tract affected formant frequencies more than in others [Carré et al., 1992].

2.2.3 Filters

Kelly and Lochbaum were the first ones to propose a physical model of the vocal tract in 1962 [Kelly and Lochbaum, 1962], along with scattering junctions for connecting the segmented tubes. Thanks to the extension of computational capacities, this model is being more discretised today to represent the real human anatomy with more accuracy. Some work considers interpolated fractional samples and truncated conical tube segments [Välimäki and Karjalainen, 1994]; other account for the nasal tract on top of the main oral tract, for radiation through the throat wall or real-time control [Cook, 1996], or for tissue impact to simulate energy losses [Titze and Alipour, 2006]. The shape thus designed controls the wave propagation and the two first formants of the vowel emitted. Extensions in two and three dimensions such as [Speed et al., 2013, Zhang, 2016] allow to take into account anatomic asymmetry and to model higher formants.
However, some knowledge about the VTS is needed to ensure that these models conform to reality at some minimum level. Magnetic Resonance Imaging (MRI) has been utilised to acquire such information for different vowels [Story et al., 1996]. Extracting the main features from these vowels, e.g. by Principal Component Analysis, can provide parametric models of the VT and allow the interpolation of its shape between different classified vowels [Story, 2005].

2.3 Theory

In this chapter, following on the voice synthesis methods mentioned here above, we develop the underlying theories and write down the necessary equations for air propagation, air excitation, absorption, dissipation and other distorting phenomena.

2.3.1 Source

Let's recall that a GP describes the opening and closing phases of the glottis, that are caused by the mechanic periodic abduction and adduction of the glottis muscles and that are scaled by the pressure released by the lungs at time t.

Time Envelope

An envelope can be modelled more or less artificially. Three methods are here presented: artificial, reality-based and hybrid. The first one - the most artificial one models the utterance transitory parts with a square sinus. It is defined as in Eq. 2.2 over angles in $[0, pi/2]$ rad and $[pi/2, pi]$ rad for fading-in and fading-out respectively, and parameterised with two different durations t_{fi} and $t_{utt} - t_{fo}$ respectively, where t_{utt} is the duration of the utterance.

$$f(t) = sin(\theta_t)^2 \quad \text{with} \begin{cases} \theta_t = [0 : dt : pi/2] & \text{if} \quad 0 < t \leq t_{fi} \\ \theta_t = pi/2 & \text{if} \quad t_{fi} < t < t_{fo} \\ \theta_t = [pi/2 : dt : pi] & \text{if} \quad t_{fo} \leq t \leq t_{utt} \end{cases} \tag{2.2}$$

The second type of envelope used can be extracted from *real* speech utterances. Given a small database of vowel series pronounced by different persons, the envelopes are extracted by detection of the minima in intensity over the low-pass filtered (1900-tap Hilbert filter) "sentence"; after redressing (so that the first and the last samples are zero), they are normalised and standardised in duration for being freely adaptable.

The third method is based on the modelling of real envelopes. In this case, the three envelope parts are identified and extracted from the real speech signal as in the previous method; then, they are modelled - for instance using the Matlab function "*polyfit*". Continuity between the three sections must be enforced - e.g. by using the Matlab interpolation function "*interp1*" (interpolation method: spline). These intermediary sections are called *patch* and are parameterised by their duration (they overlap and replace the adjacent initial sections). Finally, redressing the signal ensures that the first and last samples are 0.

Analytical functions

We develop now on analytical functions. Despite being similar, all analytical GP methods differ in literal expressions and parameterisation. Especially, the time of maximum opening (t_p) is the most vital parameter (indicated by a vertical red line on Fig. 2.4). As an example, one can take $t_p = .4T = 0.4\frac{1}{f_0}$. The time of first closing after t_p, referred to as t_e (indicated by a vertical green line on Fig. 2.4), is the instant at which the GP reaches its maximum amplitude. The only constraint about t_e is to be located *after* t_p in time. It can be further controlled in using the OCR defined such as

$$OCR = \frac{pte}{1 - pte} \in [0, 1]. \tag{2.3}$$

In all cases, a gain G_t is applied to relate the function to reality. In [Sulter and Wit, 1996], G_t is defined based on the average glottal flow which is 140 $\text{cm}^3.s^{-1}$.
The literal expression for method gp-klglott88 follows. Note that here t_e and t_p are linked:

$$g_K(t) = \begin{cases} G_t\left[\left(\frac{t}{t_p}\right)^2\left(3 - 2\frac{t}{t_p}\right)\right] & \text{if} \quad 0 \leq t \leq t_e = \frac{3}{2}t_p \\ 0 & \text{if} \qquad t_e < t < T \end{cases} \tag{2.4}$$

It is then low-pass filtered. In method gp-rosenB, t_e is user-defined:

$$g_R(t) = \begin{cases} G_t \left[\left(\frac{t}{t_p}\right)^2 \left(3 - 2\frac{t}{t_p}\right) \right] & \text{if} \quad 0 \leq t \leq t_p \\ G_t \left[1 - \left(\frac{t-t_p}{t_e-t_p}\right)^2 \right] & \text{if} \quad t_p \leq t \leq t_e \\ 0 & \text{if} \quad t_e < t < T \end{cases} \tag{2.5}$$

For method gp-fant79 finally, an additional parameter K is used to tune the slope of the function's closing part:

$$g_F(t) = \begin{cases} G_t \, \frac{1}{2} \left[1 - \cos(\pi \frac{t}{t_p}) \right] & \text{if} \quad 0 \leq t \leq t_p \\ G_t \left[K \cos\left(\pi \frac{t-t_p}{t_p}\right) - K + 1 \right] & \text{if} \quad t_p \leq t \leq t_e = t_p \left(1 + \frac{1}{\pi} \arccos \frac{K-1}{K} \right) \\ 0 & \text{if} \quad t_e < t < T \end{cases} \tag{2.6}$$

Some other analytical functions are complementary defined with their derivative and additional parameters.

Glottal Noise

The produced vocal flow rapidly alternates periods of non-flow and periods of maximal flow. This may cause perturbations - turbulence - in the air flow direction under certain conditions related to the dimensions of the duct neck - here the glottis - and to the properties of the environment - here the air. An index exists that measures this flow characteristic: Reynolds number *Re*. This index is dimensionless and can adapt to any fluid flow by considering the right characteristic length L_c. In our case:

$$Re = \frac{u\,\rho}{\eta\, L_c} \tag{2.7}$$

where u is the instant flow, ρ is the air volumic mass, η is the air dynamic viscosity, and $L_c = L_{VF}$ is the VF length [Samlan and Story, 2011]. A low value of *Re* will characterise a *laminar* flow - without noise, while a high value will characterise a *turbulent* flow - with additional noise. Depending on the application, the value of the threshold Re_c between both state changes. For voice production, $R_e \approx 1200$ [Samlan and Story, 2011].
In the case of a turbulent flow, a noise component shall be calculated. Referring to [Fant, 1960], it is generated in the following form:

$$U_{nois} = \begin{cases} N_f (Re^2 - Re_c^2)\, G_s & \text{if} \quad Re > Re_c \\ 0 & \text{if} \quad Re \leq Re_c \end{cases} \tag{2.8}$$

where N_f is a broadband 0-centered noise signal that has been band-pass filtered between 300–3000 Hz (2^{nd} order Butterworth) and $G_s = 4.10^{-6}$ a scaling factor set as in [Samlan and Story, 2011].

2.3.2 Resonator

Once the excitation signal produced at the glottis, it propagates through the VT until it reaches the extremities. On the way, the signal is shaped in both time- and frequency-domain.

VT modelling

Considering that the sound wavelength is great compared with the cross-dimensions of the VT and therefore assuming a one-dimensional wave propagation along the VT and under observation of adequate matching conditions, a stack of N straightened uniform cylinders based on the VT cross-sectional area functions A can be used to solve the simplified wave equation, as illustrated on Fig. 2.5. N is typically fixed according to the following formula:

$$N = \frac{L_{VT}}{dx} \text{ with } dx = \frac{c}{2\, f_s} \tag{2.9}$$

where $f_s = 44100$ Hz is the sampling frequency, $c = 350\, m.s^{-1}$ is the sound velocity (higher than the usual $340\, m.s^{-1}$ due to air temperature in a human body), $L_{VT} \approx 17$ cm is the average VT length for a male individual. In practice, $N = 44$ according to these parameters' values.

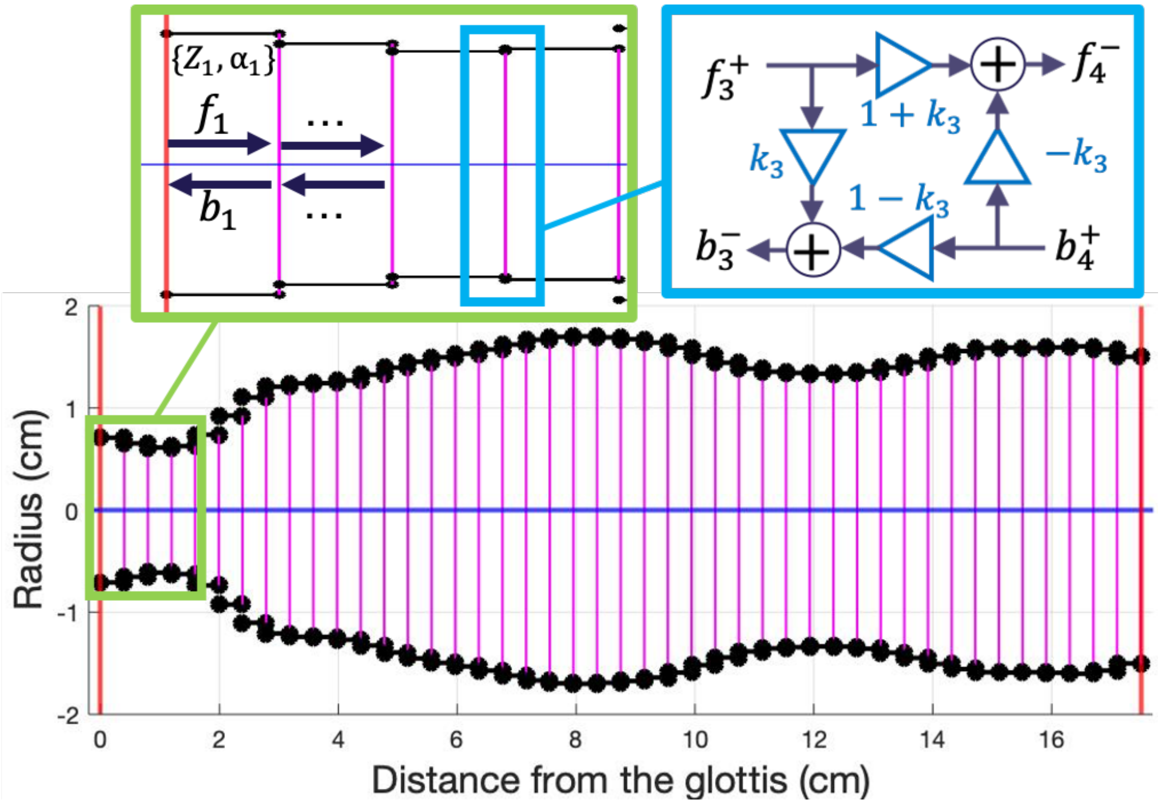

Figure 2.5: Discretised vocal tract obtained for vowel /ɛ/. Bi-directional delay-line are observable in the green quadrant in segment 1; scattering junctions for segmented tubes are illustrated in blue quadrant between segments 3 and 4. f_i and b_i are the forward and backward acoustical pressures for segment i, respectively. Notations Z, α and k correspond to impedance, attenuation, and reflection, in this order.

Propagation

The voice is an acoustic wave, meaning that is depends on both time t and space x when propagating.Under the hypotheses presented here above, it can be approximated in one dimension as follows:

$$u_{tt}(t, x) = c^2\, u_{xx}(t, x) \tag{2.10}$$

where u, t and x represent the acoustic displacement, the time variable and a single space variable, respectively, and the subscripts indicate the 2^{nd} order discretisation operator.This is the 1D-waveguide equation expressed for u.
According to general theory for wave propagation, the solution to the 1D-waveguide equation can be modelled by two time- and space-dependent bi-directional waves propagating forward and backward (resp., f and b).This solution is illustrated on Fig. 2.5 (top left quadrant) and is of the form:

$$u(t, x) = f_i(x - ct) + b_i(x + ct) \tag{2.11}$$

Impedance

In acoustic tubes, acoustic pressure p is related to flow velocity u through *acoustical impedance* Z such as:

$$Z = \frac{p}{A\,u} = \frac{\rho\,c}{A}, \tag{2.12}$$

with ρ the volumic mass - typically 1.147 kg.m^{-3} at 35 deg Celsius. The notion of impedance quantifies the effect - p - of a phenomenon relatively to its cause - u. Conservation laws tell us that at any point *inside* a tube,

$$\begin{cases} p^+ = Z\,u^+ \\ p^- = -Z\,u^- \end{cases} \quad \text{and} \quad u = u^- + u^+, \tag{2.13}$$

where the superscripts $+$ and $-$ indicate the sense of wave propagation. After discretisation, Eq. 2.13 becomes

$$\begin{cases} p_i^+ = Z_i\,u_i^+ \\ p_i^- = -Z_i\,u_i^- \end{cases} \quad \text{and} \quad u_i = u_i^- + u_i^+, \tag{2.14}$$

where u_i and p_i are displacements and acoustic pressures in segment $i \in \{1, ..., N\}$. In the system considered here, Z_i is constant in every segment, but may differ from one segment to another.

Junctions in Line

In order to account for the sudden change of cross-sectional area between segments, *scattering junctions* are introduced (see Sec. 2.3.2). Such a junction is observable in Fig. 2.5 (top right quadrant). In the following, all expressions are expressed for the junction between cylinders i and $i+1$, referred to as junction i; however, the index i is unmarked to prevent mixing with intersection indexes. Furthermore, note that we now use the superscripts $+$ and $-$ as indicators of incoming or outgoing wave at the junction, respectively.
Due to continuity laws at the junctions, for a junction i and its J intersections:

$$\forall j \in \{1, ...J\}, \quad u_j = u_J \quad \text{and} \quad \sum_j^J p_j = 0 \tag{2.15}$$

with u_J a constant. After injecting equations 2.13 into Eq. 2.15:

$$u_J = 2\,\frac{\sum_j^J Z_j\,u_j^+}{\sum_j^J Z_j} \quad \text{and} \quad u_j^- = u_J - u_j^+. \tag{2.16}$$

Knowing the displacement at the junctions, the f_is and b_is can now be determined. These segments being of various sections, and based on displacement continuity at the junctions, part of the acoustic energy propagating is transmitted to neighbouring segments while the rest is retained. The *reflection coefficient* k measures this effect and is expressed at the junction i for two incoming waves as follows:

$$\forall i \in \{1, ..., N-1\}, \quad k_i = \frac{Z_i - Z_{i+1}}{Z_i + Z_{i+1}}. \tag{2.17}$$

This coefficient can also be defined from the A_i since the area is the only varying parameter in Z_i's definition.

Junctions in Parallel

Thanks to this system of junctions, additional ducts for speech sound propagation can be added to the model, such as the nasal tract, or any other cavity impacting the voice. Formulae 2.15 are inverted in this case, meaning that for a junction i and its J intersections:

$$\forall j \in \{1, ...J\}, \quad p_j = p_J \quad \text{and} \quad \sum_j^J u_j = 0 \tag{2.18}$$

with p_J a constant. Following, Eq. 2.16 and 2.17 also change:

$$p_J = 2\,\frac{\sum_j^J Y_j\, p_j^+}{\sum_j^J Y_j} \quad \text{and} \quad p_j^- = p_J - p_j^+, \quad \text{with} \quad Y_j = \frac{1}{Z_j} \quad \text{and} \quad k_i = \frac{Y_i - Y_{i+1}}{Y_i + Y_{i+1}}. \tag{2.19}$$

See in [Bilbao, 2004, Ch.1] for a comprehensive description of the alterations.

Attenuation

Another phenomenon occurring during propagation is acoustical *attenuation*, due to the nature of the tissue constituting the VPS: the contact between the air and the tissues is not frictionless and implies that part of the energy is wasted locally at the level of the boundary layer between the two elements. It is usually named α and can be arbitrarily defined from a *damping* coefficient d usually set in $[0.001, 0.01]$ as follows: $\forall i \in \{1, ...N\}$,

$$\alpha_i = 1 - \frac{1 - d_i}{\sqrt{A_i}}. \tag{2.20}$$

It can also be modelled physically by using aerodynamic physics and knowing the properties of the VT tissues. The Reynolds number impacts this loss too (see Eq. 2.7).
In the case of a constriction inside the VT - not at the extremities, Stevens (reported in [Maeda, 1982]) proposes to model laminar resistance in the concerned segments of the VT such as $R_i = (8\,\pi\mu)/A_i^2$. where μ is the air viscosity. This formula is valid in a circular duct only.

2.3.3 Boundary Conditions

Our system has two types of extremities: the ones overlooking on the outside - lips and nostrils - and the internal one - the glottis. Their nature sets the boundary conditions (BC). The BC imply a certain loss of acoustical energy and restrain global energy to skyrocket to infinity. These BC are basically represented by a radiation impedance (or "radiation load"), characterised by a resistance and inductance.

Lips and Nostrils

Rabiner's acoustic model of lip radiation - mentioned in [Airaksinen et al., 2014] - assumes radiation from an infinite plane baffle. In Digital Signal Processing (DSP), it comes down to a first-order FIR filter alike $R(z) = 1 - \alpha z^{-1}$ where $\alpha \to 1^-$ and whose solution's root is slightly pulled inside the unity circle to guarantee the filter's stability. In practice, values in between [.98, .999] are being used [Airaksinen et al., 2014] even though this author criticises the use of a constant α coefficient for it causes distortion at low frequencies, particularly in the closed phase. The author therefore proposes a method to automatically adjust α to address the problem. Other vibrating piston shapes have been experimented, e.g. spherical, but the difference didn't prove significant [Maeda, 1982].

Glottis

The glottis can be considered as a boundary or as an interface. As a boundary, it is a more or less dissipating element that reflects the remaining of the acoustical energy into the VT. In [Maeda, 1982], the VT is equivalently modelled with an electrical circuit. In this context, the glottis is modelled as the resistance R_{gl} such as:

$$R_{gl} = \frac{12\,\eta\,L_{VF}^2}{A_{gl}^3} T_{VF} + k_c \frac{\rho\,u_{gl}}{2\,A_{gl}^2} \tag{2.21}$$

where A_{gl} is the glottal area, L_{VF} and T_{VF} are respectively the length and the thickness of a rectangular duct representing the glottis, and k_c is a coefficient having a typical value of 1.38. The first term refers to laminar resistance due to air viscosity and the second one accounts for turbulence occurring at the glottis.
As an interface, the glottis connects the trachea and the VT [Story, 2013] through the following formula:

$$A^* = \left(\frac{1}{A_{-1}} - \frac{1}{A_1}\right)^{-1} \tag{2.22}$$

where A^* is the glottis effective VT area for acoustic loading and A_{-1} is the first sub-glottis cross-sectional area in the discretised representation of the trachea. In this case, the lower BC is moved to the bottom of the trachea, and can be fixed as a fixed low-pass filter for instance.

2.3.4 Coupling

The theory presented here above is based on the hypothesis that the waveform and the VTS are independent. In practice, this is questionable. According to [Maeda, 1982], the waveform of glottal area is supposedly independent from the vocal tract shape for some vowels (example: /i/, /a/, /u/), but the shape influences the waveform of the glottal flow. Consequently, integrating some coupling between the exciter and the resonator would make the model more faithful to reality.

2.3.5 Voice Synthesisers - Examples

Numerous voice synthesisers have been developed, especially over the last decade. We briefly present here two pieces of software that can be taken as references, either for the theory used, the selected parameterisation, or the manageability of the final product.

LeTalker (LeT)

LeTalker, named after Lumped-element Talker model [Story, 2013], has been completed on Matlab [Oppenheim and Schafer, 1998] by Story. The excitation method used is the three-mass-model (3MM, see Sec.2.3.1) based on [Story and Titze, 1995], and the resonator is a classical cylindrical discrete VT. It aims at synthesising sounds through the manipulation of physiological parameters such as the lungs air pressure, the contractions of cricothyroid and thyroartenoid muscles.

Pink Trombone (PkT)

PkT [Thapen, 2017] is an experiment aiming at shaping a sound into an utterance in real time by manipulating the shape of the VPS. Its aim is to be manageable by any average person, i.e. to be a "bare-handed speech synthesis" (PkT's descriptive motto). In practice, it means that all complex parameterisation is hidden. The excitation is based on [Lu and Smith, 2000] and the resonator is based on [Story, 2005].

		LeT	PkT
Modelling Features	**Resonator**	ϕM	ϕM
	Excitation	3MM	LF-85
	Nasal tract	No	Yes

Table 2.1: Synthetic comparison of two speech synthesizers: LeTalker and Pink Trombone regarding modelling. The feature LF-85 refers to Liljiencrants-Fant analytical model (1985) [Fant et al., 1985]

2.4 Ageing Voice

Ageing is a natural process that is caused by time passing. In humans, ageing represents the accumulation of changes over time, encompassing - amongst others - physical, psychological, and social changes. Obviously, the voice which is produced by physical organs is affected by such changes, and these changes are audible.

2.4.1 Voice over Time

This part summarises the current knowledge of the causes and effects of time on the voice. The reader is invited to refer to both [Makiyama and Hirano, 2017, Rojas et al., 2020] for an overview of the topic and for detailed information.

Physiological Changes - Overview

Let us consider two speech signals, selected (or produced) to be representative of two different age groups, for instance 30-40 and 60-70 years old respectively. The following observations can be made, when characterising the "older" speech signal relatively to the "younger": fluctuations in the period (jitter) and in peak amplitude (shimmer) increase, f_0 shifts [Brown et al., 1991, Stathopoulos et al., 2011] the noise at the glottis augments [Ferrand, 2002] and spectral characteristics [Eichhorn et al., 2017] and general voice level evolve [Stathopoulos et al., 2011].
Reasons for such alterations are that all the organs involved in voice production endure the passing of time: the lungs, the VF, the cavities and the extremities. Observable symptoms are reduced respiratory power, decrease of the VF function and VF bowing (presbylarynges), and weakening of the motor function of resonant organs such as the palate, lips, and tongue. At a microscopic level, this may be attributed to histological degeneration of the organs; for example for the VF, atrophy and sparse distribution of the elastic fibres can occur in the intermediate layer of the lamina propria.

Ageing Symptoms

When analysing a sustained speech signal featuring a vowel, some parameters are symptomatic of the ageing phenomenon. A selection among the most conspicuous ones is presented here: pitch, vibrato, tremolo, formants, vocal noise, general voice level and speaking rate. The general evolution of these parameters - gender-specific most often - is briefly described below; however, the reader should bear in mind that these alterations are highly individualised in reality.

Mean f_0 Evolution The intonation of a speaker is rarely monotonous, even over one sentence. This is due to small variations of f_0 around its mean. However, the average f_0 participates in characterising a person at a given age. We now use the *mean vowel fundamental frequency* (mvf0) as an ageing feature. Literature agrees on the general decrease of mvf0 for women over time, which is mostly related to hormonal change after menopause; on the contrary, mvf0 evolution for men is more uncertain [Eichhorn et al., 2017]. Some results for four authors are illustrated in Fig. 4.1. Plus, mvf0 variations increase even more at old age and for women [Stathopoulos et al., 2011, Rojas et al., 2020].

Tremolo The tremolo characterises local variations of amplitude in voice. The influence of age on tremolo lacks consensus across existing studies [Makiyama and Hirano, 2017]. According to this study however, tremolo seems to increase significantly with ageing, with an effect more pronounced for women (83% for male subjects *vs* 93% for female subjects). Tremolo appears at variable ages and at variable intensities for both genders.

Vibrato The vibrato characterises local variations of period in voice. The influence of age on vibrato lacks consensus across existing studies [Makiyama and Hirano, 2017]. According to this study however, it seems to be rather steady than evolutive in a direction or the other. The opposite variations between tremolo and vibrato imply that the amplitude of the glottal flow may be affected more by degeneration in the respiratory system and tissue changes of the vocal folds than vibration frequency.

Spectral Characteristics The evolution of formants would need further research. Overall, formants frequencies do no systematically decrease with age and different vowels may follow different trends. Nonetheless, existing work shows a general diminution of F_1 frequency - especially in women [Eichhorn et al., 2017] - and of higher harmonic levels [Makiyama and Hirano, 2017]. As for bandwidths, they are determined physically by the combined effects of radiation, compliance of the vocal tract walls, viscosity, heat conduction, and glottal opening [Kent and Vorperian, 2018]. No clear effect is known, for existing results diverge - most probably due to differing experimental conditions and data processing methods.

Vocal Noise The vocal noise relates to the quality of voice in terms of roughness (whether the timbre sounds smooth or broken up) and breathiness (whether the timbre is well defined or lacks consistency due to too much air in the voice). The ageing process seems to cause an increase in noise within the high frequency range especially, which is related to decreases in higher harmonic levels with ageing. It is said to be gender-dependent and to touch women

Voice Loudness In [Stathopoulos et al., 2011], voice loudness seems to increase linearly with age and similarly for both genders. A large variability is observable between individuals and is said to steadily decline then increase again passed 60 years.Other results reported in [Stathopoulos et al., 2011, Makiyama and Hirano, 2017] show different trends (voice loudness stagnation or decrease with age). Overall, this aspect of the voice doesn't seem to reflect the declining laryngeal system; it may instead be an effect of the declining auditory system.

Speaking Rate Speaking Rate is a supra-segmental features contrarily to all the previous ones. As shown in [Mokhlesin et al., 2017], it is influenced by age; furthermore, it seems to be a determinant clue for age estimation, with an impact stronger than f_0 [Harnsberger et al., 2008]. It is also the main voice feature naïve speakers who wish to artificially age their voice will use [Skoog Waller and Eriksson, 2016].

2.4.2 Ageing Voice Synthesisers - Experimentation

Most synthetic voice models are configurable in terms of vowel, gender, even muscle stress etc., but not in terms of age. And yet, the voice changes over a lifespan. Some experimentation has

	Child → Adult	4p-interp
Author	[Story et al., 2018]	[Schotz, 2006]
General method	ϕM	Formant synthesis
Main Feature	Vocal tract length	23 parameters: f_0, f_i,...
Original data	VTs of Children and adults	Voices of 4 related persons of different generations
Procedure	Derivation of length warping and cross-dimension scaling functions	Interpolation

Table 2.2: Synthetic comparison of two ageing voice synthesizers

been done however, and although little literature about this topic is available, two pieces of work are worth mentioning here - and are summarised in Table 2.2.

Childhood to Adult voice This study, developed by [Story et al., 2018], has constructed a developmental and sex-specific version of a parametric vocal tract area function model, representative of male and female VTS ranging in age from infancy to adult age.They analysed the VTS of adults and children and derived a general transformation law over time. This study focused on physical cues (the VTS) but did not consider the evolution of the source (the VF) or physiological transformations (tissues). However, the model was assessed and validated through three experiments comparing children and transformed adult vocal tracts.

Age interpolation for voice synthesising This model, developed by [Schotz, 2006], is an analysis-synthesis ageing model based on formant synthesis and age-weighted linear interpolation of 23 parameters - among which those six described previously. It aims at simulating an age between the ages of any two of four differently aged female reference speakers belonging to the same family. Several assessments have revealed satisfying similarity between real and synthesised words but also several weaknesses, such as distortion owing to large differences in the adjacent formant frequency and voice parameter values.

Other Voice-Time Related Projects Another branch of voice research that relates to some extent to our topic is the evolution of the voice at a larger time scale, i.e. thousands of years through human evolution. It aims at understanding whether our ancestors could have talked, and in the case of a positive hypothesis, what sounds they could have uttered [Boer and Fitch, 2010]. Other animal species are also under study; note that computer models exist in this specific branch [Wilkinson, 2016].

Chapter 3

Requirements

Based on the research and problem statements, a list of basic specification items was composed to use as the guiding thread during the implementation phase.

Purpose

This project aspires to provide a means of experimentation on the phenomenon of ageing: whether for the purpose to discover the main aspects of the ageing voice or to apply them to their own voice.
Although the phenomenon of ageing is quite disparate among the population, some symptoms are consistently mentioned into the medicine- and speech-oriented literature [Makiyama and Hirano, 2017, Rojas et al., 2020], such as the evolution in pitch, tremolo, vibrato, loudness, glottal noise, spectral characteristics and speaking rate. This "macro"-model therefore aims at transcribing this medical knowledge - based on signal analysis - into computable and manipulable features.We assume that the quality and credibility of the synthetic voice are of primordial importance. Indeed, the user needs to be able to acknowledge the virtual speaker before he starts considering to vary his age.

Features

- the voice is human-like, seems natural and vowels are identifiable - the quality of the synthesised speech is refinable through low-level parameterisation
- the observation of the ageing phenomenon is facilitated through the concatenation of differently-aged samples and/or to real-time changes of the user-defined, along with the existence of an ageing mode on at least one support
- the ageing phenomenon is applied to recorded voice for a better appreciation.
- the parameterisation is flexible and addresses various levels of technicality
- several formats of code correspond to different levels of modelling complexity (from low- to high-level)
- the code is optimised to reduce the computation time
- an analysis feedback and/or visualisation are provided
- the analysis crosses several methods to strengthen the results
- the parameterisation and the corresponding modelling and results can be stored for future verification or comparison
- the raw data in use (VTS, age-parameter relation) can be checked before use

Chapter 4

Design / Implementation

In this chapter, we present the different steps and actions that were undertaken and connected within a framework to create the model.

4.1 Data

In this section, we present the methodology employed to make up realistic raw data to feed our algorithm with, or to analyse our computed data against.

4.1.1 For Synthesis

This project being about *physical* modelling - at least in part -, we need to know *how* the voice is produced, i.e. the vocal apparatus dimensions and properties. This subsection addresses the data needed for Source modelling and for Filter modelling accordingly with the source-filter theory.

Excitation As the analytical method was chosen for this element, the three GPs presented in Sec. 2.3 were implemented. The parameters *ptp* and *pte* were approximated from illustrations encountered in the literature, particularly [Hézard, 2013].

Time envelope The three types of envelope were implemented. The data of 12 French-speaking persons (6 male, 6 female) were exploited. They are aged between 23 and 85, clustered into 3 generations centred on 26.9, 60.3, and 85 years old; mean: $\{43.3 \pm 22.6\}$ y.o.). The vowels pronounced were /a/,/ø/,/i/,/o/,/y/,/ɛ/,/u/. From this data, the average fading-in duration was estimated at $\approx 13\%$ of t_{utt} and the average fading-out duration at approximately twice as much. They are used in all three implementations. In practice, the fading values can be chosen to be taken randomly in a small interval around the values aforementioned. The second and third envelope types were also extracted and computed.

Noise at the Source It was implemented as in subsection 2.3.1.

Filter The vocal tract shape is the resonator of the vocal organ, in which the excitation signal propagates. A model of vocal tract was hence essential.
Two types of VTS were taken from the literature and investigated, both from Story work. The first one [Story et al., 1996] is based on MRI and averaging on a consequent number of images for a given gender and sound: consequently, it is supposed to be an accurate reproduction of the average human vocal organ. The second set of VTS [Story, 2005] is the result of some

analysis on the VTS of the first set and feature reduction by projection of the data on two dimensions; the dataset thus obtained is an approximation of the average VTS for each gender and every sound. In this second case, the resulting dataset is not a set of VTS but is made of three variables and constants instead (see Eq.4.1). At the basis of the model is the neutral VTS Ω, used for all utterances. It is altered with the variables $\phi_m(x)$ that characterise the mode m at position x in the VT. Finally, some coefficients q_m are applied to reconstruct the different vowels.

$$V(x) = \frac{pi}{4}\left[\Omega(x) + \sum_{m=1}^{M} q_m \phi_m(x)\right]^2 \tag{4.1}$$

where $M = 2$ is the number of modes considered. We refer to these two models respectively as vts-mri and vts-pca from now on.

Nasal tract Data for modelling the NT was taken from [Xi and Longest, 2009] where it was originally obtained with MRI. It was branched to the main VT at ≈ 8 cm from the glottis with a parallel junction [Bilbao, 2004].

4.1.2 For Ageing

In this subsection, we address the methods employed to model the age-feature relations.

mvf0 Many different sources were available, with relatively different curves due to different experimental setups. All data was estimated from data pictures in the literature in sampling the curve at critical points (e.g. gradient zeroing). The resulting approximation seems to be fair in comparison with the discrepancies observed between data (see Fig. 4.1). In particular for the data extracted from [Stathopoulos et al., 2011] work, the approximation done in the original paper was already surprisingly coarse. This participates to explain the large difference in frequency range between this author and the others. When testing these functions in practice, the fact that the combination Linville-female in [Linville, 1996] was not available was a hindrance to compare all authors. All authors are still available however, but [Brown et al., 1991] and [Makiyama and Hirano, 2017] are recommended rather than the two others. Additionally, some variations were applied to f_0 to model the *pitchsigma*, i.e. the standard deviation of f_0 between different utterances (or in a sentence). The data used for this purpose was [Peterson and Barney, 1952]. This was thought essential to attenuate the monotony of the series of vowels synthesised.
Independently of pitch sigma, it is important to notice that these two functions are *not* convex! Therefore, there may exist several solutions to an equation. For instance, when looking for the corresponding age to a voice fundamental frequency of 118 Hz, three values are possible: 31, 54, and 68 in Makiyama model. This only criterion being insufficient to determine age from a frequency with absolute certainty - or conversely, other parameters need to be put into consideration.

All other ageing effects are supposed to appear from a certain age, the *pivot-age* a_p, randomly drawn in [50-75] y.o.. Furthermore, most curves can be approximated as linear or quadratic - possibly piecewise - functions [Stathopoulos et al., 2011].

Tremolo The tremolo is measured by the *shimmer* metrics (see 5). It is applied through a sinus function dependent on age a, gender g and time t:

$$f_{trem}(a, g, t) = A_{0,trem}(a, g) \sin\left(2\pi f_{0,trem}(a, g, t)\, t\right) \tag{4.2}$$

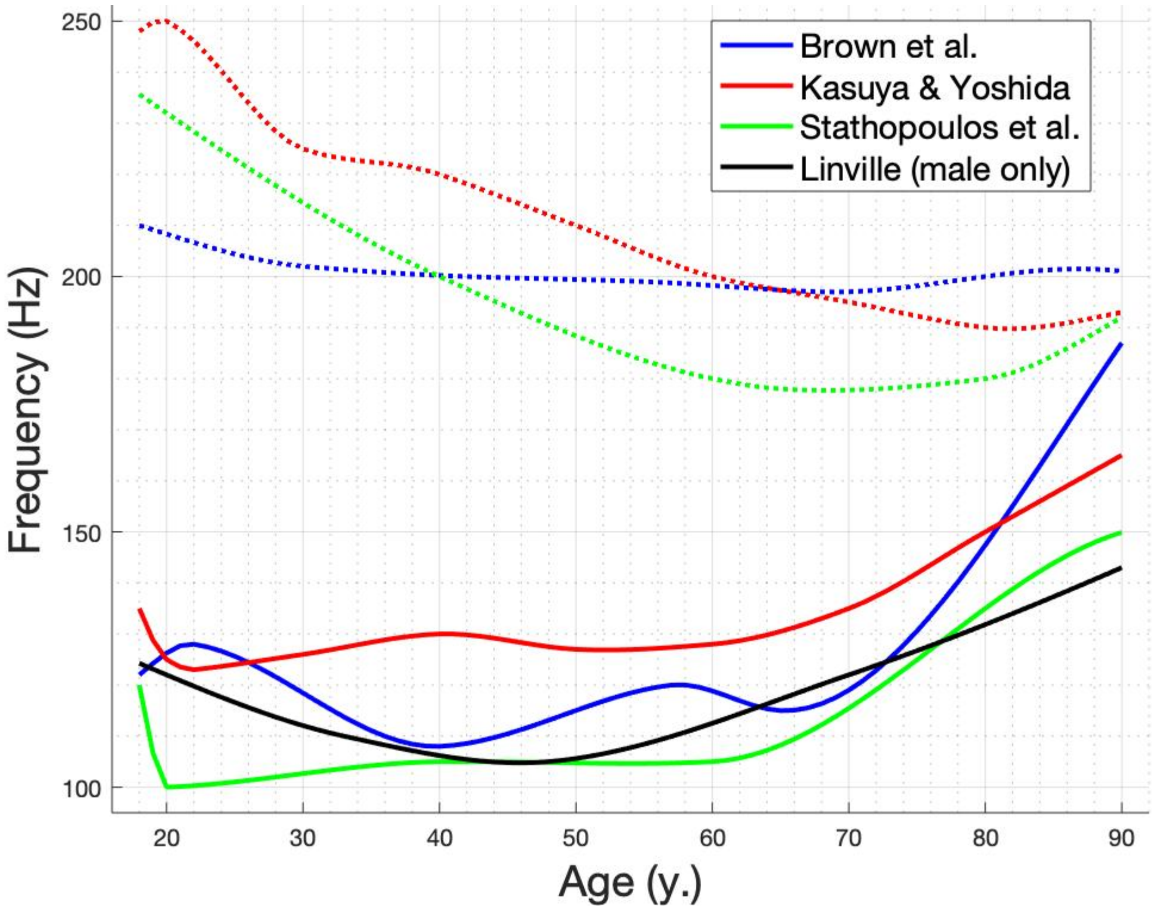

Figure 4.1: Average mvf0-age relations determined by four authors: [Brown et al., 1991, Makiyama and Hirano, 2017, Stathopoulos et al., 2011, Linville, 1996]. Data was visually extracted from cloud points or regressions and interpolated with a modified Akima cubic function. Continuous line: male; dotted line: female.

where $A_{0,trem}$ and $f_{0,trem}$ represent the modulation amplitude and frequency, respectively, and are themselves functions of age. These two parameters are deducted from the expected shimmer (in %), *shim*, which is itself dependent on age and gender. How to determine *shim* for a given age and gender?
The shimmer is set as a piecewise function, f_{shim}, on the model of [Makiyama and Hirano, 2017, p.32]. Note that the plots in the document in question display longitudinal data for two different *individuals* and do not intend to show a generality; therefore, a certain amount of randomisation is introduced in the definition of the function here (noted by a superscript star $*$ in all equations). As non-pathological values for shimmer are below 3% for the sustained phonation in young adults [Teixeira et al., 2013], the first part of f_{shim} is fully described by the constant C_{shim}, randomly picked in [0.3-1.5]%. The second section is an affine function whose constant slope dC_{shim} is taken in [0.10-0.17]% for male voices and in [0.10-0.25]% for female voices. As a result, the function controlling f_{shim} as a function of age is:

$$f_{shim}(a, g) = \begin{cases} C^*_{shim} & \text{if} \quad a < a_p \\ C^*_{shim} + dC^*_{shim}(g) * a & \text{if} \quad a \geq a_p \end{cases} \tag{4.3}$$

Given this function and the user-defined age a_0 and gender g_0, the model collects the appropriate value $shim_0 = f_{shim}(a_0, g_0)$. This value is used to fetch in a data pool the corresponding values of $A_{0,trem}$ and $f_{0,trem}$. Note that the calculation of *shim* is highly dependent on f_0- for a certain f_0. This correspondence was obtained by running the Matlab script for a large range of values of $pf_{0,trem}(\%) = f_{0,trem}/f_0$ (in [0.001-0.03)]%), $A_{0,trem}$ (in [0.05-0.5], normalised) and f_0 (Hz)

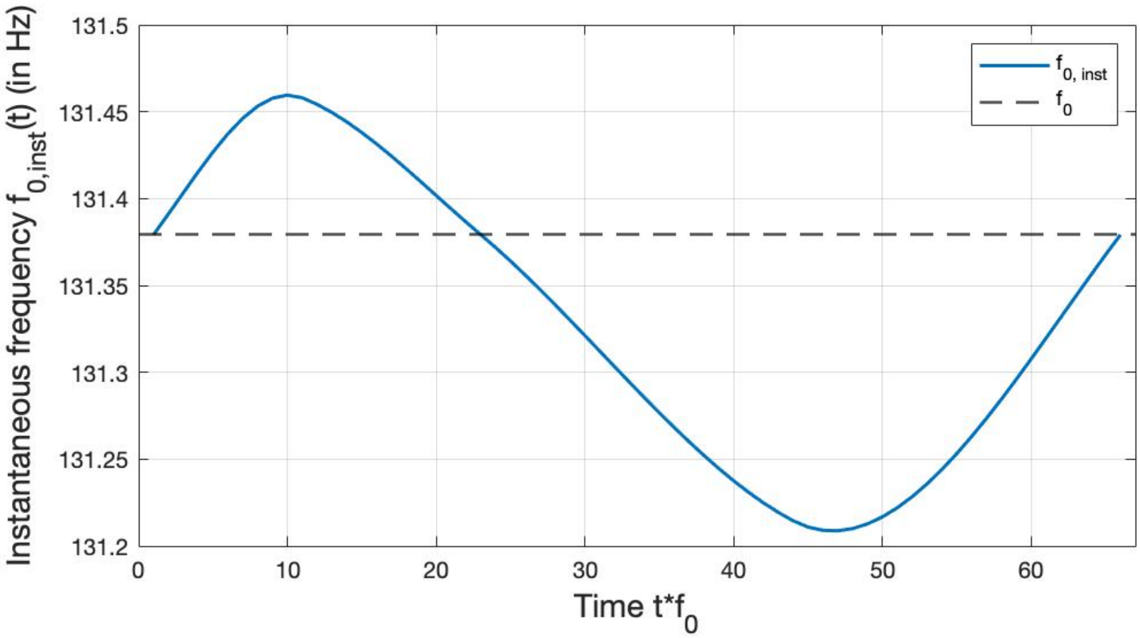

Figure 4.2: Irregular vibrato for a male speaker aged 50 y.o.. $f_0 = 131.3792$ Hz, $f_D = 0.0084.f_0$, $n_{marks} = 3$, $dnutt = 0.8918\%$.

(in [75-250]).

Vibrato The vibrato is measured by the *jitter* metric which depends on f_0. Typical value of jitter measured during sustained phonation on young adults are in [0.5-1.0]% of f_0 [Teixeira et al., 2013]. However, in [Awan, 2006], the jitter calculated on a population of women aged from [18-79] and without known voice pathology cover the interval [0.36-0.82]% of f_0. Unlike for the tremolo, the age-vibrato relation is modelled under two *explicit* forms.
The first form consists in using the Matlab frequency modulation (FM) function "*comm.FMModulator*", which is fed solely with f_s and the frequency deviation f_D: this parameter quantifies how much f_0 varies from its average at a short-term scale. In our situation, $f_D = jitt(\%) * f_0$.
The second form was implemented from scratch and aims at producing irregular vibrations in frequency (an irregular FM). It is used during the construction of the GP. On top of the necessary parameters used in several functions, it takes three specific parameters as input. First is the frequency deviation f_D which is the amplitude of the FM. The second parameter is an integer and is indirectly related to the frequency of the FM: in practice, it encodes the number of times the instantaneous frequency $f_{0,inst}$ crosses the line f_0 within an utterance. Let's call it n_{marks} for it counts the number of time markers. The third parameter is a percentage and is also connected to the frequency of the FM. It defines the standard deviation of the interval length (in samples) between successive occurrences of f_0. In other words, this parameters moves the time markers along the time axis. As such, n_marks and this third parameter are interdependent. Let's call this second parameter *dnutt*, for it applies small variations in samples to the subparts of an utterance.
Once all three parameters are set, all $f_{0,inst}$ are calculated by interpolation between the different time markers at frequencies calculated within $[f_0 - f_D, f_0 + f_D]$. Eventually, the final GP is obtained by concatenating multiple GPC that differ not only in their $f_{0,inst}$ but also in the manner they are grouped on both sides of f_0 occurrences. It was taken care that all GPC were non-truncated.

Vocal Noise The noise is measured by the *HNR* metric. The vocal noise over lifespan is modelled as an affine function dependent of age according to [Stathopoulos et al., 2011, Fig.3], f_{snr}, varying from [23-26] dB for female speakers and from [21-25] for male speakers from age [18-90] y.o. To the value picked on the correct curve (male or female) at the selected age a_0 is added a coefficient that combines the standard deviation of *snr*, dependent on age, which is moreover slightly varied thanks to a random term.

Utterance Duration The duration of a diphthong (in "*white*" and "*light*") was found to be shorter by 22 ms (144 ms *vs* 166 ms) in favour of young male speakers against their elders [Harnsberger et al., 2008]. These values can *not* be transposed here directly since we consider monophthongs (i.e. there is only one vowel sound in a syllable), and moreover because too short durations would prevent apprehending the ageing effects that need longer portions of signal to be audible (e.g. tremolo). However, similar ratios between old and young speakers were tested out:

$$t_{utt,old} = r_{utt}\, t_{utt,young} \quad \text{with} \quad r_{utt} \in [1.2 - 1.4] \tag{4.4}$$

In this project, the default utterance duration was set at 0.5 second (young speakers), and the definitive ratio at 1.4. The utterance lengths in a sentence are also altered by the Vibrato method (see above).

Loudness Loudness is *not* gender-dependent and is here measured with SPL. Loudness is modelled based on this metric, from [Stathopoulos et al., 2011, Fig.2] as an affine function of age. A similar evolution was thus applied to the sub-glottal pressure. Set default at 7840 dyn/cm^2 for young speakers, it is multiplied by a scaling coefficient in [1-1.3] linearly associated to ages in [18-90] y.o.. Note that other results reported in [Makiyama and Hirano, 2017] show different trends (SPL stagnation or decrease with age).

4.2 Framework and Model

In this section, we present the framework that supports the ageing model that was created. There are several ways to build a model in signal processing: in the time domain, in the frequency domain, or both.For the voice in particular, it is worth reflecting about the question, since a dynamic and fast-varying signal - the GP - is filtered by an equally fast evolving system, the vocal tract. The choice was made to stay in the time domain for the synthesis, but to go to the frequency domain for the analysis.

4.2.1 Framework

All this project was realised in Matlab [Oppenheim and Schafer, 1998], mostly on versions R2019b and R2020a. Data was stored in Matlab structures 'struct'. The implementation was adapted to endure different types and lengths of structures for most inputted parameters. A main script, an interface, libraries, and related functions were mostly implemented by myself, except for some Matlab functions ('findpeaks', 'envelope', 'convhull') and for the f_0 estimation library fastF0Nls [Nielsen et al., 2017]. All elements are integrated in the framework described in Fig. 4.3.
The model was developed under three formats: a script (working document), an interface and a (fixed-age) real-time plugin. The interface and main script behave similarly, but differ in the flexibility they offer to the user, the main script being far more parameterisable while the interface being more user-friendly. They both possess an *ageing mode*. The real-time plugin, despite being deprived of such mode, enables the user to try out different parameters (among

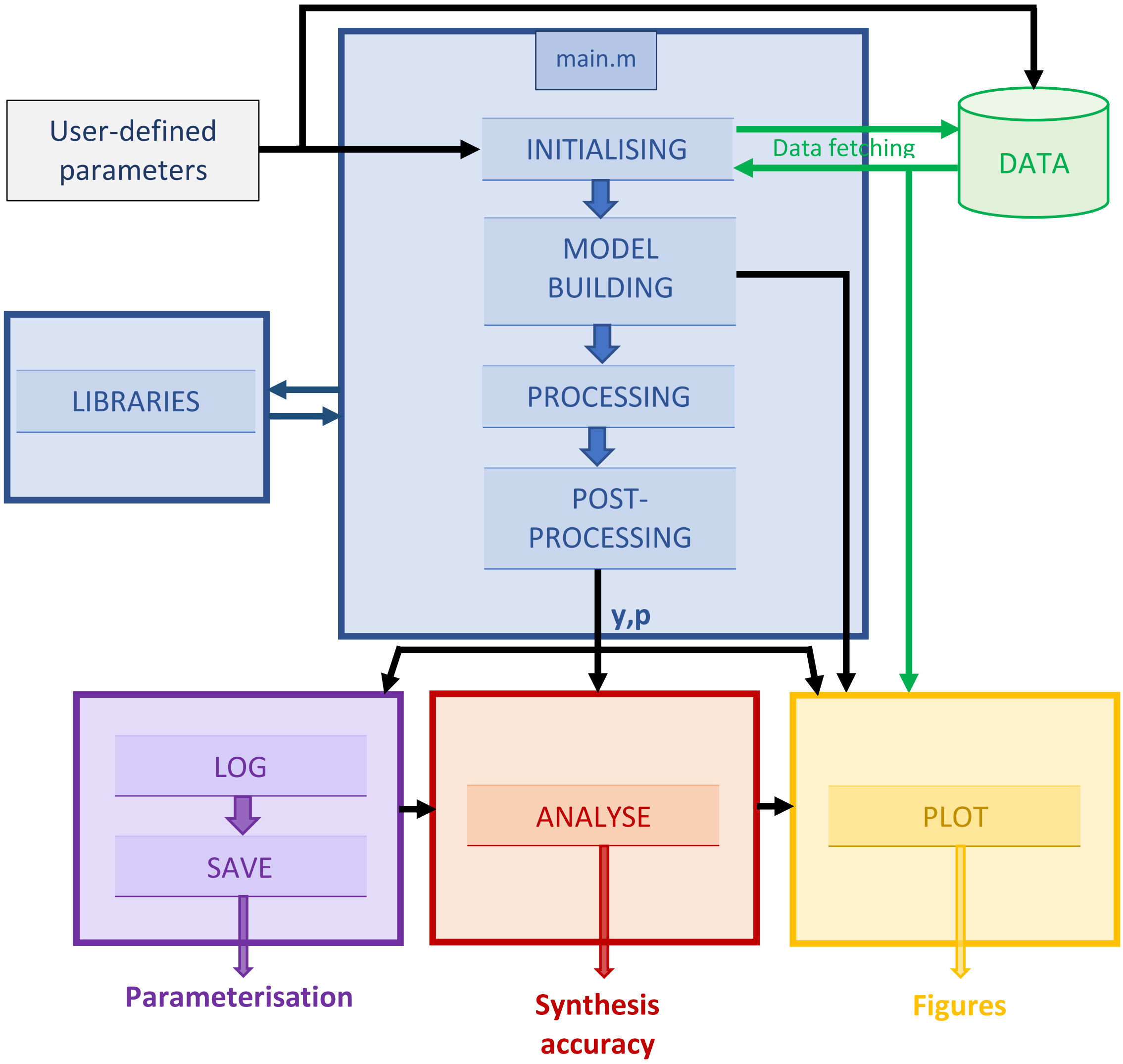

Figure 4.3: Framework - y: audio file, p: parameters.

which those responsible for ageing) and to understand their impact independently. Data bases for VTSs, GPs, mvf0s and shimmer were prepared and stored on path for a quick access during pre-processing. This data was made for both *a priori* and *in process* use: a set of functions was created for observing, analysing and understanding how different occurrences of aforementioned parameters behave.
The core process of this project can be divided into three main parts: initialisation, synthesis, and post-processing which are presented in the following subsections.

4.2.2 Initialisation

This stage consists for the manipulator in tuning the primary parameters the model will be fed with. The quality of the model depends on the fine-tuning of these parameters; this parameterisation is therefore a key stage.
A substantial amount of inputted parameters was used in the model that was created. Only part of them is available from the interface - as visible on Fig. 4.4. For legibility and understanding, these parameters are divided into three categories as follows:

- [B] Basic human-characterising or user-defined: that contain very common and easy-to-get concepts; e.g. fundamental frequency f_0 (as an approximation of pitch), gender, vowels (see recap Table 4.1).

Phonetics	Key-word	Word example (en)	Word example (fr)
/i/	ii	*heed*	*île*
/ɪ/	ih	*hid*	*clé*
/ɛ/	eh	*head*	*mère*
/æ/	ae	*had*	N/A - *patte*, *pâte*
/ʌ/	ah	*ton*	N/A - *patte*, *ceux*
/ɑ/	aa	*hod*	*pâte*, *sort*
/ɔ/	aw	*paw*	*sot*, *sort*
/o/	oo	*hoe*	*sot*
/ʊ/	uh	*hood*	N/A - *sot*, *ceux*
/u/	uu	*who*	*coup*, *tu*

Table 4.1: Vowels summary table. Word examples are given in English, and approximate equivalents are provided in French.

- [M] Model-related: all necessary parameters that control the excitation or the nature of the signal; e.g. VTS from [Story et al., 1996] or the one from [Story, 2005].
- [E] Effect-related: all optional parameters that affect the resulting speech sample.

Additionally, some analysis-related parameters [A] were added to facilitate analysis and/or rendering, and a batch of running *modes* [R] was created to permit relevant combinations of parameters. For a comprehensive list of categorised parameters, see Appendix B.

Finally, here are a few words about how to set basic parameters. For having collected more data for male than for female, it is recommended to keep the gender to 'male' (id=1). Furthermore, male gender was also tested in priority because it represents the "easy" case, with lower fundamental frequencies and bigger frequency intervals in the log-frequency domain which prevents too much proximity between f_i.

4.2.3 Synthesis

The synthesis is two-fold: first step is about creating the model derived from user-defined parameters; second step is about the processing of the model.
For model building, the parameters inputted are now used as key-words by the software to fetch corresponding data. Given the vowel(s) chosen, one or several VTS will be selected - example on Fig. 4.4 is given on top picture with vowel /i/. New physical parameters, such as Z, k, α are then computed for use during processing. Regarding the nasal tract, it can be activated in the script; doing so changes the propagation environment in adding a second duct, connected to the first one at about the half of this latter via a parallel scattering junction.
The glottal pulse is computed according to the excitation method selected. Either it computes a parametric GP based on a parametric analytical function; either it loads a glottal *area* whose samples will be processed successively to produce the glottal pulse.
The tremolo and vibrato are optional.
This data is finally processed in applying the theory presented in Sec. 2.3.

4.2.4 Post-Processing and Logging

After processing, all parameterisation **p** and an audio file **y** are outputted from the system. **y** is stored three times: as is, normalised, and multiplied by an envelope to render the sound

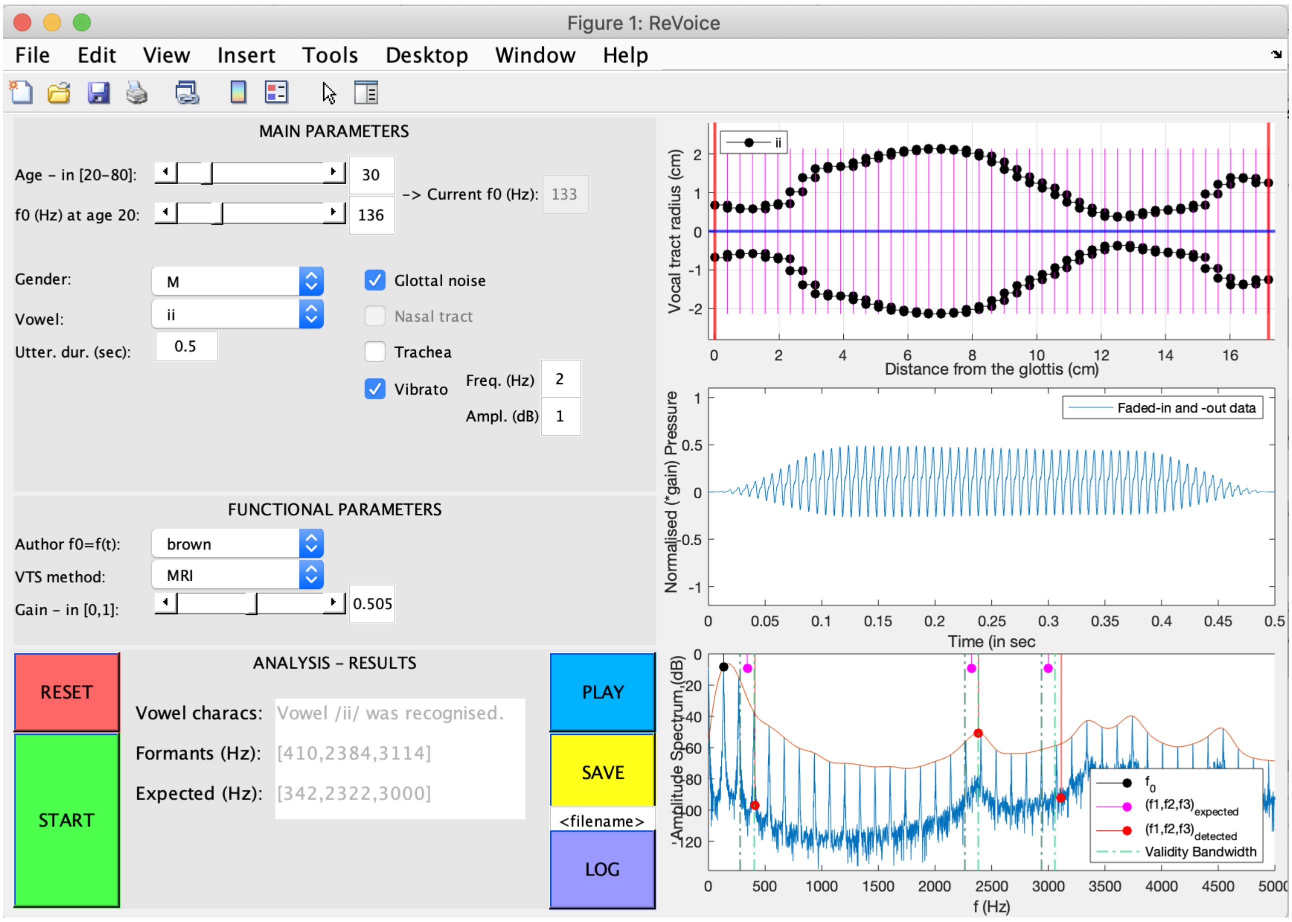

Figure 4.4: User Interface.

more natural. Additionally - and optionally, the *sentence*[1] is smoothed in applying overlap-add technique between vowels with - in case there were several vowels. If the user wishes to keep track of past parameterisations and corresponding audio files, all data can be logged in three structures connected by an index common to all. One of these structures is a 'wav' file, another is a quick description of the main parameters used (e.g.: vowel, gender), and the last one is the whole set of parameters. Based on this data, the user can ask for its analysis, in which case formants will be estimated and compared to existing knowledge (see Sec. 4.1). Finally, the user can also choose to display figures related to the data used (e.g.:VTS) or to the results obtained (e.g. spectrum, spectrogram).

4.3 Age Morphing

It is one thing to apply ageing to an all-synthesised speech signal; use it on recorded voice is another, and seems relevant to this project to assess the efficiency of the ageing effects on actual real voice. Some work exists in this field [Skoog Waller and Eriksson, 2016, Rupal and Seth, 2017]. *Age morphing* is the expression further used to refer to this second manipulation. However, the voice handling process for both these applications is highly different. In the first case, the developer controls the signal completely and absolutely within the framework, and from an initial neutral signal it is effortless to manipulate it to force a different perception - in our case: ageing. In the second case, the recorded signal comes from outside of the aforesaid framework with its default and natural variations.

[1]Sentence: succession of several sounds.

In order to apply ageing to a real speech signal, several steps are to be considered. The first optional step is speech segmentation - i.e. the identification and separation of different utterances - and the extraction of all the vowels. Indeed, only vowels are of use here. This stage is unnecessary if the speech signal is a mere vowel; however, the information extracted in this case will also be limited.
The second stage - strongly recommended - is about analysing the vowels extracted and computing ageing metrics (see Sec.5). An estimation of the age of the speaker can then be gauged. The third stage - strongly recommended too - consists in *neutralising* the signal and to bring it to a neutral profile. In practice, it means removing all effects responsible for ageing. In case the speaker is young enough, these effects would be limited and this step can be skipped. The last stage involves applying the changes corresponding to a given age to this neutral audio file. A prototype script was developed in Matlab following the steps:

- record or read an audio file of real speech
- apply *comm.FMModulator* to simulate a regular vibrato;
- apply *audioTimeScaler* to extend or reduce the length of the utterance;
- apply *shiftPitch* to shift pitch according to a given age-f_0 relation;
- use simple sinus function to model the tremolo.

Unfortunately, the successive transformations introduced their share of artefacts and prevented the formation of a truly realistic image of the speaker and of their age.

Chapter 5

Evaluation

We present here the different aspects of evaluation of the framework developed.

5.1 Voice Evaluation - Overview

This section presents the tools and methods usually employed for voice analysis and perception in the literature, and the related caution when using the results.

5.1.1 Acoustic Voice

Voice analysis has a considerable history that we outline here; however, a good synthesis of the topic in literature is available in the review [Kent and Vorperian, 2018].

Formant Estimation The main features extracted in literature are usually the fundamental frequency f_0 (in Hz) and the central frequencies f_i (in Hz) of the formants F_i. The bandwidths b_i (in Hz) and amplitudes a_i (in dB) complete the spectral description of the F_i.
To estimate formants, visual and automatic detection methods are applied on FFT-based computation (spectrogram, spectrum, cepstrum) or on LPC. When looking for formants in a speech signal, it is recommended to select a period of time during which the formant pattern is static, to prevent averaging on very different data that can occur in transitory periods, which would introduce estimation error. Since the signal properties can vary significantly in different nearby regions, such as the closed and open phases [Gray and Markel, 1976], synchronising analysis frames with the instants of glottal closure proved to yield highly consistent estimates of the formant frequencies [Yegnanarayana and Veldhuis, 1998]. Among other speech analysis tools, let's mention Praat [Boersma, 2002] as a reference for voice analysis since it was used in a number of works [Eichhorn et al., 2017, Pépiot, 2014, Vinceslas, 2011]. It uses LPC with various algorithms (e.g. auto-correlation, Burg [chiller, 1978]).
Besides the f_i, there has been a consequent amount of time and energy spent over formant bandwidths since the late 1950^{ies}: notably [Dunn, 1961, Fant, 1972, Hawks JW, 1995] and many others who determined that formants frequencies and bandwidths depend on each other and on the glottis configuration, and researched on methods to measure them accurately. One of these relations is that bandwidths increase for both male and female speakers as the formant frequencies become higher; and it is to be noted that the bandwidths for females are wider with greater variances than those for males [Yasojima et al., 2006].

Related Features Additional features, stemming from the estimation of the f_i are commonplace too: for instance the distance between successive formants - or inter-formant distance - and the *spectral tilt*, defined as the slope (in dB) between successive harmonics and supposedly implicated in voice quality. This latter has been used in [Kreiman et al., 2014, Garellek et al., 2016, Samlan and Story, 2011].

Representation Methods Different visual representations of the formant space exist, the most common one being the - normalised or not - F_1-F_2 space, as in [Peterson and Barney, 1952, Story and Bunton, 2017, Berisha et al., 2014]. Numerous metrics reported in [Kent and Vorperian, 2018] were developed based on these representations, such as various vowel space areas (VSA) [Berisha et al., 2014 the formant centralisation ratio [Sapir et al., 2010], the vowel space density [Story and Bunton, 2017], and the convex hull [de Boer, 2009].

Known Related Issues Several issues can hinder the voice analysis process [Kent and Vorperian, 2018]. For instance, the spectrum is influenced by f_0 when it contains more than one pitch period: formant analysis is therefore made more difficult as f_0 increases and the spacing between harmonics becomes wider. According to [Vallabha and Tuller, 2002], this aspect produces an error of 10% f_0 on formant detection. Furthermore, neither LPC- or FFT-based methods are absolute in formant-frequency or formant-bandwidth accuracy of estimation.As an example, the frequencies of F_1 and F_2 are said to be estimated within an interval of $\pm$60 Hz ($f_0/4$ at best) [Kent and Vorperian, 2018]. According to [Vallabha and Tuller, 2002], four factors impact this accuracy among which an incorrect choice of the LP filter which yields an error of $10 - 80$ Hz on formants, whatever the method employed. Another problem is the proximity between f_0 and the f_i or in between the f_i - as for f_2 and f_3 in vowel /i/: in such case, the pair of peaks is likely to hide one peak among the two. Adjustments of some analysis parameters may be useful in this regard: in the case of a FFT-based analysis method, visual assessment and decrease the number of FFT points will help; as for LPC-based analysis, an increase of the number of filter coefficients will do instead. Nonetheless, there is no perfect and absolute solution for a perfect formant estimation; at the end, the quality of the analysis is highly dependent on the analyst himself who must use his *a priori* knowledge of the signal to process it at best. As mentioned in Sec. 2, the f_i arrangement determines the nature of the vowel pronounced and is also indicative of the identity of the speaker: indeed, it is said that a shift of 5% of the three lower f_i annihilates the personality of the speaker [Kuwabara and Ohgushis, 1987]. When it comes to bandwidth - and according to this same author - either modifications in bandwidths of higher order (> 3) or their uniform scaling by 5 or one-fifth also alter the voice personality. To conclude about formants, we insist on the fact that both formant frequencies and bandwidths are gender-dependent: formant frequencies *and* bandwidths are higher for women than for male: for bandwidths, the following relation was stated in [Hawks JW, 1995]: $b(female) = 1.25 * b(male)$.

5.1.2 Age Estimation

Although not as predominant as visual cues, time alterations in the voice are generally audible and provide another person with enough information to estimate the speaker's age, along with other characteristics such as gender, height, social category, etc. In practice, age estimation (AE) from voices is fairly accurate, but its precision is subject to several parameters as reported in [Moyse, 2014] and summarised here below.
The *expert* effect regroups three sub-effects: a person Y will be more accurate in estimating age of person X if they share the same ethnic group or if Y regularly mixes with persons of same age as X [Vestlund, 2004]. In case Y and X are approximately the same age, an additional effect enters the category: the own-age bias [Moyse et al., 2014] detected especially for older adults when estimating age from voices. The *listener's age* effect is about young adults who

tend to outperform older adults at voice-based AE (VBAE) [Linville, 1996]. The effect known as *speaker's ethnicity* is however uncertain in the case of VBAE. The *speaker's gender* affects AE too. Different trends are reported for male and female: VBAE is particularly accurate for female voices [Krauss et al., 2002]. Regarding the stimulus duration, longer stimulus were found to yield better performance [Schötz, 2005]. Some work in VBAE [Schötz, 2005] finally suggests that the age of younger people is often overestimated, contrarily to the age of older people which is generally underestimated. All in all and in a single-modality setting, VBAE is achieved more accurately from faces than from voices, within a 10-year [Ptacek and Sander, 1966] uncertainty interval. Results of a VBAE experiment, based on data characterising the perceived age (PA) of the speaker are usually compared to the chronological age (CA) of the speaker when it is known, by calculating the signed and absolute difference between these two.

5.2 Voice Analysis

In this section, we present the features that were selected for voice analysis in this project.

5.2.1 General Voice

Fundamental Frequency Given an *a priori* knowledge, the f_0 is estimated through peak detection, a correlation-based method and parametric techniques: harmonic summation (HS) and Fast Nonlinear Least Squares Estimation (fastF0Nls) [Nielsen et al., 2017].

Formant Estimation Three methods are used. A LPC-based function implemented and tuned by the author so that the order adapts to every sample by using the *a priori* knowledge of formant frequencies through the following metric:

$$E_f = \sqrt{\sum_{i}^{I=3} \left(f_{ref,i} - f_{meas,i}\right)^2} \tag{5.1}$$

where $f_{ref,i}$ is the expected value of formant i and $f_{meas,i}$ the estimated one. Second, a function based on the detection of maxima in the spectrum was implemented, which considers *a priori* knowledge to resolve some of the issues mentioned previously. For example, the length of the smoothing envelope (hilbert) is parameterised by f_0. Finally, Praat [Boersma, 2002] parameterised such as it detects 6 formants in the range [0-5500] Hz. In both LPC-based methods, the size of the sliding window is not of utmost importance as soon as it is shorter than the utterance duration, for the sound is stationary. The 3 methods are referred to as *myLPC*, *findpks* and *Praat* respectively. These estimations are compared to expected values [Hillenbrand et al., 1995, Peterson and Barney, 1952] or to expected intervals of validity (e.g., frequency within ±60 Hz [Kent and Vorperian, 2018], see the lowest right-side quadrant in 4.4).

Inter-formant Distance Amongst the three characteristics of the formants, the central frequencies hold the predominant role. However, the *absolute* values of these frequencies are not of primary interest, since they depend on f_0. Instead, the *inter-formant distances* are calculated (in Hz) such as:

$$df_i = f_{i+1} - f_i \tag{5.2}$$

where i is the index of the lowest of the two formants considered. Visually, the spectrum and the spectrogram were used. The results of this metric feed the error operator:

$$E_{df} = \frac{1}{2N_v}\sqrt{\sum_{n}^{N_v}\sum_{i}^{I} \left(df_{ref,i} - df_{meas,i}\right)^2} \tag{5.3}$$

where N_v is the number of vowels under consideration and $I = 2$ the number of formant intervals under consideration.

VSA The calculation of VSA in the F_1-F_2 space permits to observe the positioning of vowels relatively the others, and to see how well vowels are reproduced relatively to the literature. In this space, all vowels of interest V are defined as points and are associated with their coordinates (f_V^1, f_V^2). The triangular VSA (tVSA) is described by 3 corner vowels: /u/,/i/,/ɑ/. The quadrilateral VSA (qVSA) is described by 4 corner vowels (/u/,/i/,/æ/,/ɑ/). All vowels are listed clockwise in the F_1-F_2 plane. The formulae for qVSA and tVSA rely on simple geometry. Their results are surfaces and are expressed in Hz^2 or kHz^2. The expressions follow, as reported in [Kent and Vorperian, 2018]:

$$\begin{aligned} qVSA &= \frac{1}{2}\Big|f_i^2.f_æ^1 + f_æ^2.f_a^1 + f_a^2.f_u^1 + f_u^2.f_i^1 - (f_i^1.f_æ^2 + f_æ^1.f_a^2 + f_a^1.f_u^2 + f_u^1.f_i^2)\Big| \\ tVSA &= \frac{1}{2}\Big|(f_u^2 + f_i^2).(f_u^1 - f_i^1) - (f_u^2 + f_a^2).(f_u^1 - f_a^1) - (f_a^2 + f_i^2).(f_a^1 - f_i^1)\Big| \end{aligned} \tag{5.4}$$

The calculation of these two metrics is of course dependent on the detection of the four necessary vowels. Therefore, it is highly sensitive to errors of formant detection. The result of this metric feeds the error operator:

$$E_{VSA} = VSA_{ref} - VSA_{meas}. \tag{5.5}$$

Hull The vowel space representation considers only two formants. For higher dimensional representations, convex hulls and their corresponding volumes can be computed using the *convhull* Matlab function. It was done up to dimension 3, thus corresponding to the calculation of a volume (in Hz^3) in the F_1-F_2-F_3 space. The result of this metric feeds the error operator:

$$E_{hull} = hull_{ref} - hull_{meas}. \tag{5.6}$$

All these metrics can be represented alone or along with the reference to enable a visual comparison.

5.2.2 Ageing Voice

When analysing a sustained speech signal featuring a vowel, some features can be extracted to characterise the voice in an ageing perspective. Among the selection of metrics presented in 2, the ones specific to ageing are developed here.

Jitter The jitter can be calculated under different forms, depending on the intended unit of the metric and on the time range considered. Here following are two expressions of jitter that measure the average absolute difference between *two* consecutive periods, in seconds (jitta) and in % (jitt):

$$jitta\,(s) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{5.7}$$

$$jitt\,(\%) = 100\,\frac{jitta}{\frac{1}{N}\sum_{i=1}^{N-1}T_i} \tag{5.8}$$

Expressions for additional jitter measures are available in [Teixeira et al., 2013].
It is agreed that common values for *jitt* belong to the interval [0.5-1.04]%.

Shimmer The shimmer characteristics can be calculated under different forms, depending on the intended unit of the metric and on the time range considered. Here following are two expressions of shimmer that measure the average absolute difference between the amplitudes of *two* consecutive periods, in % (*shim*) and in dB (*shdB*):

$$shim\,(\%) = 100\,\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N-1}A_i} \tag{5.9}$$

$$shdB\,(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\,log_{10}(\frac{A_i}{A_{i+1}})\right| \tag{5.10}$$

Expressions for additional shimmer measures are available in [Teixeira et al., 2013].
The shimmer becomes pathological for values of *shim* above 3.81% [Teixeira et al., 2013].

HNR The Harmonic-to-Noise Ratio (HNR) measures the vocal noise by quantifying the ratio between periodic components and non periodic component in a segment of voiced speech. For a stationary signal x where the noise is supposed white, the HNR expression is:

$$HNR\,(dB) = 10\,log_{10}\frac{r_x(0)}{r_x(0) - r_x(T)} \tag{5.11}$$

where r_x is the auto-correlation of signal x and $T = 1/f_0$. Note that HNR values may differ depending on the window size that is used.

Sound Pressure Level The voice level is measured with the SPL as in Eq. 2.1. Variations in SPL are modelled as declining until reaching 60 y.o. then increasing again passed 60 years.

5.3 Perception

The credibility of the model has also been evaluated perceptually during a three-fold within-subjects test where each stage was aiming at assessing an aspect of the model: voice identity, vowel accuracy, and ageing simulation. Free expression has been made available at every stage. Out of concern for the homogeneity of the results, it has been decided to look for respondents sharing the same mother-tongue. The largest population group available is French speaking; hence, French has been used throughout the test.

5.3.1 General Voice

This first stage is about characterising perceptively the general voice. The stimulus used is a series of ten 0.5-second-long different vowels, the "sentence" *S1*. The participants have been tasked with characterising *S1* in terms of *identity*, *homogeneity* and *naturalness*. More precisely, they have been estimating the gender, age, stature (height and corpulence), the number of speakers and origin of the audio file (natural, synthetic or natural processed voice). A confidence rating and the report of the (audio) pointers have finally been requested in order to help the respondents' answers. The independent variables are the utterances characteristics, namely its nature, f_0 and duration. The dependent variables are the perceived features relative to quality, identity and homogeneity of the voice synthesised, and the related confidence ratings. Simple statistics are applied to the data, and redundant and relevant remarks made by the participants are extracted.

5.3.2 Vocalic Accuracy

The second stage aims at evaluating the synthesis accuracy, or in other words at determining whether the vowels synthesised were perceptively recognisable or not. The stimuli used (S2) are 10 separated 0.5-second-long different vowels. The total duration of this stimulus ($\approx$ 7 sec) aims at providing the participants with a stimulus long enough to create a personal image of the speaker, especially of its age [Schötz, 2005]. The participants have been asked to identify the vowels synthesised *S2* by selecting the word in a list (in French) that is closest in pronunciation (see Table 4.1). The list and the equivalence between the French and English pronunciation had been established based on an online IPA chart. The independent variables are the nature of the sound uttered. The dependent variables are the perceived phonetics and the related confidence ratings. The responded vowel attribution is compared to the expected vowel attribution, and the existence of clusters - grouping of vowels that seem to share a similar timbre - is considered.

5.3.3 Ageing Perception

The third and last stage addresses age estimation and aims at answering the two following questions. Does a sample made to sound as a 70-year-old person sound as such? Are there predominant features? Two types of stimuli have been used at this point. The stimuli *S3* are altered versions of *S1*: they contain all ageing effects at ages 25, 55, and 80.This age selection has been chosen to cover lifetime and to enable the presentation of multiple modelling representations. For this purpose and in accordance with the pre-selection made in 5.4, three stimuli have used the age-f0 relation from [Makiyama and Hirano, 2017] and three have used the one from [Brown et al., 1991]. The stimuli *S4* have the same basis as *S1* except that they are applied only one aspect of ageing among f_0, speaking rate, vibrato and tremolo at three different amplitudes. The stimuli have been randomly presented in every question. The respondents have been invited to estimate the age of each stimulus *S3* independently (part 1) and *S4* relatively (part 2), within 20-100 years old (y.o.). As such, the independent variables are the ageing effects previously described in 2.4 that define the modelled age (MA). The corresponding dependent variables are the perceived age (PA) and the related confidence ratings. In part 1, simple statistics are calculated and used to compare the two age-f_0 relations against each other and against the neutral track *S1*. Additionally, the signed and absolute differences between PA and MA are calculated. In part 2, simple statistics are computed per group of feature and compared against the others.

5.4 Pre-Tuning

Due to the large amount of parameters, some pre-tuning has been deemed necessary in order to diminish the manageability complexity of the model and to provide enlightening about the validity range of some parameters and metrics.

5.4.1 Parameters

Some parameters have been selected for this purpose. Evaluation at this stage has been done through attentive listening and the computation of adapted metrics when relevant, on all vowels available and considering only the three first formants.

1) Damping coefficient Several damping coefficients in [0.990-0.999] have been tested and analysed perceptively and analytically.

2) Reflection at the lips Even if [Airaksinen et al., 2014] recommends to set r_{lp} in between $[.98, .999]$, values in [0.80,0.99] have been tested perceptively and analytically.

3) Excitation method & Trachea consideration The three analytical GPs introduced in Ssec. 4.1.1 have been compared perceptively for several OCR and $tp = .40T$, with an equal parameterisation otherwise. Note that only one OCR is available for gp-klglott88 once t_p is fixed. Additionally, the impact of the trachea in associated with different excitation methods has been analysed perceptively.

4) VTS obtention method Vocal tracts obtained with MRI and PCA have been compared perceptively and analytically, with an equal parameterisation otherwise.

5) Noise addition The impact of noise addition as implemented in 4.1.1 has been perceptually evaluated.

6) Listener position & Loudness The position of the listening position has been tested out to verify the basic expectation according which intensity decreases along the VT from the glottis to the lips.

7) Time envelope The temporal envelope of the signal has been worked on, in order to attenuate some of the audible artefacts. The three envelope methods mentioned in 2.3 have been tested, with the sinus-based one being used as baseline. The hybrid method has been tuned in terms of patch duration.

5.4.2 Metrics

The two vibrato functions have been tried out, and the interdependence between the metrics *vibrato*, *tremolo* and *vocal noise* has been tested. These three effects had been implemented under a strong and debatable hypothesis: **H1** The system is transparent (i.e., given an expected effect, the measured effect is the same). Such an hypothesis gave ground to using the results available in the literature (the values of the metrics associated to these effects) to set their parameterised implementations (see 4.1). However, **H1** is known to be false, since the vibrato, tremolo and noise are interdependent from each other [Awan, 2006]. A 6-way ANOVA has been conducted on each metric relatively to the combined 6 basic parameters controlling the effects: $df0, dh0, snr, f_D, dnutt, n_{marks}$. The main trends have been observed.

Chapter 6

Results

The results for every aspect listed in Chapter 5 are presented here.

6.1 Pre-Tuning

The observations and decisions taken about specific parameters of the model are presented in this section.

6.1.1 Parameters

1) Damping coefficient Most results sounded relevant; however, for too high damping coefficients ($d \geq 0.997$), artefacts were clearly audible under the form of high frequencies, and observable on the spectrogram as more concentrated energy lines and on the F1-F2 space (changing shape). After testing, a standard value of 0.995 has been judged adequate. Global error comparison gives $E = 179$ Hz for $d = 0.999$ against $E_{df} = 39$ Hz for $d = 0.995$.

2) Reflection at the lips In terms of perception, one can see in Table 6.1 that any value for r_{lp} below .95 sounded acceptable. However, according to metrics displayed here, namely error E and qVSA, values up to 0.90 are acceptable. This observation shows that these two metrics are not embracing sound quality, especially in terms of noise. A value of 0.99 was finally set for r_{lp}.

r_{lp}	.999	0.99	0.95	0.90	0.85	0.80
Error(Hz)	39.7	40	40	42	45	123
qVSA(kHz^2)	.304	.304	.308	.308	.379	.380
Perception	ok	ok	ok	b.	b.	bb.

Table 6.1: r_{lp} fine-tuning. Perception is done through listening; b.: "buzzy", bb.: very "buzzy", ok: of acceptable naturalness.

3) Excitation method & Trachea consideration For $OCR \leq 1$, both functions have produced sounds of debatable quality, with an effect of instability as when voice breaks. For $OCR > 1$, functions take different directions. Large OCRs smooth vowels for gp-rosenB but introduce high frequencies for gp-fant79. All results are summarised in Table 6.2. Furthermore, note that the gp-klglott8 method is particularly slow due to the call of the Matlab function

OCR	≤ 1	1	1.50	1.66	2.24
gp-fant79	br.	ok	ok	HF	HF
gp-klglott88	/	/	b.	/	/
gp-rosenB	br.	ok	ok	ok	ok

Table 6.2: Glottal pulse fine-tuning at $tp = .40T$. Perception is done through listening; b.: "buzzy" voice, br.: breaking voice effect, HF: presence of high frequencies, /: not available.

"*designfilt*". For easiness in control of OCR and acceptable perceived sound quality, gp-rosenB has finally been selected with $OCR = 1.5$.

In practice, although some changes are observable in the spectrum, considering the trachea doesn't yield any audible difference when analytical functions are used. However, it does when another excitation method, which is largely inspired from LeT 'calcflow' function, is employed. This latter method offers the benefit to *couple resonator and excitation* for each sample, instead of looking into a data base for the next glottal sample as is the case for analytical functions. Resulting sounds are conclusive with this method when the trachea is activated; otherwise, they sound "buzzyer". Nonetheless, in both cases, some vowels don't get detected and error is generally higher than with the former method ($E > 60$Hz vs $E \approx 40$Hz).

4) VTS obtention method The vts-mri and vts-pca methods have produced very distinguishable sounds; not in terms of quality, but rather in terms of timbre. vts-mri yielded vowels that sounded nasal. Metrics have acknowledged this differenciation, as shown in Table 6.3. This "nasalisation" may be due to the sharp angles occurring in a natural VT, compared to the vts-pca which is made of the bones of vts-mri. In other words, vts-pca has removed the detail

VTS method	**vts-mri**	**vts-pca**
Error(Hz)	55	39
qVSA(kHz^2)	.466	.303
Perception	nasal	ok

Table 6.3: Vocal tract fine-tuning. Perception is done through listening; ok: of acceptable naturalness.

- or high frequencies - of the vts-mri. Both VTS were kept in the interface as playable parameters, so that the user could realise that approximation (understand, vts-pca) is sometimes more enjoyable - if not realistic - than reality itself.

5) Noise addition Noise addition, with the values announced in 4.1.1, didn't yield significant audible changes.

6) Listener Position & Loudness The expected decrease in intensity is illustrated for all vowels on Fig. 6.1. It means that a sound that is recorded right in the middle of the VT would already be identifiable, though, not as well as at the lips (it may merely be a question of habit). The fact is that the largest deformations in the VT occur at the remote extremity of the VT, close to the lips. Especially sampling the vocal signal before the mouth region removes a lot of information. figure In addition, the differentiated loudness caused by different vowels has turned out to completely dominate the loudness effect related to ageing. Out of audible clearness, the loudness ageing effect has finally been let aside and the sentence's amplitude has been normalised.

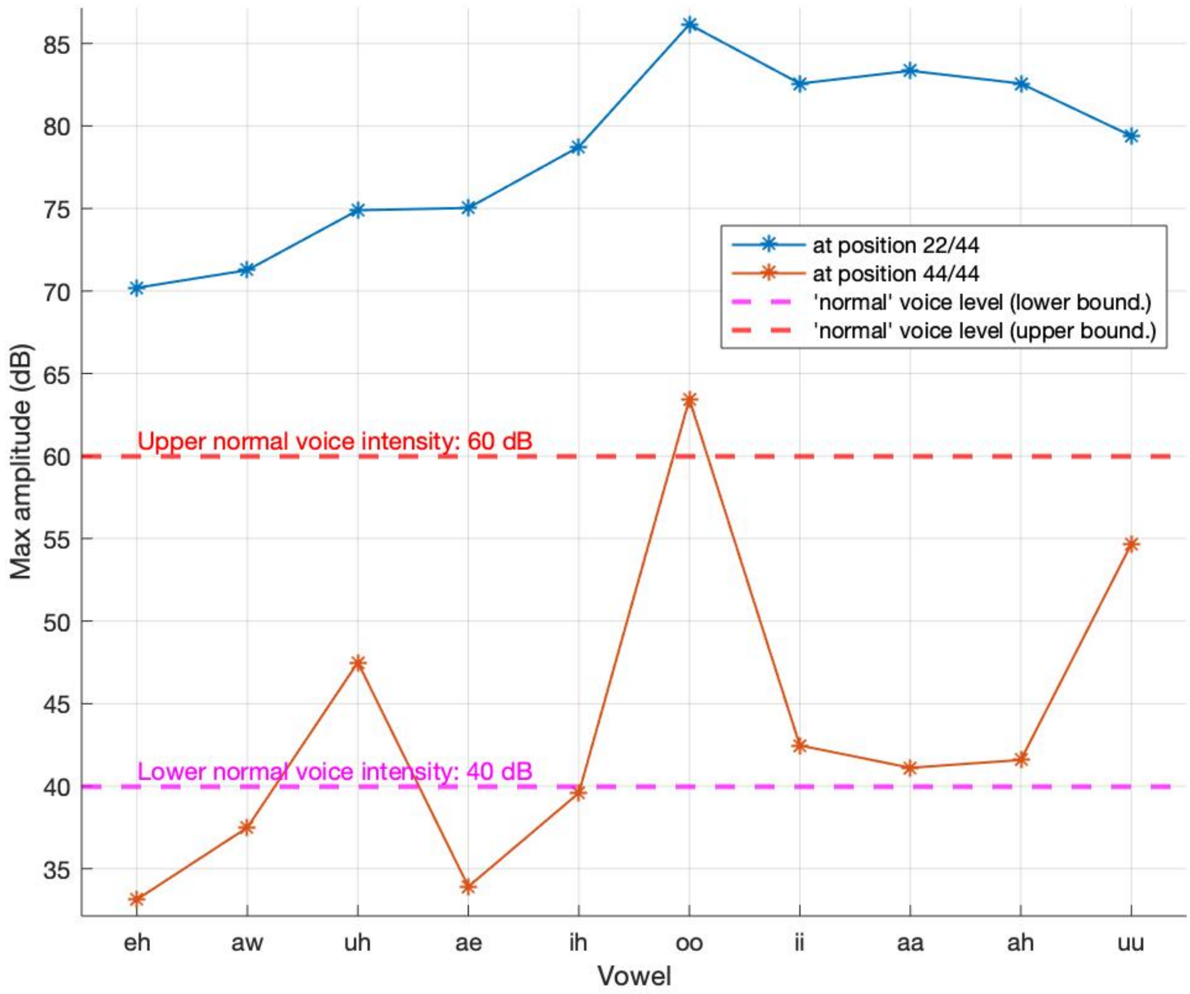

Figure 6.1: Loudness evolution along the vocal tract. Dash-lines: expected delimitation of loudness for a "normal" voice at the lips. Continuous lines: listeners positioned in the middle of the VT (in blue) and at the lips (in orange).

7) Time envelope After testing values in [4-10]% of t_{utt}, it has been set at 7% as a compromise between too smooth an envelope (higher values) and the creation of additional audible artefacts (lower values). Overall, even though the natural and hybrid envelopes have made the sounds less artificial and have managed to attenuate the buzzing artefacts, they have added artefacts of another kind, namely high frequencies. After comparison, the sinus-based envelope method has been judged better.

6.1.2 Metrics

The Matlab function "*comm.FMModulator*" proved to alter gravely the very nature of the speech signal and has thus been abandoned.

Shimmer Shimmer correlates significantly with f_D (p< .001), n_{marks} (p< .001) and snr (p< .001), and non-significantly with the others. Among the significative interactions between parameters, those including f_D and n_{marks} are dominant.

Jitter Jitter correlates significantly with f_D (p< .001), $dh0$ (p< .01) and snr (p< .001), and non-significantly with the others. Among the significative interactions between parameters, those including f_D and snr are dominant.

HNR HNR correlates significantly with f_D (p< .001), n_{marks} (p< .01), $dh0$ (p< .001) and snr (p< .001), and non-significantly with the others. Among the significative interactions between parameters, those including f_D, $dh0$ and snr are dominant.

It follows that, as expected, the parameters $df0, dh0, snr, f_D, dnutt, n_{marks}$ are not exclusively associated to the metric they have been designed for, since they impact the other metrics too. Consequently, the values of the 6 parameters have been adapted manually (scaling or summing additional terms) to yield acceptable metric values.

6.2 Vocalic Synthesis Acoustics

The three formant detection methods have been applied to S_1. The LPC-based methods seem similar in accuracy (see Table 6.4) while and generally better for low frequencies.

Metric	**myLPC**	**findpks**	**Praat**
n_{off}	0	2	0
E_{df} (Hz)	86	195	78
E_{qVSA} (*10^3 Hz^2)	-117	251	0.6
E_{tVSA} (*10^3 Hz^2)	-71	-240	35

Table 6.4: Acoustic Analysis for the sentence S_1. Various metrics (see Ch.5) are displayed for three analysis methods, relatively to [Hillenbrand et al., 1995].

$F_1 - F_2$ Space Fig. 6.2 features the $F_1 - F_2$ space, in which the quadratic vocal space, its area and vowels coordinates from [Hillenbrand et al., 1995] (in red) and from the three analysis techniques used are represented. In the case of '*findpks*', it can be read that vowels were not all faithfully reproduced. For instance /o/: F_1 has totally been missed during detection - hence its position in the far top right angle of the $F_1 - F_2$ space instead of the bottom left angle. The same happened for vowels /i/ and /u/. Consequently, both the quadrant itself and its area are impacted and the latter is unemployable for compared analysis. Based on the formants values, an identification of the vowels is made; in the case displayed here, only 5 vowels were formally recognised by the software: /ɛ/, /ɪ/, /ʌ/, /ɔ/ and /ʊ/. It proves how unreliable '*findpks*' can be. For '*myLPC*', the general shape of the quadrant is approximately maintained even so its gravity centre has moved a certain distance. This method generally underestimates all formants frequencies, resulting in a reduced qVSA.
The question is now: is it due to the analysis method or to the synthesis method? The method '*Praat*' now serves as baseline. On one hand, the following discussion relies of course on Praat software accuracy. On the other hand, we understand that the point is not to find exactly the formants of the literature but rather to get close enough so that they are *perceptively* relevant. Now that these two points are made, let's compare [Hillenbrand et al., 1995] data with Praat results. Any difference can be interpreted as the action of the system. The synthesiser hence tends to produce formants that spread out on the spectrum along F_1 dimension, and on the contrary that are compressed along F_2 dimension, with a stressed effect on F_1. The resulting qVSA is accordingly bigger - though not dramatically.
When comparing these to the formants estimated by *findpks*, it appears that *findpks*, when not ignorant of some formants, is more accurate than its counterpart '*myLPC*'. This may be due to an insufficiently precise and adapted parameterisation of the LPC-based method. Therefore, the FFT-based method *findpks* deserves to be researched further and been made more accurate.

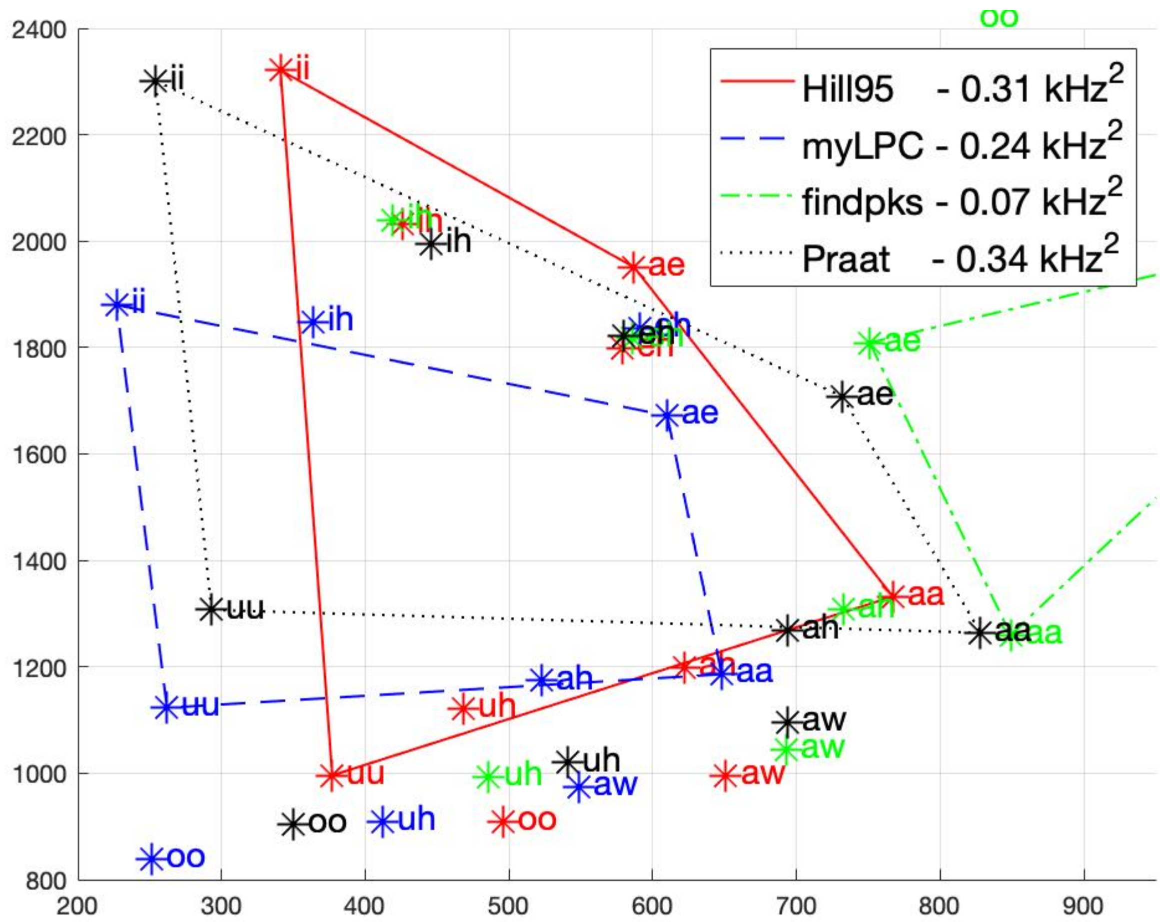

Figure 6.2: F1-F2 space: *Abscissa*: F_1 (Hz), *Ordinate*: F_2 (Hz). Features of VSA and qVSA between [Hillenbrand et al., 1995] (reference) and the 3 formant detection techniques.

Male *vs* Female In average for male, errors on f_1 sum to ≈ 0 while those on f_2 have been either largely underestimated or very similar, as illustrated in Table 6.5 (first row). In comparison, females have been less well detected. It must be taken into account that these figures have been computed for a male and female aged 30, on 9 and 5 vowels respectively, as a consequence of the vowels that have been detected by the algorithm for each gender. Still, we can observe similar trends for both genders regarding f_1 and f_2, so to say: slight underestimation for the former and larger overestimation for the latter.

	Formants	F_1	F_2	F_3
Men	**Mean**(Hz)	3.44	-58.4	51.9
	Std(Hz)	69.2	86.9	106
Women	**Mean**(Hz)	24.8	-166	-43.4
	Std(Hz)	62.5	280	176

Table 6.5: Error on Formants reproduction for the 3 first formants and for 9 recognised vowels for male, for 5 recognised vowels for female. To be noticed: reference material comes from [Hillenbrand et al., 1995] where expected amplitudes were not provided.

6.3 Voice Perception

The survey has been completed by 15 French-speaking participants (6 men, 9 women), aged 39 in average (std: 16 y.o.), of average size 1.70 m (std: 7.5 cm) and medium corpulence (60%, thin 20%, overweight 20%). It is to be noted that even though all participants are fluent in French, they may have different accents depending on their region of residence.

6.3.1 Voice Identity

The voice has been predominantly attributed to a man (93%), aged 36 in average (std: 9 y.o) with an average stature both in size (normal 80% *vs* big 20 %) and build (normal 73% *vs* thin 27%). 93% of the participants agree on the voice belonging to a single person while one indicated a mix of 3 different voices. Also, 60% of the participants define the voice as synthetic against 40% who report hearing it as a recorded voice that has been processed. When asked about the clues that helped them respond, the participants have mentioned the pitch/timbre (60%) and the intonation (stable, clear and confident voice, 20%). Two persons have not provided any exploitable answer while one has described the voice as "robotic". The fact that the "natural voice" option has been absolutely excluded from the participants selection highlights a lack of quality in the synthesis. However, this aspect doesn't seem to prevent the voice to be representative from a human being. Particularly, age estimation reveals a high consistency between participants about the age of the speaker, since AE is expected to vary within an interval of ± 10 years relative to the chronological age of a real speaker [Schötz, 2005]. It seems that this voice may be related Interestingly, such a precision in AE combined with the fact that the participants' average age is 39 may be interpreted as an (unwanted) illustration of the own-age bias; in which case, it would mean that this sample was particularly representative of a certain age category. Another interpretation is also possible: this voice sample is "age neutral", in the sense that it could be attributed to a speaker of any age - and the easiest answer to give is then one's own age. These two options will be kept in mind when analysing the other parts of the perceptive test. Combined with the resulting perceived stature information, this voice seems however to possess a certain *identity* even though its quality is mediocre.

6.3.2 Vowel Reproduction Accuracy

The table 6.6 reports the attribution of the sounds produced by the model (left) to the selected list of French words (top). The phonetic elements that are starred is recognised by the system during analysis. It is not exactly a confusion matrix, for the relation between the English phonetics and the French words is not one-to-one. For example, although /ε/ is common to both languages, U is not. In the observations and analysis to come, the aim is to measure the capability of the system to produce the sounds parameterised by the user. It is vital to realise that various sources of error can occur at this stage so they can be identified: error during synthesis, error during analysis, phonetic translation inaccuracy, language accent and human factor errors.

Analysis-based error One can observe a good match for some of the vowels shared by both languages: /ɪ/ (93%), /ε/ (92%), /i/ (100%), even though the sound /i/ has not been formerly identified during the analysis phase. The error here stems from analysis. The situation of /æ/ is somehow different. Indeed, when listening to its international pronunciation, it seems in-between /ε/ and /ɑ/.

Synthesis-based error The vowel /u/ has been associated to the word "tu", whose proper pronunciation in the international phonetic alphabet (IPA) is /y/. Although it seems the formants were accurately produced relatively to [Hillenbrand et al., 1995], some unwanted elements of the system designed may have interacted with the sound and amplified perceptively its higher harmonics.

Perception-based approximation In spite of that, when looking at the F_1-F_2 space, it appears closer to /ε/ (almost superimposed). Let's come back to the results: for French listeners, this occurrence of /æ/ definitely sounded like /ε/. However, in this situation precisely it is difficult to attribute the error with certainty (synthesis, English phonetics or phonetic transcription).
Two pairs of words were systematically brought together: {sot, sort} and {patte, pâte} corresponding to two "clusters" of vowels: {/o/, ʊ} and {/ɑ/, ʌ} respectively. These blurred boundaries between words highlight similar pronunciations in French, which can be explained by the proximity of the vowels in the F_1-F_2 space (the proximity in question may be due to synthesis approximation).

	Word identified by the listeners									
Intended vowels	clé	coup	mère	île	sot	sort	ceux	patte	pâte	tu
/ɪ/*	**93**									
/u/		**29**					36			36
/ε/*	8		**92**							
/i/				**100**						
/o/		60			**30**	7				
/ɔ/*		7				**7**	7	7	71	
/ʊ/*		7			33	60	**0**			
/ʌ/*							25	**50**	25	
/æ/			100						**0**	
/ɑ/								64	36	**0**

Table 6.6: Relation matrix adapted to French words. The digits are percentages. They quantify the number of times a given vowel has been attributed a certain word. The phonetic elements that are starred have been recognised by the system during analysis.

The decision to use French as main language all along the test - except for the nature of the vowel synthesised - has proved to be an error on my part. It may have made the test easier for the participants - and I do think the difficulty of the test demanded such a decision; however, it made the analysis and interpretation parts much more challenging.

6.4 Ageing Voice Perception

The observations made on and both parts of the perception-based test and their interpretations are presented here.

6.4.1 Absolute Age Perception

The aged stimuli S_3 and the neutral track S_1 have been evaluated in age independently from each other. S_1 yielded AE close to the former one: 37 ± 11 y.o., which seems to validate the first hypothesis about this stimulus: this sample is particularly representative of a certain age category. The raw data for S_3, averaged over participants, is available in Table 6.8. It is essential to notice that the standard deviation (std) of age on any sample is especially large and prevents the mean ages to differ significantly (the age intervals largely overlap). This may be due to too small a testing pool. However, this is also consistent with the low global confidence rated on all six estimations ($18 \pm 14\%$), and with some remarks of the respondents: some have mentioned the feeling to listen to the same person over and over, without clear distinction (26%). When looking at the individual responses of the participants, it appears that two samples could often not be discriminated against each other (but not always the same pair). Furthermore, all age evolution trends have been observed across the participants. As such, it seems that the timbre of the voice and the defective quality coupled with a rather individualised perception of ageing have hindered all interpretable ageing effect. However, an interesting fact relates to f_0. Recall

Expected age (y.o.)	80	55	25
mean f_0 (Hz) - **Kasuya et al.**	150	128	125
mean f_0 (Hz) - **Brown**	145	120	125

Table 6.7: Mean fundamental frequencies of the two voice sentences used in the Ageing Perception Test - Part 1.

that f_0 is the only ageing parameter whose evolution is not monotonous with age (refer to Fig. 4.1 and see the Table 6.7 reporting the mean $f_0 used here$). in addition, it is supposed to present the largest changes in this case since the Kasuya and Brown methods differ solely by their age-f0 relation.On one hand, stds are similar for a given expected age; on the other hand, a monotonous evolution of f_0 yields a non-monotonous evolution of AE (Kasuya) and conversely (Brown). These observations may infer that f_0 impact cannot be understood all alone and presents some interaction with the other ageing parameters.

	Age: mean ± std (in years)		
Expected age	80 ± 10	55 ± 10	25 ± 10
Kasuya et al.	50 ± 21	40 ± 16	44 ± 18
Brown	49 ± 23	55 ± 17	41 ± 17
Avg (estim)	49 ± 22	48 ± 18	42 ± 17

Table 6.8: Age estimation (mean and std). Stimuli: twice 3 samples have been produced at ages 80, 55, 25 with the age-$mvf0$ relations of [Makiyama and Hirano, 2017] and [Brown et al., 1991].

6.4.2 Feature Impact Identification

Four groups of three samples have been produced on the following model. A fixed-age basis has been altered by a certain feature ($mvf0$, speaking rate, tremolo or vibrato) at three different levels. As a general observation, it can be seen that the stds almost homogeneously increase with the amplitude of the effects, and that all parameters do not cause the same average ageing effect: f_0 yielded an average AE of 42 y.o. *vs* 53 y.o. for $dfvar$. This is connected to the fact that most of these features appear passed the pivot-age. The AE about f_0 confirms the observations made right above: a monotonous evolution of f_0 does not imply the same monotonous evolution of age. As such, it is not sufficient a feature to perceptively estimate a voice accurately. At fixed tremolo amplitude, an increasing $df0$ means a faster tremolo. As

	Sample 1	Sample 2	Sample 3
f_0 (Hz)	130	138	145
AE (y.o.)	40 ± 14	47 ± 16	40 ± 24
$df0$ (%)	4	15	28
AE (y.o.)	44 ± 16	48 ± 17	56 ± 21
t_{utt} (s)	0.5	0.6	0.7
AE (y.o.)	33 ± 9	47 ± 16	58 ± 22
$dfvar$ (%)	1	10	15
AE (y.o.)	42 ± 15	48 ± 13	70 ± 19

Table 6.9: Age estimation (mean and std). Stimuli: groups of 3 samples produced with features $mvf0$, tremolo, speaking rate, and vibrato (in this order)

expected, such trend has been perceived as synonym of ageing in average (even though the intervals overlap due to large stds). We observe a monotonous evolution of t_{utt} as a function of age. However, this progression of std is the most substantial one. At $t_{utt} = 0.5$ seconds, the std matches the expectations for VBAE (within ± 10 y.o.); at higher t_{utt}, AE is similar to those of the other parameters. Although we could have thought that longer stimuli would provide the listeners with more material about the voice, it seems to have prejudiced them instead. A reason may be that the audibility of artefacts increases as time passes and removes part of its humanity to the voice. At fixed vibrato frequency, an increasing $df0$ means deeper variations of frequency (amplitude of the FM). As expected, $dfvar$ increases monotonously as a function of age. Yet, over a certain value, this feature seems to be the most characteristic of old age, as AE for sample 3 seems to prove.

These parameters therefore different effect on age perception, and except for f_0, their interactions are not known yet.

6.5 Ageing Voice Morphing

Ageing voice morphing is about applying the ageing effects implemented in this project to real voice. A short pilot test has been conducted: participants have been asked to listen to a sample of processed voice and to estimate the age of the speaker. Free collection has been collected. Then, the participant has listened to the unprocessed speech signal and has been inquired about its perception of the speaker: identity of the person, Finally, general impressions have been collected. Unfortunately, the successive transformations have introduced their share of artefacts and have prevented the formation of a truly realistic image of the speaker and of their age, and have damaged the expected sensation of identity of the speaker.

Chapter 7

Discussion

An experimental ageing voice model has been designed, documented and evaluated.

A fixed-age voice model combining both physical- and analysis-based elements has been implemented. Although it is relatively human sounding, it lacks the quality of a natural voice: audible artefacts such as high frequencies and buzzing sound disrupt the voice perception. The aforesaid quality is controllable to a certain extent through the parameterisation of the model However, the speaker created is perceived as being a single male human being aged ≈ 36 years old. Furthermore, the vowels he says are mostly recognisable and classified as existing sounds.

A framework to support the voice synthesis model has been designed. Voice synthesis can be realised through three tools coded in Matlab: a script, an interface and a real-time plugin. The script is "low-level" and needs some taking-in-hand, but benefits from large possibilities in terms of parameterisation, including the design of the model or analysis and visualisation settings. The interface and the plugin are high level and benefit from a graphic interface. The parameterisation of the interface is intermediary between the ones of the script and of the plugin. Visualisation is provided for all formats to convey the form of the data used and/or the results of the analysis. Part of the data (for instance the age-f_0 relation or the two types of VTS) can also be visualised outside of the framework. The analysis method is two-fold by using both a LPC- and a FFT-based functions, the purpose being to cross the results and strengthen the analysis capabilities of the framework. The final state of the system can be logged for future use and/or analysis. When it comes to optimisation, all code formats have been profiled using the Matlab profiler. Some experimentation on Matlab executable functions (.mex) has been conducted on the propagation loop, without real necessity however: the principal cause of latency in this program is the model initialisation, not propagation. As proof, the real-time plugin runs without difficulty (i.e., without the creation of audible artefacts).

As an optional additional unit to the fixed-age model, ageing effects can be considered separately (script, interface, plugin) or as a whole via an ageing mode (script, interface). This ageing mode works differently in the script and the interface. In the interface, it is applied to a single vowel and several occurrences of this vowel at different ages are concatenated to provide a direct hindsight over the ageing effect. In the script, it takes the whole sentence of vowels, calculates and applies the corresponding ageing effect for a user-defined age. Further tests on ageing have been experimented on real voice, with the purpose to *age-morph* a real speaker. They have not proved particularly conclusive, for the audio thus processed contains audible artefacts.

Chapter 8

Conclusion

After having brought basic knowledge about voice, especially its characteristics and production system, physically-inspired and analytical modelling techniques have been introduced in the frame of the source-filter theory. After presentation of two existing voice synthesisers and a brief introduction to voice analysis, the theory underlying these theories has been detailed. The data necessary to achieve a simple voice synthesiser has then been referenced, along with ageing-related and voice analysis data. Afterwards, the framework supporting this project has been introduced, and all steps have been explained - from parameterisation to post-processing and logging via model creating and processing. The analysis of the model has been undertaken with a tool specifically designed besides the synthesis model, and complementary verification on Praat has been made. The metrics taken into consideration in evaluation include the vocal space area and an error calculated on the distance between formants frequencies.

In practice, in order to create the model, diverse data bases have been created to gather glottal pulses, the mean vocal fundamental frequencies as functions of age from various authors, and two types of vocal tracts obtained through different procedures. The excitation and boundary conditions have been implemented in an analytical fashion while the resonator is fully physical. The resulting model can be qualified of physical-inspired model. A pre-tuning of a selection of parameters has clarified the impact of some parameters and simplified the model; an analysis tool has been developed and special care has been taken for rendering visually results legible and clear, especially through a comprehensive framework including an interface, a real-time plugin and good management of the data outputted from the system. Ageing has been integrated through the parameter-dependent features tremolo, vibrato, vocal noise, speaking rate, fundamental frequency as functions of age. A procedure to apply it to real voice have also been developed. The ageing effect on synthetic voice has been evaluated perceptually. Even though differences are audible especially between young age and old age, they remain slight and dissimulated by omnipresent artefacts such as whistling and buzzing sounds.

Nevertheless, this implementation sets the ground rules for further development, especially regarding formants detection, voice quality control and ageing effect credibility. A software that would fulfil all these criteria would open the door to a large public by triggering interest in individuals for their personal leisure and in audiovisual media for manipulating the voice of their actors at lower costs. In the case of a particularly accurate age-morphing software, it could also find an application in medicine as a ”voice prosthesis” for severe age-caused speech-disabled people.

Bibliography

[Airaksinen et al., 2014] Airaksinen, M., Bäckström, T., and Alku, P. (2014). Automatic estimation of the lip radiation effect in glottal inverse filtering. In *Proc. Annual Conf. Intl. Speech Communication Assoc. (INTERSPEECH)*, pages 398,402. International Speech and Communication Association.

[Alku et al., 2006] Alku, P., Story, B., and Airas, M. (2006). Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatrica et Logopaedica*, 58(2):102,113.

[Awan, 2006] Awan, S. N. (2006). The aging female voice: Acoustic and respiratory data. *Clinical Linguistics & Phonetics*, 20(2-3):171–180.

[Berisha et al., 2014] Berisha, V., Sandoval, S., Utianski, R., Liss, J., and Spanias, A. (2014). Characterizing the distribution of the quadrilateral vowel space area. *The Journal of the Acoustical Society of America*, 135(1):421–427.

[Bilbao, 2004] Bilbao, S. (2004). *Wave and Scattering Methods for Numerical Simulation.* Wiley.

[Boer and Fitch, 2010] Boer, B. D. and Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, 18(1):36–47.

[Boersma, 2002] Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341,345.

[Brown et al., 1991] Brown, W., Morris, R., Hollien, H., and Howell, E. (1991). Speaking fundamental-frequency characteristics as a function of age and professional singing. *Journal Of Voice*, 5(4):310,315.

[Carré et al., 1992] Carré, R., Chennouk, S., and Mrayati, M. (1992). Vowel-consonant-vowel transitions, analysis and modeling. *The Journal of the Acoustical Society of America*, 92(4):2413–2413.

[chiller, 1978] chiller (1978). *Modern spectrum analysis.* IEEE Press selected reprint series. IEEE Press, New York.

[Cook, 1996] Cook, P. R. (1996). Singing voice synthesis: History, current work, and future directions. *Computer Music J.*, 20:38–46.

[de Boer, 2009] de Boer, C. (2009). Recent Books in the Field of Public Opinion Research. *International Journal of Public Opinion Research*, 21(2):255–260.

[Dudley, 1964] Dudley, H. (1964). Thirty years of vocoder research. *The Journal of the Acoustical Society of America*, 36(5):1021–1021.

[Dunn, 1961] Dunn, H. K. (1961). Methods of measuring vowel formant bandwidths. *The Journal of the Acoustical Society of America*, 33(12):1737–1746.

[Eichhorn et al., 2017] Eichhorn, J., Kent, R., Austin, D., and Vorperian, H. (2017). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice*, 32.

[Fant, 1960] Fant (1960). *Acoustic theory of speech production : with calculations based on X-ray studies of Russian articulations /*. Description and analysis of contemporary standard Russian ; 2. Mouton, The Hague, Netherlands.

[Fant, 1972] Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 2/3:28–52.

[Fant et al., 1985] Fant, G., Liljencrants, J., and Lin, Q.-G. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 4(Oct-Dec):1,27.

[Ferrand, 2002] Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, 16(4):480,487.

[Garellek et al., 2016] Garellek, M., Samlan, R., Gerratt, B. R., and Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, 139(3):1404–1410.

[Gray and Markel, 1976] Gray, A. H. and Markel, J. D. (1976). *Linear prediction of speech / J. D. Markel, A. H. Gray, Jr.* Springer-Verlag, Berlin.

[Harnsberger et al., 2008] Harnsberger, J. D., Shrivastav, R., Brown, W., Rothman, H., and Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, 22(1):58,69.

[Hawks JW, 1995] Hawks JW, M. J. (1995). A formant bandwidth estimation procedure for vowel synthesis. *The Journal of the Acoustical Society of America*.

[Hézard, 2013] Hézard, T. (2013). *Voice production : exploration, models and analysis/synthesis*. Theses, Université Pierre et Marie Curie - Paris VI.

[Hillenbrand et al., 1995] Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America*, 97(5):3099,3111.

[Hirano, 1974] Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica et Logopaedica*, 26(2):89–94.

[Hirano et al., 1975] Hirano, M., Koike, Y., Mihashi, S., Kasuya, T., and Okada, M. (1975). Structure of the vocal cord as a vibrator. *The Journal of the Acoustical Society of America*, 58(S1):S13,S13.

[Ishizaka and Flanagan, 1972] Ishizaka, K. and Flanagan, J. L. (1972). Acoustic properties of a two-mass model of the vocal cords. *The Journal of the Acoustical Society of America*, 51(1A):91,91.

[Kelly and Lochbaum, 1962] Kelly, J. L. and Lochbaum, C. C. (1962). Speech synthesis. page p. Paper F7.

[Kent and Vorperian, 2018] Kent, R. D. and Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, 74:74 – 97.

[Krauss et al., 2002] Krauss, R. M., Freyberg, R., and Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, 38(6):618,625.

[Kreiman et al., 2014] Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, 1(1).

[Kuwabara and Ohgushis, 1987] Kuwabara, H. and Ohgushis, K. (1987). Role of formant frequencies and bandwidths in speaker perception. *Electronics and Communications in Japan (Part I: Communications)*, 70(9):11,21.

[Linville, 1996] Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, 10(2):190 – 200.

[Lu and Smith, 2000] Lu, H.-L. and Smith, J. O. (2000). Glottal source modeling for singing voice synthesis. In *ICMC*.

[Maeda, 1982] Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1:199–229.

[Makiyama and Hirano, 2017] Makiyama, K. E. and Hirano, S. E. (2017). *Aging Voice*. Singapore: Springer Singapore.

[Mithen,] Mithen, S. *The singing neanderthals : the origins of music, language, mind, and body.* Harvard University Press, Cambridge, Mass.

[Mokhlesin et al., 2017] Mokhlesin, M., Ahmadizadeh, Z., Kasbi, F., and Ahadi, H. (2017). Effect of age, gender and task on speech rate of farsi speakers. *Majallah-i 'ilmi-i danishgah-i 'ulum-i pizishki-i Simnan*, 19(2):327–332.

[Moyse, 2014] Moyse, E. (2014). Age estimation from faces and voices: A review.(invited review)(report). *Psychologica Beligica*, 54(3):255–265.

[Moyse et al., 2014] Moyse, E., Beaufort, A., and Brédart, S. (2014). Evidence for an own-age bias in age estimation from voices in older persons. *European Journal of Ageing*, 11(3):241–247.

[Nielsen et al., 2017] Nielsen, J. K., Jensen, T. L., Jensen, J. R., Christensen, M. G., and Jensen, S. H. (2017). Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient. *Signal Processing*, 135:188–197.

[Oppenheim and Schafer, 1998] Oppenheim, A. V. and Schafer, R. W. (1998). *Discrete-Time Signal Processing*. Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition.

[Pépiot, 2014] Pépiot, E. (2014). Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. In *Speech Prosody 7*, pages 305–309, Dublin, Ireland.

[Peterson and Barney, 1952] Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175,184.

[Ptacek and Sander, 1966] Ptacek, P. H. and Sander, E. K. (1966). Age recognition from voice. *Journal of speech and hearing research*, 9(2):273,277.

[Rojas et al., 2020] Rojas, S., Kefalianos, E., and Vogel, A. (2020). How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of speech, language, and hearing research*, 63(2):533–551.

[Rupal and Seth, 2017] Rupal, P. and Seth, M. G. (2017). Aging a text-to-speech voice.

[Samlan and Story, 2011] Samlan, R. A. and Story, B. H. (2011). Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling.(report). *Journal of Speech, Language, and Hearing Research*, 54(5):1267–1283.

[Sapir et al., 2010] Sapir, S., Ramig, L., Spielman, J., and Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech: Research note. *Journal of Speech, Language and Hearing Research (Online)*, 53(1):114–125.

[Schotz, 2006] Schotz, S. (2006). *Perception, Analysis and Synthesis of Speaker Age*. PhD thesis, Phonetics. Defence details Date: 2006-12-02 Time: 13:15 Place: Hörsalen, Humanisthuset, Språk-och Litteraturcentrum, Helgonabacken 12, Lund External reviewer(s) Name: Möbius, Bernd Title: Associate Professor Affiliation: Institut für Maschinelle Sprachverarbeitung, Stuttgart University — The information about affiliations in this record was updated in December 2015. The record was previously connected to the following departments: Linguistics and Phonetics (015010003).

[Schötz, 2005] Schötz, S. (2005). Effects of stimulus duration and type on perception of female and male speaker age.

[Skoog Waller and Eriksson, 2016] Skoog Waller, S. and Eriksson, M. (2016). Vocal age disguise: The role of fundamental frequency and speech rate and its perceived effects. *Frontiers in Psychology*, 7:1814.

[Speed et al., 2013] Speed, M., Murphy, D. T., and Howard, D. M. (2013). Three-dimensional digital waveguide mesh simulation of cylindrical vocal tract analogs. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):449,455.

[Stathopoulos et al., 2011] Stathopoulos, E. T., Huber, J. E., and Sussman, J. E. (2011). Changes in acoustic characteristics of the voice across the life span: measures from individuals 4-93 years of age.(report). *Journal of Speech, Language, and Hearing Research*, 54(4):1011–1021.

[Stebbins, 1983] Stebbins, W. (1983). *The Acoustic Sense of Animals*. Harvard University Press.

[Story, 2005] Story, B. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation. *The Journal of the Acoustical Society of America*, 117:3231–54.

[Story, 2013] Story, B. H. (2013).

[Story and Bunton, 2017] Story, B. H. and Bunton, K. (2017). Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):EL458–EL464.

[Story and Titze, 1995] Story, B. H. and Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 97(2):1249,1260.

[Story et al., 1996] Story, B. H., Titze, I. R., and Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *J. Acoustic Soc. America*.

[Story et al., 2018] Story, B. H., Vorperian, H. K., Bunton, K., and Durtschi, R. B. (2018). An age-dependent vocal tract model for males and females based on anatomic measurements. *J. Acoustical Soc. America*, 143:3079–3102.

[Sulter and Wit, 1996] Sulter, A. M. and Wit, H. P. (1996). Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. *The Journal of the Acoustical Society of America*, 100(5):3360,3373.

[Teixeira et al., 2013] Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal acoustic analysis – jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112,1122.

[Thapen, 2017] Thapen, N. (2017). Pink trombone.

[Titze and Alipour, 2006] Titze, I. and Alipour, F. (2006). *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech.

[Vallabha and Tuller, 2002] Vallabha, G. K. and Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech Communication*, 38(1):141 – 160.

[Vestlund, 2004] Vestlund, J. (2004). Åldersbedömning av ansikten – precision och ålderseffekter (estimation of age from faces – precision and age effects).

[Vinceslas, 2011] Vinceslas, L. (2011). Formant bandwidth and resilience of speech to noise. Internship report, IRCAM, Paris, France.

[Välimäki and Karjalainen, 1994] Välimäki, V. and Karjalainen, M. (1994). Improving the Kelly-Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques. In *Proc. Intl. Conf. Spoken Language Processing*, pages 615–618.

[Wilkinson, 2016] Wilkinson, W. (2016). A synthesis model for mammalian vocalization sound effects.

[Xi and Longest, 2009] Xi, J. and Longest, P. W. (2009). Characterization of submicrometer aerosol deposition in extrathoracic airways during nasal exhalation. *Aerosol Science and Technology*, 43(8):808,827.

[Yasojima et al., 2006] Yasojima, O., Takahashi, Y., and Tohyama, M. (2006). Resonant bandwidth estimation of vowels using clustered-line spectrum modeling for pressure speech waveforms. pages 589 – 593.

[Yegnanarayana and Veldhuis, 1998] Yegnanarayana, B. and Veldhuis, R. (1998). Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6(4):313–327.

[Zhang, 2016] Zhang, Z. (2016). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of the Acoustical Society of America*, 139(4):1493,1493.

Appendix A

Acronyms and Notations

Here below are the notations and acronyms used through this paper. This is meant to help the reader differentiate some very similar appellations that may confuse him.

A.1 Acronyms

:

- ϕM: Physical Model
- **3MM**: Three-mass model
- **AE**: Age Estimation
- **BPW**: Backward-propagating wave
- **CA**: Chronological Age
- **FAM**: Fixed-Aged Model
- **FFT**: Fast Fourier Transform
- **FPW**:Forward-propagating wave
- **FWF**: Formant wave functions
- **GP**: Glottal pulse
- **HNR**: Harmonic-to-Noise Ratio
- **LPC**: Linear Predictive Coding
- **LeT**: LeTalker
- **MRI**: Magnetic Resonance Imaging
- **OCR**: Opened-to-Closed Ratio
- **PA**: Perceived Age
- **PCA**: Principal Component Analysis
- **PkT**: Pink Trombone

- **Pra**: Praat analysing software
- **SPL**: Sound pressure level
- **VBAE**: Voice Based Age Estimation
- **VPS**: Voice Production System
- **VSA**: Vocal Space Area
- **VT**: Vocal Tract
- **VTS**: Vocal Tract Shape
- **dB**: Decibel
- **mvf0**: mean vowel fundamental frequency
- **qVSA**: quadratic Vocal Space Area
- **tVSA**: triangular Vocal Space Area

A.2 Notations

- $\mathbf{F}_i$: Formant indexed i.
- $\mathbf{H}_i$: Harmonic indexed i. Corresponding frequencies are calculated from f_0 such as $f(H_i) = i * f_0$.
- Nature (Type) of the data: "detected" or "expected".

- $X_{i,det}$: feature X indexed i associated at the type "detected"; $f_{i,exp}$: feature X indexed i associated at the type "expected".
- f_i: frequency associated to $\mathbf{F}_i$.
- df_i: frequency difference between f_i and f_{i+1} of same nature, such as $df_i = f_{i+1} - f_i$.
- Δf_i: frequency error, such as $\Delta f_i = f_{i,det} - f_{i,exp}$.
- Δdf_i: frequency error, such as $\Delta df_i = df_{i,det} - df_{i,exp}$.

- a_i: amplitude associated to formant $\mathbf{F}_i$.
- da_i: frequency difference between a_i and h_{i+1} of same nature, such as $da_i = a_{i+1} - a_i$.
- Δa_i: amplitude error, such as $\Delta a_i = a_{i,det} - a_{i,exp}$.
- Δda_i: frequency error, such as $\Delta da_i = da_{i,det} - da_{i,exp}$.

- h_i: amplitude associated to harmonic $\mathbf{H}_i$.
- dh_i: amplitude difference (dB) between h_i and h_{i+1} of same nature, such as $dh_i = h_{i+1} - h_i$.
- Δh_i: amplitude error, such as $\Delta h_i = h_{i,det} - h_{i,exp}$.

- Δdh_i: frequency error, such as $\Delta dh_i = dh_{i,det} - dh_{i,exp}$.

Convention:

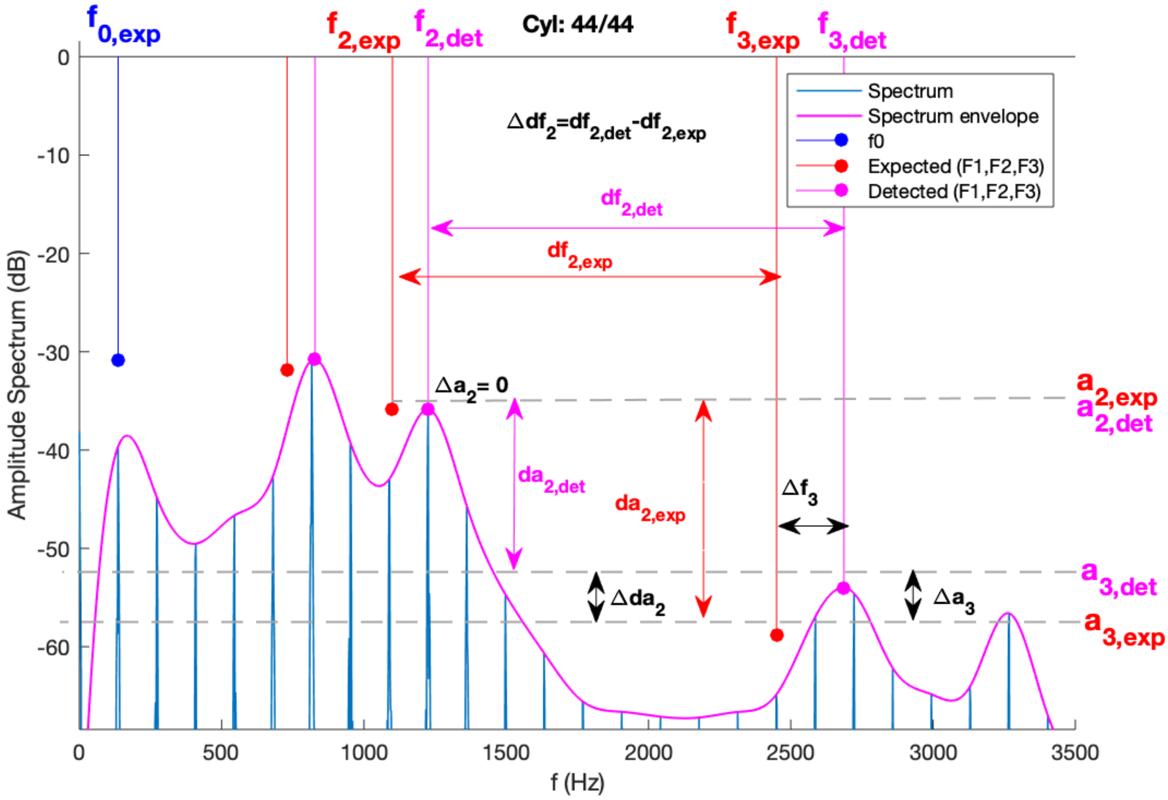

Figure A.1: Convention to characterise formants in terms of frequencies and amplitudes. Example is given on vowel /a/ at the lips (segment 44/44).

Appendix B

User-defined Parameters

B.1 Definitions

Categories:

- [B] Basic human-characterising or user-defined: that contain very common and easy-to-get concepts; e.g. fundamental frequency f_0 (as an approximation of pitch), gender, vowels (see recap Table 4.1).
- [M] Model-related: all necessary parameters that control the excitation or the nature of the signal; e.g. VTS from [Story et al., 1996] or the one from [Story, 2005].
- [E] Effect-related: all optional parameters that affect the resulting speech sample.
- [A] Analysis-related.

[B] Basic parameters (Human-dependent parameters)

- f_0: fundamental frequency - interdependent with 'age'.
- *gender_idx*: index for selecting the gender: male (1) or female (2).
- *age*: in the range [18,90].
- *vowel*: in ['eh', 'aw', 'uh', 'uu', 'ae', 'ih', 'eh', 'ii', 'aa'].
- *which_bcd* (consonants): in ['ll','mm','nn','ng','pp','tt','kk','ss','sh','th','ff',].
- *where_bcd*: (consonants' position) in the "sentence".

[M] Model-related parameters

- *vts_idx*: index designing a set of VTS to build the vocal tract - 1 refers to VTS determined by MRI, 2 refers to modelled VTS after feature reduction.
- $IStrachea$: boolean to select the trachea (1) or not (0).
- *excit_meth*: excitation method - either based directly on the glottal flow , or based on the glottal area which is further processed.
- *ocr*: opened-to-closed ratio. Characterises the duration of opened glottis relatively to duration of closed glottis. Larger OCR produce softer voices.

- *gp_func*: GP function, to select among Rosenberg-B, Fant, and KLGLOTT88 analytical methods.
- *exc*: sub-struct containing parameters to create the GP (gain, proportion duration of opening, proportion time of closing, gain).
- *lips_meth*: choose between constant reflection coefficient at the lips (1), or adaptative (2).
- *f0_author*: to select an author of mvf0s among Brown's, Makiyama's, and Stathopoulos' data.
- *r_lp*: reflection coefficient at the lips.
- *l_e*: length of the vocal tract.
- *len_vf*: length of the vocal folds.
- *th_vf*: vocal folds thickness.
- *damp*: damping coefficient - related to the attenuation of the sound due to wall viscosity e.g.
- *temp_C*: air temperature inside the body - supposedly constant and homogeneous.
- *tcos*: time threshold to initiate the surpression below the glottis.
- *pl*: lungs' pressure or sub-glottal pressure - depending on the consideration of the trachea or not.
- *af*: air flow.
- *t_utter*: duration over which the parameter settings are held constant.

[E] Effect-related parameters

- *ISsnr*: to add noise to the glottis source (1) or not (0).
- *snr*: (in dB) ratio between periodic components and non periodic component (advised to use values in [-40-Inf])
- *ISvib*: to add variation of amplitude (1) or not (0). Impacts the shimmer.
- *df0*: frequency of the modulation, in % relatively to f_0 (advised to use values in [0-.04])
- *dh0*: amplitude of the modulation (advised to use values in [0-1])
- *ISdf0*: to add variation of f_0 (1) or not (0). Impacts both shimmer and jitter.
- *dfvar*: frequency range of variation (in %) (advised to use values in [0-0.1])
- *dnutt*: variation in local utterance length (in samples) (advised to use values in [0-.04])
- $n_s tamps$: number of times the initial value of f_0 is used during the f_0 progression in the utterance (use integers in [1-8], a proper decision also depends on dnutt).

[A] **Analysis-related**: these parameters are located in the "init_params" script.

- *pListen*: location of the listener in the VT.
- $fstop$: arbitrarily, to define the range of frequencies when displaying figures.
- n_fm: number of formants. Shall be ¡ 3.
- *snd2anaOrDisp*: selection of a single sound to be used for analysis when only one is permitted.
- *pos2anaOrDisp*: for the display, position selection among the pListens.

[R] **Running modes**

- *wannaAge*: to activate, deactivate or let be ([ ]) all parameters related to age and to set them according to the only parameter *age*.
- $wannaProfile$: to profile "*process_vt.m*"(1) / everything (2)
- $wannaTest$: to use a test script that loops on user-defined parameters and extracts user-defined metrics. Not recommended to use.
- *all_bool*: to activate (1), deactivate (0) or let be ([ ]) all booleans related to display and logging.
- *all_sense*: to activate (1), deactivate (0) or let be ([ ]) all booleans related to display.
- *all_save*: to activate (1), deactivate (0) or let be ([ ]) all booleans related to logging.
- *save_parts*: to save the resulting speech signal only as a sentence (0) or also as separate files for each vowel (1).
- $fwvn$: "file_write_var_name" to name the resulting audio file with a specific keyword.
- $wannaDefault$: to use the default parameterisation (1).
- *wannaGeom*: to activate (1), deactivate (0) or let be ([ ]) all parameters related to geometry. Default: ([ ]).
- $wannaEffect$: to activate (1), deactivate (0) or let be ([ ]) all effects related to ageing (but without dependence on *age*). Default: ([ ]).
- $wannaDIY$: bool for manipulating a few basic parameters yourself(gender_idx,f_0,gender_idx,vowel) via a GUI
- *wannaPause*: bool for waiting 2 seconds between the apparition of the figures
- *wannaAna*: for analysing the data that was computed. Strongly advised.
- *wannaDB*: bool for storing the audio. Depends on *all_save*
- *wannaLog*: to log the data that was computed. Depends on *all_save*; strongly advised for future checking.
- *wannaDisp*: bool for enabling the display of messages in the command window.

- *wannaSeeAna*: bool for producing the analysis figures
- *wannaSeeData*: bool for producing the figures to observe the raw data used
- $selec\_figs$: list of figures (data, analysis) to be produced.
- $wannaHear$: to listen to the audio produced.
- $wannaTalkSmooth$: to concatenate in an overlapping manner the vowels produced; does not overwrite the original file.

- $IShuman$: to activate (1) "humanising" features (variations in utterance length and in f_0)
- $mono\_f0$: to vary (0) or not (1) f_0 around its average in the sentence.
- $mono\_utt$: to vary (0) or not (1) the utterance length around its average in the sentence.
- $real\_shape$: controls the nature of the temporal envelope: artificial (0), natural (1), hybrid (=natural modelled, 2)
- $pPatch$: active only if $real\_shape$ in hybrid mode. Advised in [4-10]%.

B.2 Default parameterisation

$f0 = [];$
$gender\_idx = 1;$
$age = 30;$
$vowel = \{'eh','aw','uh','ae','ih','eh','oo','ii','aa','ah','uu'\};$
$bcd = \{'ss','mm'\};$
$loc\_bcd = [1, 3];$
$ISnose = 0;$
$vts\_idx = 2;$
$IStrachea = 0;$
$ISnoise = 1;$
$ocr = 1.5;$
$opened - 2 - closedratioexcit\_meth = 1;$
$gp\_func =' pg\_rosenB';$
$exc.Gt = 1;$
$exc.ptp = .40;$
$exc.pte = ocr/(1 + ocr);$
$exc.K = 20;$
$lips\_meth = 1;$
$f0\_author =' brown';$
$r\_lp = -0.99;$
$l\_e = [];$
$l\_vf = [];$
$th\_vf = [];$
$damp = .995;$
$temp\_C = 35;$
$tcos = 0.002;$

$pl = [];$
$af = [];$
$df0 = .05;$
$dh0 = 2;$
$hnr = 10;$
$t\_utter = .5;$
$pListen = [1];$
$fstop = 5000;$
$n\_fm = 3;$
$vow2anaOrDisp = 1;$
$pos2anaOrDisp = length(pListen);$

Appendix C

Voice Data

C.1 About Formants

(in Hz)	f1	f2	f3
'eh'	530	1850	2500
aw'	570	850	2400
uh'	440	1000	2250
'uu'	300	850	2250
'ae'	660	1700	2400
ih'	400	2000	2550
ii'	270	2300	3000
aa'	730	1100	2450

Figure C.1: Values of male formants' frequencies
[Peterson and Barney, 1952]

C.2 Voice modelling

C.2.1 Glottal pulses

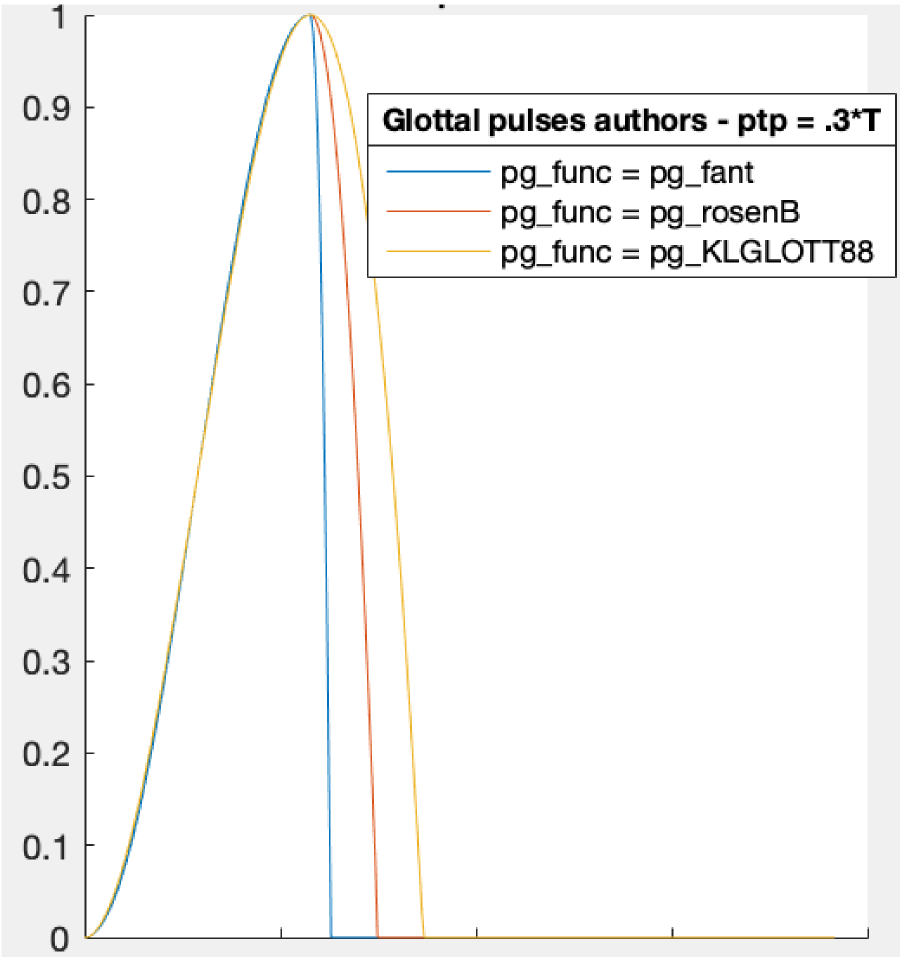

Figure C.2: Comparison of glottal pulses for $ptp = .3 * T$, $pte = 1.3 * ptp$, $K = 30$, $f_0 = 130Hz$, $f_s = 44100Hz$ for three different authors. Abscissa is normalised on T, $G = 1$.

C.2.2 Vocal tract shapes

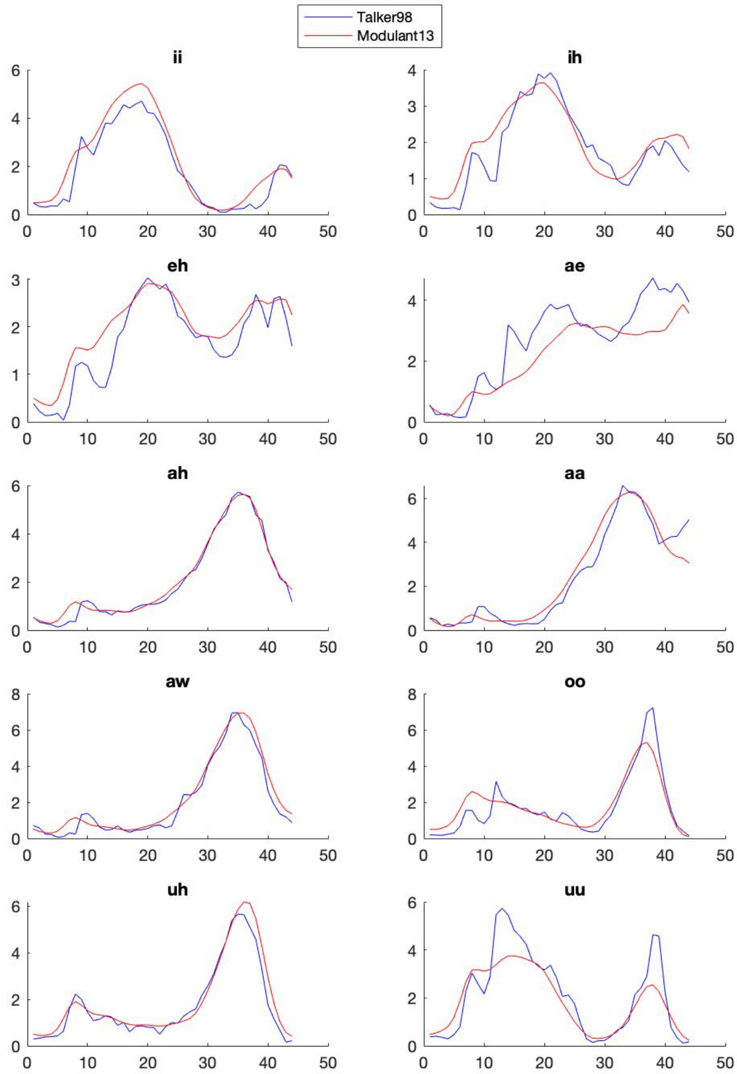

Figure C.3: Comparison of Vocal Tract Shapes estimated by MRI (story98) and feature reduction (Modulant13)