Realtime and Accurate Musical Control of Expression in Voice Synthesis

Nicolas d'Alessandro

A dissertation submitted to the Faculty of Engineering
of the University of Mons, for the degree of Doctor of Philosophy

Abstract

In the early days of speech synthesis research, understanding voice production has attracted the attention of scientists with the goal of producing *intelligible* speech. Later, the need to produce more *natural* voices led researchers to use prerecorded voice databases, containing speech *units*, reassembled by a concatenation algorithm. With the outgrowth of computer capacities, the length of units increased, going from diphones to non-uniform units, in the so-called *unit selection* framework, using a strategy referred to as "take the best, modify the least".

Today the new challenge in voice synthesis is the production of *expressive* speech or singing. The mainstream solution to this problem is based on the "there is no data like more data" paradigm: emotion-specific databases are recorded and emotion-specific units are segmented.

In this thesis, we propose to restart the expressive speech synthesis problem, from its original voice production grounds. We also assume that expressivity of a voice synthesis system rather relies on its interactive properties than strictly on the coverage of the recorded database.

To reach our goals, we develop the RAMCESS software system, an analysis/resynthesis pipeline which aims at providing interactive and real-time access to the voice production mechanism. More precisely, this system makes it possible to browse a connected speech database, and to dynamically modify the value of several glottal source parameters.

In order to achieve these voice transformations, a connected speech database is recorded, and the RAMCESS analysis algorithm is applied. RAMCESS analysis relies on the estimation of glottal waveforms and vocal tract impulse responses from the prerecorded voice samples. We

cascade two promising glottal flow analysis algorithms, ZZT and ARX-LF, as a way of reinforcing the whole analysis process.

Then the RAMCESS synthesis engine computes the convolution of previously estimated glottal source and vocal tract components, within a realtime pitch-synchronous overlap-add architecture. A new model for producing the glottal flow signal is proposed. This model, called SELF, is a modified LF model, which covers a larger palette of phonation types and solving some problems encountered in realtime interaction.

Variations in the glottal flow behavior are perceived as modifications of voice quality along several dimensions, such as tenseness or vocal effort. In the RAMCESS synthesis engine, glottal flow parameters are modified through several dimensional mappings, in order to give access to the perceptual dimensions of a voice quality control space.

The expressive interaction with the voice material is done through a new digital musical instrument, called the HANDSKETCH: a tablet-based controller, played vertically, with extra FSR sensors. In this work, we describe how this controller is connected to voice quality dimensions, and we also discuss the long term practice of this instrument.

Compared to the usual prototyping of multimodal interactive systems, and more particularly digital musical instruments, the work on RAMCESS and HANDSKETCH has been structured quite differently. Indeed our prototyping process, called the *Luthery Model*, is rather inspired by the traditional instrument making and based on embodiment.

The Luthery Model also leads us to propose the *Analysis-by-Interaction* (AbI) paradigm, a methodology for approaching signal analysis problems. The main idea is that if signal is not observable, it can be imitated with an appropriate digital instrument and a highly skilled practice. Then the signal can be studied be analyzing the imitative gestures.

Acknowledgements

First, I would like to thank Prof. Thierry Dutoit. Five years ago, when I came to his office and told him about making a PhD thesis in "something related to music", he could understand this project, see its potential, and – more than everything – trust me.

This PhD thesis has been the opportunity to meet extraordinary collaborators. I would like to highlight some precious meetings. Prof. Caroline Traube, who definitely helped me to dive into the computer music world; Prof. Christophe d'Alessandro and Boris Doval for their deep knowledge in voice quality and long discussions about science and music; and Prof. Baris Bozkurt, for his wiseness, experience and open mind.

There is nothing like a great lab. With TCTS, I have been really lucky. I would like to thank all my workmates, for their support and availability for all the questions that I had. More precisely, I would like to thank Alexis Moinet and Thomas Dubuisson, for their involvement without boundaries, in some common projects.

This thesis also results from a strong wish of some people to enable interdisciplinary research on multimodal user interfaces. These teams share a common name, eNTERFACE workshops. I would like to thank Prof. Benoît Macq for encouraging me to come back each year with new projects. I also would like to thank FRIA/FNRS (grant n° 16756) and Région Wallonne (numediart project, grant n° 716631) for their financial support.

Family and friends have been awesome with me, during these hard times I have had when writing this thesis. Special thanks to L. Berquin, L. Moletta, B. Mahieu, L. Verfaillie, S. Paço-Rocchia, S. Pohu, V. Cordy for sharing about projects, performance and art; A.-L. Porignaux, S. Baclin, X. Toussaint, B. Carpent for reading and correcting my thesis; M. Astrinaki for these endless discussions; A. Zara for this great journey in understanding embodied emotions, and for the great picture of the HandSketch (Figure 7.1).

Finally, I warmly thank my parents for their unconditional trust and love, and Laurence Baclin, my wife, without whom this thesis would just not have been possible.

Contents

“À mon père, certainement fier
de moi, où il repose désormais.”
— Nicolas d’Alessandro

Chapter 1

Introduction

“ L’expression est le seul caractère fondamentalement irrationnel, auquel la logique ne s’oppose pas. ”

Understanding voice production mechanisms has focused the attention of scientists for many years. More precisely we can consider that signal processing and computer science people have started this story about fifty years ago with the formulation of one of the first electrical models of the speech signal [77, 82]. From this breakpoint, research has gone through two main steps: expressing voice as equations (rule-based or articulatory-based) and concatenating segments of pre-recorded voice (content-oriented). These approaches both aim at generating the palette of existing sounds in a given language, called *phones*, in such a way that the *coarticulation* (phone-to-phone transitions) is respected.

Rule-based voice synthesis started in the early seventies. It comes with the desire to encode acoustic, phonological and linguistic knowledge as a set of rules to be interpreted by the computer. These rules have successively driven different generations of voice synthesizers: parallel formant [163], sinusoidal [159], and more recently articulatory models [156]. We usually criticize this generation of systems for being quite intelligible but not natural. They still sound synthetic, offering so-called “robotic” voices.

In the beginning of the nineties, storage performances of computers became high enough to directly manipulate pre-recorded voice segments, called *units*. The idea of content-oriented voice synthesis is to use the inherent coarticulative properties of recorded sound files, instead of modeling them with mathematical rules. The total amount and individ-

ual size of these units have evolved in correlation with technological capacities, going from single instances to multiple instances, and from diphones [40] to larger sequences.

For the last ten years the speech community has come to admit that an incomparable level of intelligibility and naturalness has been reached with recent non-uniform unit (NUU) selection systems [105]. NUU-based systems are content-oriented algorithms which use large databases (hours of speech) and variable-length units. NUU systems assume that the use of a large amount of data enables the selection of an appropriate unit for any kind of synthesis target, and with limited sound transformation [34].

The weak aspect of this technique lies in its loose control of prosody and voice quality[1]. Indeed if the selection of appropriate phonetic units is efficient with a large database, it remains difficult to have a coverage of all prosodic and voice quality occurrences. This is why NUU-based techniques systematically favors the use of the longest possible units, as a way of reducing the rate of concatenation discontinuities. The consequence of this practice is that the overall prosody and voice quality of the virtual speaker more and more depend on recording conditions.

We highlight a third strategy that has recently attracted the attention of many researchers. This generation of algorithm achieves the stochastic control of production models. Voice is produced by typical rule-based synthesis models, but the behavior of controlling parameters results from the training of HMMs[2] on a large database [189,164].

1.1 Motivations

Recent commercial speech synthesizers – such as *Acapela* [89] or *Loquendo* [129] products – are very impressive. Consequently these systems are progressively taking place in our daily activities: phone helpdesks, GPS, interactive systems in several institutions, etc. However, despite recent improvements in speech quality, these systems are not ready to address applications out of this functional context: directions, instructions, etc. Indeed we often feel that the *intent*[3] of the virtual speaker remains inappropriately neutral.

These observations show that despite the efforts that have been done in order to achieve 99% of the solution to this "human substitution" problem, the remaining 1% to tackle

[1] Concepts like prosody and voice quality are described in details in Chapter 2 of this thesis.
[2] HMM: Hidden Markov Models
[3] In this introduction, we are using the "intent", from its sociological/psychological point of view: the conscient and underlying desired end of a human within a given action.

is clearly significant. This problem, called the *uncanny valley*, has first been highlighted by Mori in the early seventies, concerning robots [143]. He assumes that the acceptance of human looking avatars increases with their *likeness*. Their appearance changes into revulsion when this likeness becomes confusing. Mori also assumes that this gap can be overcome, and that perfect acceptance can be reached, as illustrated in Figure 1.1.

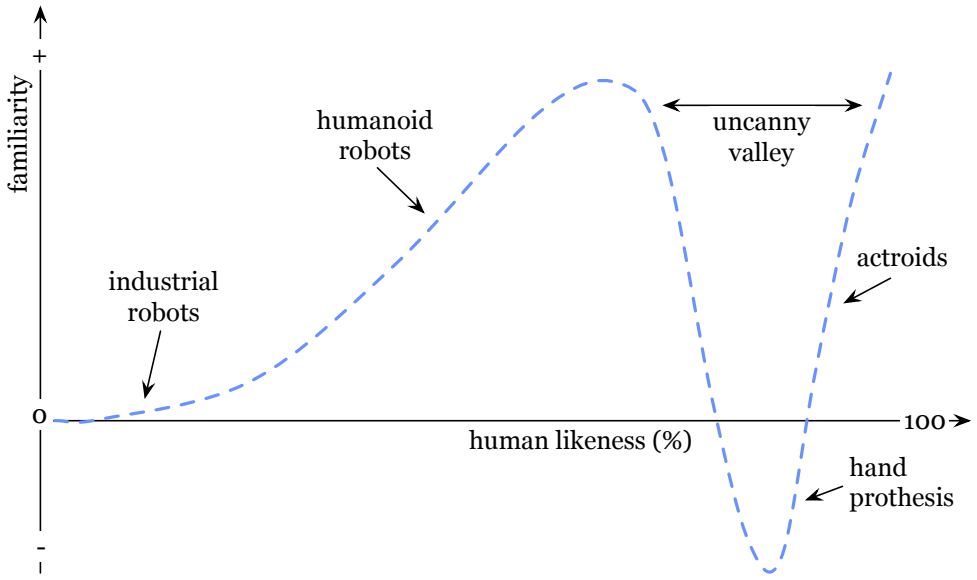

Figure 1.1: Mori's law: evolution of the acceptance of human robots by real people. We can see the uncanny valley which is a drop into revulsion when the avatar's likeness becomes confusing. It makes actroids [143] less accepted than less realistic human robots. Mori assumes that the gap can be overcome, if likeness reaches perfection.

This situation is due to various kinds of problems which can be encountered in the analysis, the modeling or the synthesis of contents, but can be depicted as a common aspect: *expressivity*. Being a transversal aspect of human behavior, there are many different definitions of expressivity in the literature, depending on research topics [39, 147]. In this work we aim at proposing a definition which stays general and flexible, thus not particularly in conflict with the state of the art.

> *Expressivity: subtle degrees of freedom that can be used in the delivery of a message in a given language, in order to transmit affective contents.*

Figure 1.2 uses the drawing of geometrical forms as an illustrative example. It shows how the units of a formal language (circle, square and triangle) can be delivered differently. The result remains clearly understandable, even if there are significant variations in the achievement: deviations, holes or exceedances.

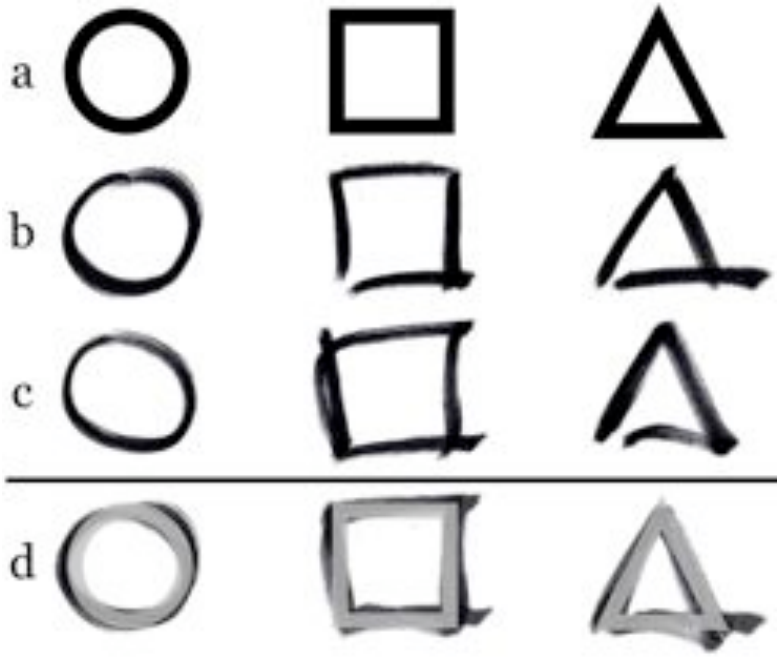

Figure 1.2: Geometrical forms are the formal language (a) and different drawn instances, first separated (b,c) then superimposed (d), give a pedagogical example of what we call expressivity: subtle degrees of freedom serving the affective contents.

Spoken languages are built on both *phonemes* and *prosody*, two aspects that are explained in Section 2.1. If the expressivity of a virtual speaker slightly differs from what we are expecting from a real human, the affective intent "sounds wrong" and the human-to-human communication behavior is replaced by a strictly functional interaction. This happens even if the message is clearly understandable, for instance in NUU systems.

This is the area in which *emotional speech research* has grown up for the last five to ten years, and the main topic of this PhD thesis. More precisely, this work generalizes from *speech* to expressive *voice* synthesis, as the musical context of this work drives us to also discuss singing issues. We also prefer *expressivity* to *emotion*, thinking it is less subjective, and more related to the speaker than to the speaker-listener interaction.

1.2 Speech vs. singing

In the voice processing community there is a long tradition of separating speech and singing into separate research areas. Speech processing labs and researchers are usually involved in fields of application related to daily life and industries, such as cellphone technologies, talking machines, GPS, help to disabled people, etc. On the other hand,

singing processing is driven by the stream of musical signal technologies, targeting music synthesis softwares, music information retrieval, live performances, etc.

However the vocal apparatus, presented in Section 2.1, is obviously the same. The main difference lies in the way this apparatus is used for speaking or singing [87]:

- differences in fundamental frequency: range and "vocabulary" (prosody/melody);
- differences in dynamics: range and "vocabulary" (accents/nuances);
- differences in timbre: behaviors of the larynx, resonators and articulators.

These differences in the use of the vocal apparatus have driven researchers to prefer separate models. Currently the speech community has some preferences for NUU selection strategies. As the sound is not deeply modified, these techniques emphasize models that represent the signal with high fidelity, as in the HNM [182], or using the waveform itself [44]. Singing synthesis solutions are rather based on more controllable spectral representations, like sinusoidal (SMS) [172] or source/filter models [19, 118].

This thesis emphasizes the use of the well-known source/filter model [77] and flexible representations of the vocal folds behavior [185, 96]. Consequently it offers the opportunity to work on voice production at a much more physiological level, and thus tackle expressive issues of both speech and singing.

The idea of working at a generalized voice production level is motivated by various reasons. First, current activities in contemporary art (music, theatre, dance) blur more and more the boundaries between speech and singing, proposing new uses of voice in performance. Add to it that in a large field of applications related to pop music, the way of singing is closer to speech, sometimes equivalent (hip-hop, slam), or even switching from one to the other in several styles. Finally, there is a significant interest (gain of time and energy) in being able to target singing applications from a speech database. This topic - called speech-to-singing conversion - starts to grow, and now challenges state-of-the-art synthesis systems such as STRAIGHT [165].

1.3 An interactive model of expressivity

In order to fill the uncanny valley of talking avatars, expressive speech synthesis research seems to converge towards applications where multiple databases are recorded (the di-

Figure 1.3: Front-ends of two successful "gigasampling" applications: Vienna Instruments™ from Vienna Symphonic Library™ (left) and SampleTank™ from IK Multimedia™ (right). SampleTank 2 provides attractive singing databases.

versification of the source), corresponding to a number of labelled expressions: happy, sad, angry, scared, etc. At synthesis time the expression of the virtual speaker is set by choosing the units from the corresponding section of the database, using unit selection algorithms [92]. Mainly this emotional labeling is done manually, managing different recording sessions with instructions given to the speaker.

Notice that a bias is sometimes introduced in these instructions in order to emphasize/exagerate a given expression, for instance requiring "joy" in order to get "surprise" units. This practice gives an overview on how the technological context can become distant from the original perspective, and from any theoretical model.

For the last years, the increase of database footprint has been quite transversal in synthesis technologies. We find similar situations in many musical signal synthesis contexts. For instance with the generalization of MIDI[4]-controlled "gigasamplers" and huge dynamic pitch/velocity matrices[5] [162]. Figure 1.3 presents two successful samplers, very representative of this evolution: Vienna Instruments™ and SampleTank™.

The idea of producing expressive variations for a given sound space can be seen as an orthogonal development of the database. For instance adding an "angry" attitude to a speech synthesizer requires new kinds of units for almost every targeted sentence. Thus it is common to multiply the size of the database by 2 or 3, in order to only produce several typical expressions. This quickly results in 5-6 hours of audio recording [35].

[4] Musical Interface for Digital Instruments: a protocol defined by keyboard/synthesizer makers in the 80's, in order to standardize the communication between electronic musical devices [127].

[5] We talk about a matrix because the MIDI protocol considers a sound as mainly driven by two parameters: the pitch (ID of the key used on the piano keyboard) and the velocity (related to the strength with which the key is pressed). It creates a two-entry table in order to represent a sample.

In the context of speech synthesis, this way of working has not really solved the inconsistency and fuzziness of the virtual speaker's intent. Recent expressive speech synthesizers propose a rather caricatural vision of expressions, comparable to *toons* or *smileys* in the visual world. Moreover these embedded expressions are related to the recording conditions (speaker and instructions), and absolutely not controlable during synthesis.

In singing voice synthesis, remarkable achievements have been reached. The algorithms proposed by *Bonada et al.* [24] provide naturalness and flexibility by organizing singing contents at a high performative level. We can also highlight singing synthesis derived from STRAIGHT [165] or HTS [164]. These approaches seem mature enough to allow the replacement of human singing by synthetic singing, at least for backing vocals.

However existing singing systems suffer from two restrictions. First they aim at mimicking singers and typical singing styles, rather than offering creative manipulation of the voice timbre. Secondly they are generally limited to note-based interactions, supposing the use of a MIDI controller, similarly to gigasampler architectures.

In this context we propose to investigate a novel approach. Along with other research topics related to the understanding of human behavior, we postulate that expression is a highly contextual characteristic of the human communication [128]. In this case, "contextual" means that an emotion can not be extracted as the absolute representation, but is rather based on context and interactions, as a continuously evolving stream.

We apply this view to voice production. It can be seen as a particular reference to the "pragmatic level" that is described in speech theory [69]. Furthermore this choice is part of a significant array of studies, encountering interactive aspects of voice quality [49].

However this assumption is quite radical. Indeed it means that providing a consistent affective stream – from the point of view of the intent – is theoretically impossible with unit selection techniques, at least if we continue to work with huge unit sizes. It also introduces the idea that the expressivity of a synthesizer is related to its refined interactive properties rather than the strict coverage of its database.

These considerations mark an important step in the way we currently work with voice synthesis, as it requires to come back to some fundamental concepts of voice production. Consequently, from the idea that an expressive system has first to be highly interactive, there are some new aspects to consider and others – often associated to obsolete speech synthesis issues – which become essential again:

- **Definition of realtime [A1]**

  A stronger definition of "realtime" has to be considered. Indeed manipulating a voice interactively requires that the granularity of the timeline decreases. Instead of syllables or even part of sentences, we have to consider working at the fundamental period (T_0) level. With typical voice parameters, it corresponds to some ms both for the latency and resolution of the timeline. This constraint immediately places this work in the context of short-term frame concatenation [169].

- **Analysis of expressivity [A2]**

  We need a better representation of the voice production, especially the behavior of the larynx. Indeed most of the research in expressive speech presents the parameters of the glottal flow as the most significant contribution in the way the expressivity is perceived by humans [120, 72]. Being able to precisely analyse the glottal flow on recorded voice is a research topic that has been tackled for many years. But we are probably in the first years where modifying/resynthesizing it with an acceptable quality seems accessible [88]. This thesis takes part to this axis.

- **(Re)synthesis of expressive contents [A3]**

  There are needs for a voice production framework that is compatible with our flexible and realtime definition of expressivity. Expressive control relies on realtime modifications of glottal source components. At the same time, intelligible voice relies on large corpus and preservation of transients. With recent voice analysis tools [26], we can expect to deeply modify properties of recorded samples. Our framework aims at taking the best compromise between corpus and rules.

- **Voice quality description [A4]**

  New mappings are required between perceptual dimensions and glottal flow parameters. These aspects are related to voice perception. Qualifying the voice timbre from the perceptual point of view, and relating *voice quality* to the analysis of signals can be seen as an important contribution to the speech processing community [102]. Specifically we target generalizing some considerations in order to better fit both speech and singing constraints, and defining perceptual spaces.

1.4 Analysis-by-Interaction: embodied research

The previous section concludes with four important axes that mainly define this thesis, [A1] to [A4]: the definition of realtime, the analysis of expressivity, the (re)synthesis of expressive contents, and the importance of perception in voice quality description.

However the most important aspect of this work is probably related to a much more transversal mechanism. Indeed a significant part of this research is related to the realtime manipulation of sounding materials, targeting creative purposes. Thus this thesis was made from daily activities which have a lot in common with the building of musical instruments: continuously mapping new sound spaces to new ideas of gestures [199].

From the point of view of traditional instrument making, it is known that a new musical instrument does not spontaneously appear first, and then is used by new performers. Instrument making is a long and close iterative process where both tools and practices evolve together. The saxophone is a particularly good example. Indeed it is today a charismatic instrument of jazz music. But it was first released in the continuity of other classical woodwinds, at the end of the XIX$^{\text{th}}$ century. Then the instrument and its corresponding practice changed progressively and simultaneously [95].

The adaptation of these activities to the technological world probably dislocated a little bit the unicity of the instrument making process, splitting it into different topics: signal processing, computer science, human-computer interaction, etc. The part devoted to the practice evolved in testing and benchmarking tasks, which usually happen at the end of an engineering iteration. This typical roadmap forgets that practicing an instrument is often much more a source of innovation than strictly a validation process.

Our point of view about digital instrument making meets Moore and Fels' research about human/device interaction, with concepts like *intimacy* or *embodiement* [142, 79]:

> *" One consequence when a person embodies a device is that expression is possible. One can conjecture that expression flows naturally when a device is completely embodied. "* — Sidney S. Fels

Consequently we think that there is an interesting space to be (re)investigated, related to this practicing activity. A significant part of this thesis has been involved in the making of finished instruments. The long term practice of them progressively sets the intimate human/object relationship and embodiment, as proposed in [79].

After a few years, the embodied use of the instrument provides an intuitive "know-how" in expressive sound production. Subtle sounding phenomena can be convincingly imitated by the performer. Applied to voice synthesis, it means that expressive voice production can be studied from the gestural point of view, giving a new lighting to usual analysis pipelines. We call this new methodology *Analysis-by-Interaction* (AbI), and it can be considered as the fifth and transversal axis of this thesis [A5].

1.5 Overview of the RAMCESS framework

RAMCESS is the name of the framework that has been built from our motivations, all along the thesis. Version 3.x is currently under developement. The various components of this software are progressively described in following chapters, but we think it is interesting to present an overall picture of the system in the Introduction.

Indeed it helps to see how the various issues reported in this thesis are imbricated together, and why some specific choices have been done: the source/filter model [77], LF-based glottal flow [78], etc. More precisely it highlights how the four motivations of Section 1.3 and the strategy of Section 1.4 create a relevant workflow.

We propose the mindmap in Figure 1.4 as a way of summarizing these ideas. It locates several important topics and directly references chapter names in the manuscript.

First we show that the musical practice (through the HandSketch, cf. Chapter 7) directly influences the voice production components: glottal flow and variant-shape vocal tract. This relation relies on the set of gestures that are applied on the instrument. This action is represented by blue dashed lines, going from the controller (A) to various synthesis modules: glottal flow generator (B), and variant-shape vocal tract (C).

We also assume that results of this ongoing practice influences database building and expressive analysis, through AbI feedbacks. It correponds to the new technological issues that are exhibited by the constant practice of any musical instrument[6]. This action is represented by forward green dashed lines going from the controller (A) to analysis steps: building a expressive database (D) and estimating glottal flow parameters (E).

One last important AbI mechanism is also highlighted: the feedback from generated voice sounds to the practice of the instrument. Indeed the desire to produce the most

[6] This situation can be depicted in every kind of musical practice. For example, a skilled guitarist would require a more accurate design of the fretboard, thus a better understanding of string behavior.

expressive sounds progressively modify the gestures themselves. This action is represented by the backward green dashed line, going from synthesis result to the user.

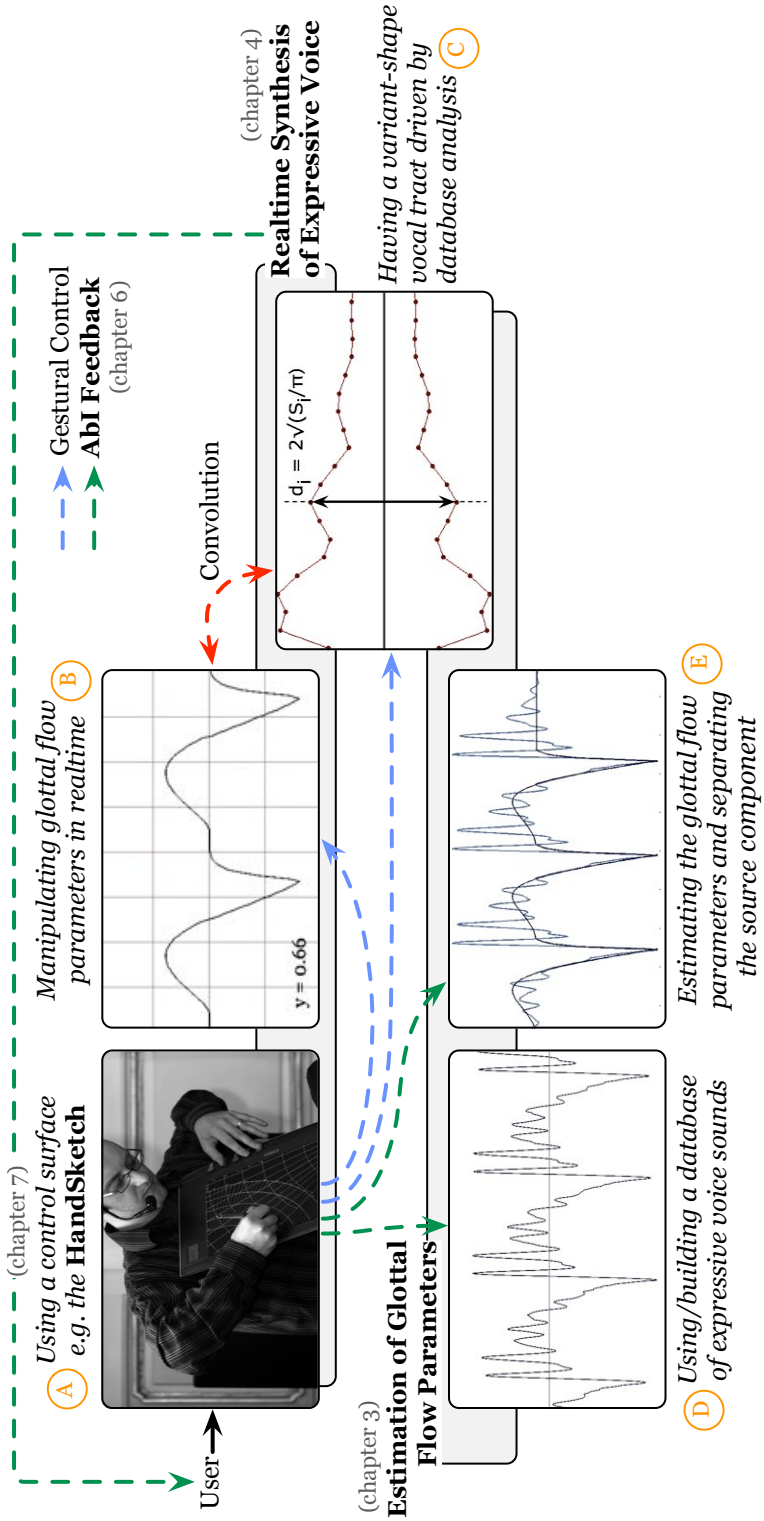

Figure 1.4: Mindmap of the RAMCESS framework.

1.6 Outline of the manuscript

Building on Sections 1.1 to 1.5, which introduced the various challenges of this PhD thesis, and particularly the mind map in Figure 1.4, we can now expose the structure of this manuscript. Chapters of this thesis are organized into three parts:

Part I – Introduction and state of the art

Following this Introduction, Chapter 2 presents a detailed state of the art in various fields: voice production, behavior of vocal folds, perceptual aspects of voice quality, glottal flow analysis and singing synthesis. It is a particularly important aspect of this thesis, as the work has been achieved in an interdisplinary way. Thus it is important to position this research respectively to each of its aspects.

Part II – RAMCESS: a framework for realtime voice quality transformations

Chapter 3 focuses on our work on the estimation of glottal flow parameters and on source/filter separation. It gathers and comments a series of existing techniques. Then it proposes the combination and the improvement of two existing algorithms for the estimation of the glottal source component on pre-recorded voice signals.

In Chapter 4 we describe the approach that is used in order to produce an expressive glottal flow signal in realtime. The complexity of this problem is discussed, and a new realtime-friendly version of the LF model is presented. A generalized mapping between voice quality dimensions and voice production parameters is also proposed.

Finally, Chapter 5 examines some recent work on extending the causal/anticausal description from voice processing to instrumental sound processing. Results related to the analysis of trumpet and violin databases are presented and discussed.

Part III – Analysis-by-Interaction methodology

One important part of this thesis concerns the Analysis-by-Interaction (AbI) methodology. Some preliminary motivations are explained in Chapter 6 and AbI is compared to current research activities in digital instrument making: sound synthesis and HCI.

Then the HandSketch digital musical instrument is presented in Chapter 7. HandSketch is the tablet-based controller which has been used in most of the experiments. The design is presented and the long-term practice of the instrument is discussed.

Finally we present the application of AbI to a case study: the synthesis of vibrato in singing, in Chapter 8. Indeed the HandSketch appears to be really expressive in the achievement of vibrating singing voice. Corresponding gestures are analyzed and AbI is used as a way of proposing a new model for the vibrato in singing.

1.7 Innovative aspects of this thesis

In Sections 1.3 and 1.4, we see that this thesis is interdisciplinary and aims at being involved in five main axes, from [A1] to [A5]. The architecture of the overall analysis/(re)synthesis system, called Ramcess, is described in Section 1.5. The structure of the manuscript, presented in Section 1.6, is made of three main parts.

In the development of this thesis, many related works are presented, discussed and often compared with our own assumptions. Thus, innovative aspects of this thesis are rather disseminated through the various Chapters. In this Section, we propose to focus on what we consider being the four original points, from [P1] to [P4]. Then these points are cited as such, when they appear in the following Sections of this manuscript:

- **New glottal flow synthesizer, adapted to realtime manipulation [P1]**

  In Chapter 4, we describe the Ramcess voice synthesizer. One important module of this synthesis engine is the realtime generator of the glottal source signal. This new generator, called SELF (Spectrally Enhanced LF), solves most of realtime-related issues encountered in existing glottal source models, such as LF or CALM.

- **Extension of ZZT-based decomposition to instrumental sounds [P2]**

  The ZZT-based decomposition has been designed for extracting glottal source and vocal tract components from prerecorded voice signals. In Chapter 5, we show that this decomposition technique can also be applied to instrumental sounds, coming from e.g. the trumpet or the violin. Modeling of these decomposition results leads to the definition of new parameters for the analysis/resynthesis of these sounds.

- **HandSketch: an expressive tablet-based digital musical instrument [P3]**

  The main aspect of this thesis is the realtime control of expression in voice synthesis. In order to reach this purpose, a new digital musical instrument, called the HandSketch, is presented in Chapter 7. This is a tablet-based controller, played vertically, with extra pressure sensors. This position and the mapping strategies lead to a remarkably expressive instrument for computer music performances.

- **Analysis-by-Interaction (AbI): a new approach for signal analysis [P4]**

  The HandSketch is widely used for performing purposes, but this thesis shows that this controller can be used for signal processing research. Indeed the HandSketch is involved in the AbI methodology, presented in Chapter 6. One important aspect of this thesis is the demonstration of AbI relevance, by describing how the HandSketch-based imitation of vibrato in singing leads to proposing a new vibrato model, through the analysis of imitative gestures (cf. Chapter 8).

1.8 About the title and the chapter quote

This thesis is entitled *Realtime and Accurate Musical Control of Expression in Voice Synthesis*. As the first interest of this thesis was more focused on the synthesis of singing, another title had initially been targeted, which was *Realtime and Accurate Musical Control of Expression in Singing Synthesis*. This title gave its name to the sofware that has been developed, with the acronym Ramcess, with a clear reference to the egyptian dynasty, particularly highlighted by versionning, such as Ramcess 2.x.

Though the title has been repurposed to voice synthesis, the acronym remains egyptian-style, in order not to confuse users too much, and to keep this interesting pun. Moreover, as we present an extension of the voice analysis tool to instrumental sounds, the first "S" of the acronym could be seen as "Sound", thus being generalized.

The straightforward english translation of this chapter quote is *"Expression is the only fundamentaly unrational behavior, which is not opposed to logic"*. It refers to a more philosophical discussion that I recently had with my visual art students. Usually artistic and scientific approaches are classified as opposed, considering that the first one is based on affective streams and the second one on reasonable analysis.

When art and science have to work together, this antagonism probably has to be slightly reconsidered. What pushes humans to express themselves is probably not rational. But expressing always consists in altering materials inside the rational world. We rather have to propose a partnership with rational topics (physics, computer science) and not oppose them. This equation also works backwards. This thesis tries to show that the intuition is not a prohibited aspect of scientific investigation, as far as it can be justified.

Chapter 2

State of the Art

> *" I felt advancing communication would advance our quality of life. "*
> — James L. Flanagan

This thesis shares boudaries with many different and heterogenous topics, such as glottal flow analysis, speech and singing synthesis, musical signal analysis, gestural control, concatenative audio frameworks, etc. However it is clear that voice production acts as the underlying and "connecting" aspect of the whole research strategy. Intentionally we try to use the term *voice* instead of reducing the scope to *speech* or *singing*.

Voice is our most flexible and expressive means for human-to-human communication. Some research seems to demonstrate that verbal skills are at the basis of human's intelligence, in many different aspects [112]. In the practice of art, especially music and theatre, voice is modulated with a lot of refinement, in order to create complex timbral gestures. Moreover different studies show that our perception of instrumental and speech sounds have significant overlapping regions [190, 18].

Consequently proposing a "State of the Art" in expressive/interactive voice synthesis is difficult to do straightforwardly. We have to think about what are the research topics that act in the same playground as this composite activity.

We start with an introduction to voice production issues, the well-known source/filter model [77] and its drawbacks (cf. Section 2.1). In Section 2.2 we give a description of the behavior of vocal folds, and discuss two existing models: LF [78] and CALM [65]. We also introduce perceptual aspects of the glottal flow, in Section 2.3. Then we give an overview

of the current situation in glottal waveform analysis and source/tract separation, in Section 2.4. Finally, as this thesis brings new insights in the interactive production of voice, we address an connected topic: the synthesis of the singing voice, in Section 2.5.

2.1 Producing the voice

In this Section we give an introduction to the main aspects of voice production. Phonation is first presented from the anatomical point of view in 2.1.1. Then we describe common assumptions that are made in the context of the source/filter model in 2.1.2.

2.1.1 Anatomy of the vocal apparatus

As Sundberg explains in his book [185] the vocal apparatus consists of all the organs that are involved in the production of vocal sounds. It is based on two main parts.

Inside the neck stands the larynx. The larynx contains the vocal folds. When the phonation of a voiced sound is desired by the locutor, the vocal folds are moved by surrounding muscles in order to block the trachea. Under the pressure of lungs, vocal folds start to achieve opening/closing asymetric movements within a given period (T_0). It produces a rich harmonic oscillation with a fundamental frequency f_0, called the *glottal flow* (GF). The obstruction of the trachea can also be partial or absent, creating unstable vibrations or air turbulances. It results in a continuum of phonation types, going from breathy vowels to fully noisy sounds, called unvoiced phonation.

The second part is located in the region of the neck above the larynx and in the head. Acoustic waves generated by the vocal folds then propagate inside the pharyngal, nasal and oral cavities. They form the *vocal tract*. These cavities can be seen as acoustic shape-variant resonators, with their own eigen frequencies. Consequently, passing through these cavities, the glottal waveforms are modified. Energy aggregates in different frequency bands of the spectrum, drawing what we call the *formants*. The main variation in the shape of the vocal tract is due to mouth articulators: teeth, tongue and jaw. Finally the sound radiates from the two outputs of the vocal tract: nose and lips openings. In this particular transition we consider that the waveforms convert from plane to spherical propagation. Figure 2.1 gives a summary of the whole mechanism.

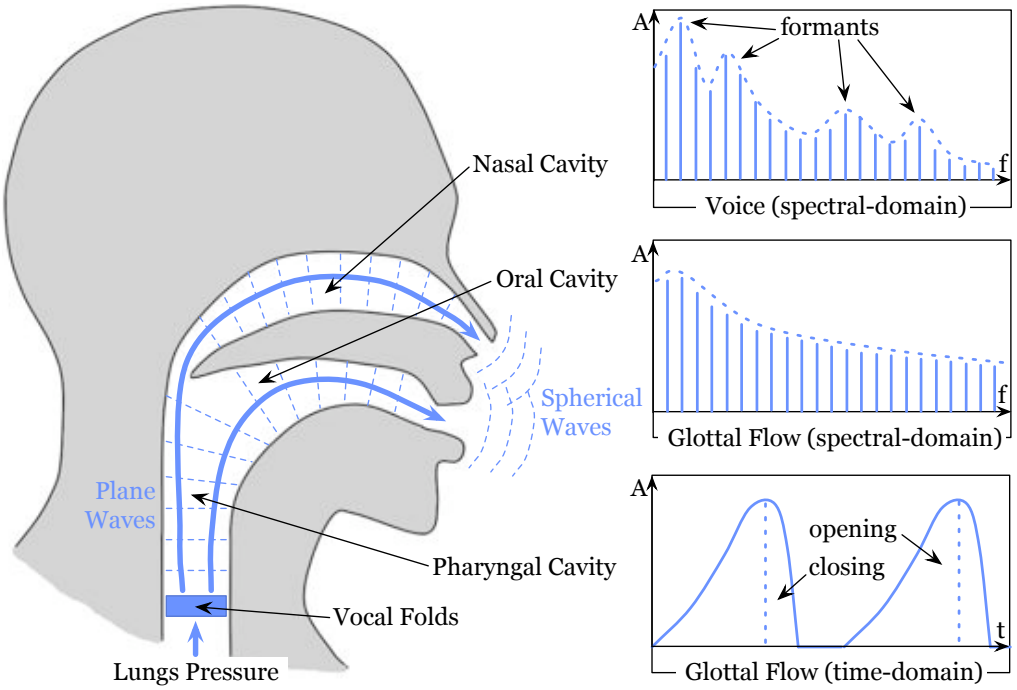

Figure 2.1: Vocal folds (inside the larynx) vibrate due to the lungs pressure. The vibration is a sequence of asymetric openings and closings (bottom graph), creating a rich harmonic spectrum (middle graph). Plane waves propagate in the vocal tract, sculpting the spectrum with formants (top graph). Finally waves radiate.

2.1.2 Source/filter model of speech

According to Fant's source/filter model [77] the production of speech can be seen as the cascading of three main operations, clearly inspired by the physiological description of the voice organ: a periodic/aperiodic generator (the source, the excitation), a vocal tract filter (for the sculpting of formants) and the radiation of lips and nose openings. Figure 2.2 gives a schematic view of the source/filter assumption.

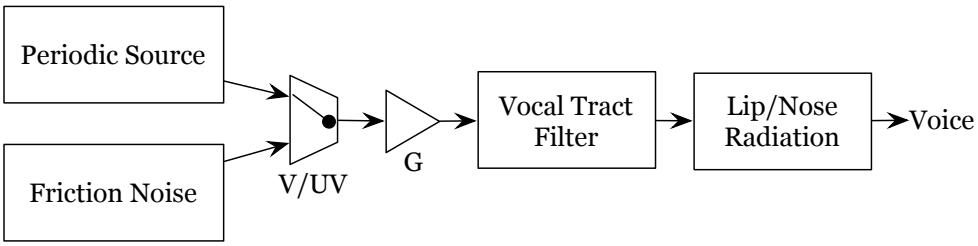

Figure 2.2: Simplified block diagram of the source/filter model of voice production: a periodic/aperiodic excitation, a vocal tract filter, and the lips/nose radiation.

The source can be modeled a mixture of a periodic signal and a noise. The voice/unvoiced switch (V/UV) defines if noise or periodic signal is sent to the following steps. In early

Linear Predictive (LP) systems [134], the source was generated with an impulse train (controlled by f_0 and a gain) and a white noise, as the behavior of the glottis, the vocal tract and the radiation were reported on the design of the filter.

Recent source/filter systems [20] now try to use a representation of the source $G(z)$ which is closer to the real behavior of the glottal flow, and generated in time or frequency domains. Another approach to highlight is the use of dictionnary-based excitation [167] using a representation of the glottal source, and called GELP systems [3].

Concerning other componants of the phonation, the vocal tract and the lips radiation, some simplifications are convenient, and remain acceptable to some extent.

- On the one hand the vocal tract filter can be modeled as an all-pole filter – see equation (2.1) – with at least 4-5 resonances ($p \geq 10$). The spectral enveloppe can be shaped with various kinds of parameters, depending on the structure of the filter, e.g. LPC, PARCOR or cepstral coefficients [42]. Even if the PARCOR representation has a geometrical meaning [134], these representations are clearly based on the spectrum. We also find 1D [45], 2D [145] or 3D [81] physical models.

$$V(z) = \frac{A}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.1}$$

- On the other hand, the effect of lips and nose openings can be simplified as the derivative of the volume velocity signal. It can be explained by representing the plane/spherical transition as a 1D signal in the direction of propagation. It is generally processed by a time-invariant high-pass first order linear filter $L(z)$ [77]. We can also highlight some research about 2D/3D models for plane/spherical transition of waves [114]. It replaces the simple derivative by a more complex acoustical model of the mouth opening, considering measured directivity patterns.

In most of the source/filter related research, these three typical steps are studied by their time/frequency domain behaviors in the field of digital signal processing. Source, filter and radiation modules are thus cascaded and periodified, if we accept the assumption of stationarity, as described equation (2.2).

$$S(z) = G(z) \times V(z) \times L(z) \times \uparrow\uparrow\uparrow_{T_0}(z) \tag{2.2}$$

$$G'(z) = G(z) \times L(z) \tag{2.3}$$

where $\uparrow\uparrow\uparrow_{T_0}(z)$ is the z-transform of an impulse train with a period of T_0.

Another common practice in speech representation consists in merging the model of the glottal flow and the lips derivation, as in equation (2.3). This leads to *glottal flow derivative* (GFD) models $G'(z)$, which are widely appreciated in the speech community, as the underlying glottal flow derivative waveform is direcly related to the speech signal waveform. This property is illustrated in Figure 2.3 for a sustained [a].

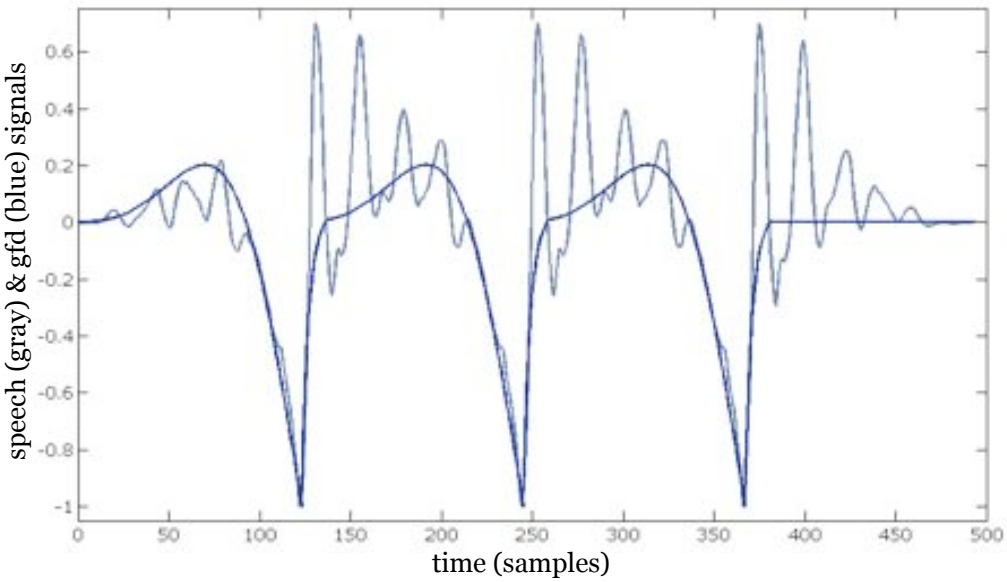

Figure 2.3: Speech waveform of a sustained [a] (gray) and underlying glottal flow derivative (blue): combined effects of the glottal flow and the lips radiation.

The three-block diagram also makes the assumption that there are no backwards influences, often called the coupling effects. Thus from the acoustical point of view, this model describes the voice production as a *forced oscillation*, with no interaction from the acoustic resonator (vocal tract) to the mechanical vibrating source (vocal folds). However we know that interactions happen, in various situations. The source/tract coupling is increased by e.g. high pitch or high impedance vowels like [u] [188].

2.2 Behavior of the vocal folds

As described in Section 2.1 the production of the glottal flow is an intentional action. It results from the displacement of the two folds, in order to block the circulation of the air flow, coming from the lungs. Childers explains repetitive movements of the folds within the myeloelastic-aerodynamic theory of phonation [42].

Under the increasing pressure, the folds slowly open and the glottal flow increases. We call it the *opening phase*. Then this system reaches an elasto-plastic limit where the returning force of the vocal folds and the lungs pressure are balanced. We reach what is called the *maximum of opening*. When that balance is exceeded, the folds suddenly achieve a closing movement, due to the Bernoulli effect, called the *closing phase*.

From the moment the two folds are touching each other, to the beginning of the next opening, we have the *closed phase*. Due to the thickness of vocal folds, the *complete closure* happen after a few time, called the *return phase*. Figure 2.4 illustrates the geometry of these opening, closing and returning movements.

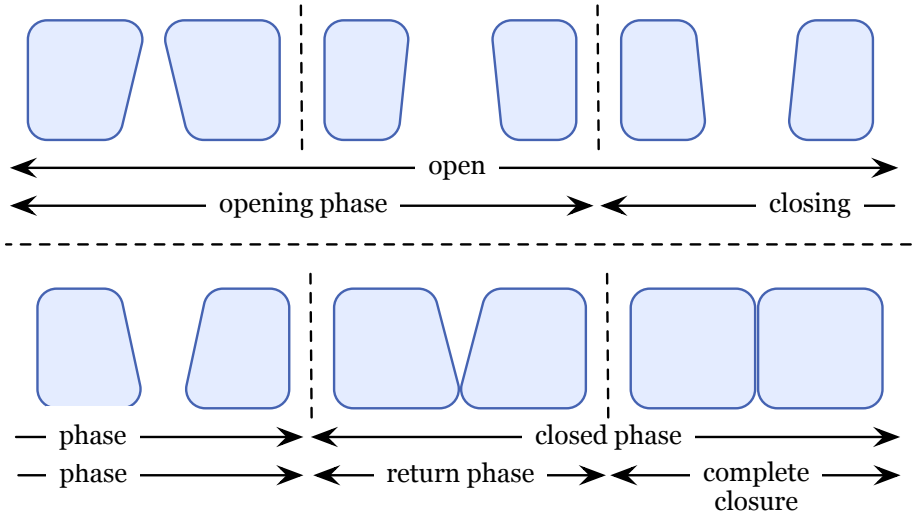

Figure 2.4: Six sagital representations of vocal folds in one period of vibration: open (opening and closing), return phases and complete closure.

The sequence of opening, closing – opening and closing phases together can be seen as the *open phase*, as the folds are separated – and closed phases is repeated during harmonic parts of voice production. This cycle can be achieved with various shapes. Indeed opening, closing and return phases can be shorter or longer within the period. They can also be affected by some noise and have different amplitudes.

> *Voice quality is the component of speech which gives primary distinction to a given speaker's voice when pitch and loudness are excluded. It involves both phonatory and resonatory characteristics.* — Mondofacto [141]

Childers defines the *laryngeal voice quality* as voice timbre variations due to the glottal source behavior [43]. We can consider that most of the voice quality variations are larynx-related, even if some studies refer vocal tract based voice quality modifications, such as the singer/speaker's formant [14]. Quantifying voice quality is a milestone for studying the analysis, synthesis and perception of expressive voice.

In this Section we present different aspects of the glottal flow (GF) and glottal flow derivative (GFD) description. First we present usual time domain parameters of the GF/GFD in 2.2.1. This leads us to the present the widely used Liljencrants-Fant time domain GFD model, in 2.2.2. Then we present the GFD from the spectral point of view in 2.2.3. We insist on the phase spectrum properties by introducing the mixed-phase model of speech in 2.2.4. An implementation of this model called CALM is presented in 2.2.5. Finally we discuss the assumption of the complete closure in 2.2.6.

2.2.1 Parameters of the glottal flow in the time domain

Through the voice quality literature and commonly used glottal flow (GF) or glottal flow derivative (GFD) models, several time domain parameters of the GF/GFD can been seen as transversal. The works of Childers/Lee [43], and Doval/d'Alessandro [64] in this field have significantly formalized the approach. This part gives a list of eight characteristics that are widely used for the description of GF and GFD:

- Open and closed phases happen in an overall sequence of length T_0. The length of this cycle is the fundamental period. The repetition of the period over the timeline produces a quasi-stationnary signal with a given fundamental frequency f_0.
- The length of the open phase within the period can be very variable. If we consider an open phase of length T_e, we can define the *open quotient* as a ratio between length of the open phase and the fundamental period, by the relation:

$$T_e = O_q \times T_0 \quad \Rightarrow \quad O_q = \frac{T_e}{T_0}$$

- The value of the open quotient has an infuence on the time-domain structure of the waveform. Indeed it drives the relative position of a particular event of the glottal flow mechanism, called the *Glottal Closure Instant* or GCI. This event happens at the end of the open phase, when the vocal folds touch each other.

- Within the open phase, the respective durations of the opening and closing of the glottis influence the symmetry of the waveform. Symmetry is often measured as the time T_p of maximum of opening, but can also be seen as a proportion of the open phase. Two different coefficients can describe this asymmetry: the *asymmetry coefficient* α_M or the *speed quotient* S_q:

$$T_p = \alpha_M \times T_e \quad \Rightarrow \quad \alpha_M = \frac{T_p}{T_e}$$

$$S_q = \frac{\textit{opening phase duration}}{\textit{closing phase duration}} = \frac{T_p}{T_e - T_p}$$

- The value of the glottal flow at the maximum of opening is an important aspect of the perception of loudness. It is called the *amplitude of voicing* A_v.

- When the glottal flow derivative is considered, another aspect is important in the scaling of the vibration. It concerns the amplitude of the GFD waveform at the GCI, usually noted E. It is straighforward to understand that E represents the sharpness of the closing and thus the velocity of the "clap".

- The *return phase* is usually modeled as a decreasing exponential in the time domain. The related time constant is used as a parameter and noted T_a.

- One last aspect which is usually added to the description of one period of glottal flow is the amount of turbulent noise. Noise appears when the closure of vocal folds is not perfect. Thus a continuous air flow propagates in the vocal tract and create sounding turbulences [77]. We can represent this mechanism with a continous V/UV, going from 0 (perfectly periodic glottal flow) to 1 (full noise).

  Figure 2.5 locates all these parameters on GF and GFD waveforms.

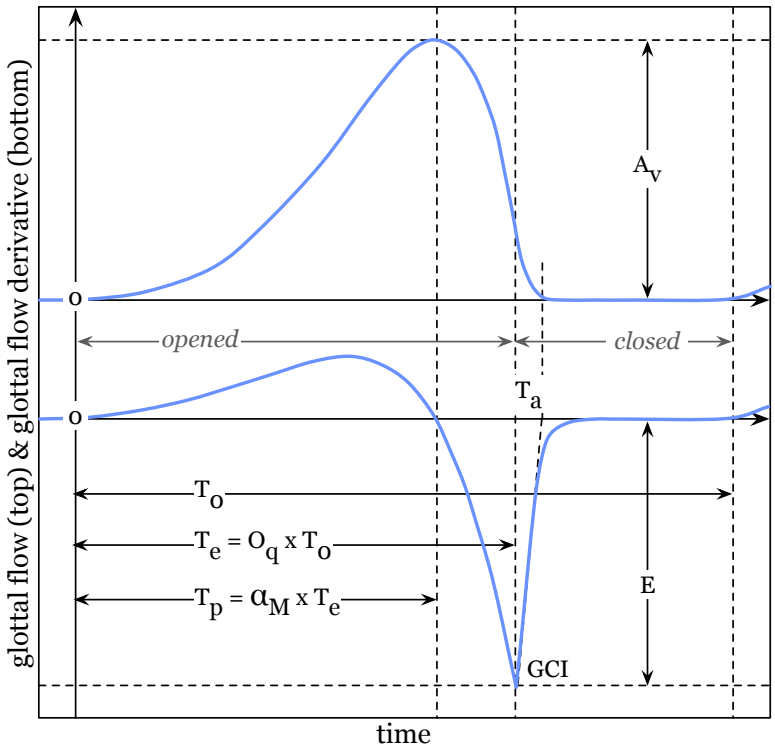

Figure 2.5: One period of glottal flow and glottal flow derivative waveforms, with parameters: T_0, T_e, T_p, O_q, α_M, T_a, A_v, E, the GCI and open/closed phases.

2.2.2 The Liljencrants-Fant model (LF)

The Liljencrants-Fant model (or LF) model defines the waveform of the GFD by adjusting two curves together in the time domain [78]. The model is driven by five parameters, that already have been presented in the general time domain description: T_0, E, T_e, T_p, and T_a. The model comes as a system of two equations. The first one describes the segment on the left of the GCI: an exponentially increasing sinusoid. The second one describes the segment on the right of the GCI: a decreasing exponential.

$$U'_g(t) = \begin{cases} -E\ e^{a(t-T_e)}\ \frac{sin\frac{\pi t}{T_p}}{sin\frac{\pi T_e}{T_p}} & \text{if} \quad 0 \leq t \leq T_e \\ -\frac{E}{\epsilon T_a}\ (e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_0-T_e)}) & \text{if} \quad T_e \leq t \leq T_0 \end{cases} \tag{2.4}$$

Two adjustements have to be verified, in order to generate a waveform which is physiologically acceptable. On the one hand, both curves must connect exactly at the GCI. Obviously the GF is a continuous variation, without any possible inflection. Conse-

quently the GFD can not have any discontinuity. On the other hand, one period of GF must correspond to a full cycle of opening and closing, thus coming from and returning to zero[1]. The integration of the GFD on the whole period must be zero.

Adjustements are done by solving a system of two implicit equations for parameters a and ϵ, as presented in equations (2.5) and (2.6). These two parameters are modifiers applied on left and right waveforms, in order to verify the physiological conditions.

$$\epsilon T_a = 1 - e^{-\epsilon(T_0 - T_e)} \tag{2.5}$$

$$\frac{1}{a^2 + (\frac{\pi}{T_p})^2} \left(e^{-aT_e} \left(\frac{\frac{\pi}{T_p}}{sin\frac{\pi T_e}{T_p}}\right) + a - \frac{\pi}{T_p} cot\frac{\pi T_e}{T_p}\right) = \frac{T_0 - T_e}{e^{\epsilon(T_0 - T_e)} - 1} - \frac{1}{\epsilon} \tag{2.6}$$

We can also obtain the equation of the GF, by integrating equation (2.4). The result is presented in equation (2.7). Some synthesizers need to generate GF pulses, e.g. as a way of modulating additive noise [52] or if the lip radiation is computed with another method [81]. This equation is also useful to get the value of A_v, in equation (2.8).

$$U_g(t) = \begin{cases} \frac{-Ee^{-aT_e}}{sin\frac{\pi T_e}{T_p}} \frac{1}{a^2 + (\frac{\pi}{T_p})^2} \left(\frac{\pi}{T_p} + ae^{at} sin\frac{\pi t}{T_p} - \frac{\pi}{T_p} e^{at} cos\frac{\pi t}{T_p}\right) & \text{if } 0 \leq t \leq T_e \\ -E\left(\frac{1}{\epsilon T_a} - 1\right)\left(T_0 - t + \frac{1}{\epsilon}(1 - e^{\epsilon(T_0 - t)})\right) & \text{if } T_e \leq t \leq T_0 \end{cases} \tag{2.7}$$

$$A_v = U_g(T_p) = \frac{-Ee^{-aT_e}}{sin\frac{\pi T_e}{T_p}} \frac{\frac{\pi}{T_p}}{a^2 + (\frac{\pi}{T_p})^2} \tag{2.8}$$

2.2.3 Glottal flow parameters in the frequency domain

Traditionally, the glottal flow has been modeled in the time domain. In [96] and [64] we find a significant breakthrough with a formalization of spectral behaviors of GF and GFD. The underlying idea is to consider that the spectral approach can be seen as equivalent to time domain only if both amplitude and phase spectra are considered.

[1] As we will in Section 2.2.6 and Chapter 4 thus assumption is more than physiological. It also forces the glottal cycle (opening and closing phases within one period) to completely close.

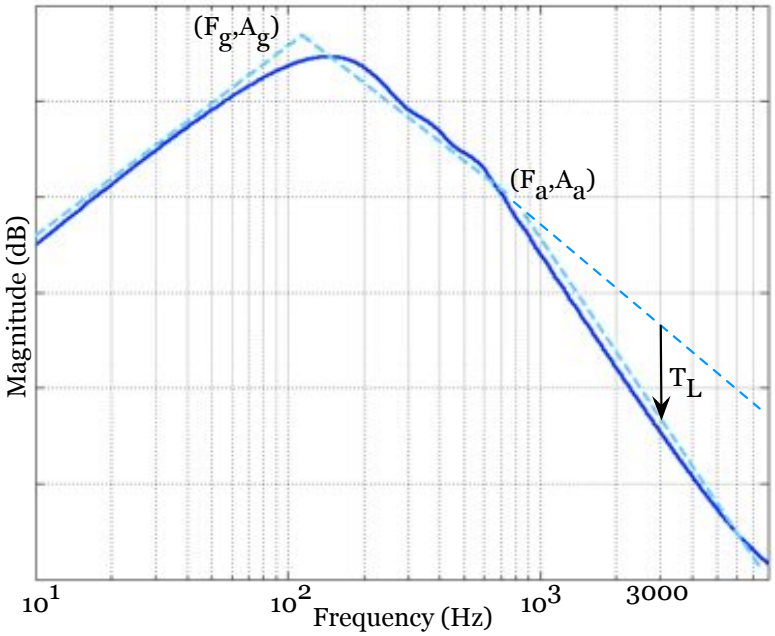

Figure 2.6: Spectrum of the glottal flow derivative: we can observe the glottal formant (F_g,A_g) and the spectral tilt (F_a,A_a), with its parametrization at 3000Hz, T_L.

Observing the magnitude spectrum of the GFD, two effects can be highlighted, as it can be seen in Figure 2.6. On the one hand, an amount of energy is concentrated in low frequencies (typically below 3 kHz). This peak is usually called the *glottal formant*. It has been shown that bandwidth, amplitude and position of the glottal formant (F_g,A_g) can change with voice quality variations [96]. On the other hand, we see a variation of spectral slope in higher frequencies ($> F_a$), called the *spectral tilt* [120].

In order to understand what are the correlations between the time domain waveform and the spectrum of the GFD, we have to introduce the concept of *causality*:

- If we observe the GFD and look for a component that exhibits a resonance – in order to explain the glottal formant – the left part, i.e. the segment of the waveform before the GCI, particularly fits the need. Indeed that segment has the shape of a second order resonant impulse response, but with a property that can be explained in two ways: the response is unstable and stops exactely at the GCI, or the response starts at the GCI and runs backwards. Spectrally both assumptions correspond to an anticausal componant: two conjugate poles oustide the unit circle.
- Working within the same model, we can highlight that the segment on the right of the GCI is shaped like a decreasing exponential. Thus it affects the magnitude

spectrum by doubling[2] (in dB) the tilt of the slope after a given cut-off frequency F_a. The longer the time constant (T_a) is, the smaller is F_a. As the spectral tilt T_L is evaluated by the decrease of energy at 3000Hz due to the return phase, decreasing the value of F_a (increasing T_a) leads to an increasing value of T_L. Figure 2.7 shows that finally the relation between T_a and T_L is non-linear and rather logarithmic.

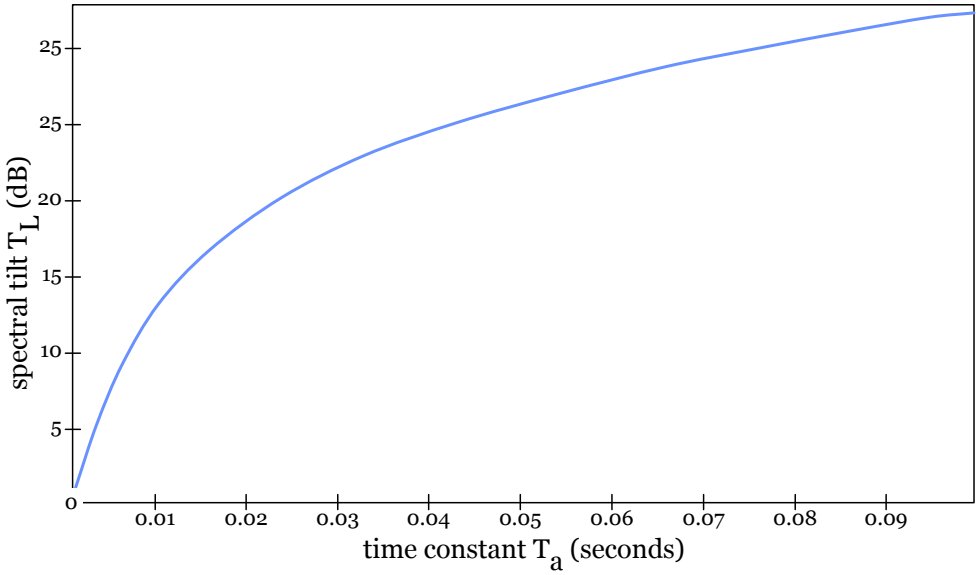

Figure 2.7: Empirical relation between the time constant of a first order impulse response T_a and the decrease of energy at 3kHz T_L compared to the spectrum of a Dirac.

Following the source/filter representation, these effects appear on the voice signal. The glottal formant influences the distribution of energy in the lowest part of the magnitude spectrum, thus within the first harmonics of the voice. Then the spectral tilt makes the voice more or less bright (amount of high frequencies in the magnitude spectrum).

This representation also leads us to consider that speech signals exhibit both minimum-phase and maximum-phase componants, which is a breakthrough considering usual LP assumptions. This defines the so-called *mixed-phase* model of speech [25].

2.2.4 The mixed-phase model of speech

In most of the speech processing literature LP analysis – and thus implicitly a minimum-phase framework – is used as a basis of work. However recent investigations have proposed a mixed-phase speech model [25], based on the assumption that speech is produced

[2] The glottal formant leads to -20dB/dec after F_g (2nd order low-pass filter, derivation). The spectral tilt filter "doubles" this slope by adding another -20dB/dec (1st order low-pass filter) after F_a.

by convolving an anticausal and stable source signal (zeros and poles outside the unit circle) with a causal and stable vocal tract filter (zeros and poles inside the unit circle). The speech signal is thus a mixed-phase signal obtained by exciting a minimum-phase system (vocal tract) by a maximum-phase signal (glottal source). An example of mixed-phase convolution applied to speech is illustrated in Figure 2.8.

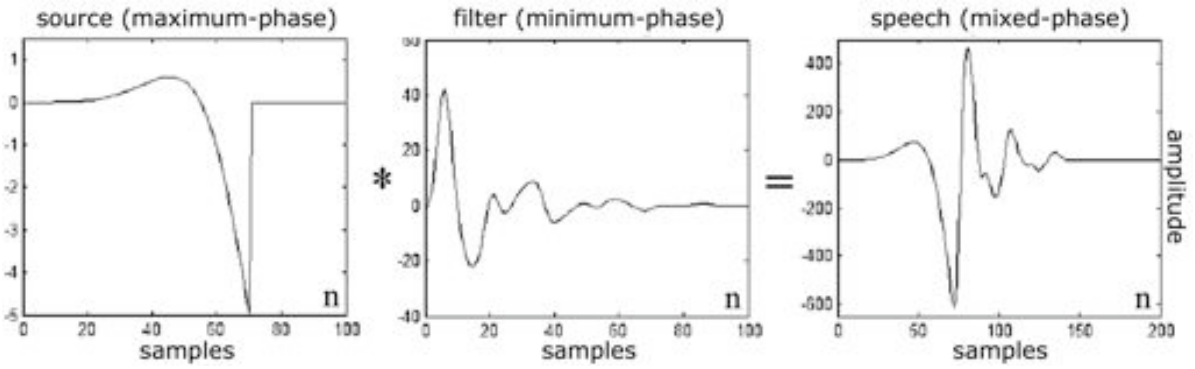

Figure 2.8: Mixed-phase representation of speech: convolution of a maximum-phase source with a minimum-phase filter, and the GCI as a singular point [25].

However considering that the source is the anticausal part and that the tract is the causal part is an approximation: a close observation of the behavior of vocal folds [96] shows us that GF waveform contains both a anticausal part (open phase) and causal part (return phase). This aspect is even clearer on GFD, where the junction between anticausal and causal parts of the waveform happens at the GCI.

Using a mixed-phase model is equivalent with the assumption that the speech signal has two types of resonances: multiple causal resonances due to vocal tract acoustics, called formants, and one anticausal resonance called the glottal formant [65].

2.2.5 The causal/anticausal linear model (CALM)

Considering the spectral representations of GF and GFD, a new model has been proposed in order to synthesize both their magnitude and phase behaviors with digital linear filters. This model is called CALM for *Causal/Anticausal Linear Model*. CALM generates the GFD signal by computing the impulse response of a cascade of two filters.

- $H_1(z)$: second order resonant low-pass at (F_g,A_g), and anticausal;
- $H_2(z)$: first order low-pass at (F_a,A_a), and causal.

The complete study of spectral features of GF in [65] gives us equations linking relevant parameters of glottal pulse (f_0: fundamental frequency, O_q: open quotient, α_M: asymmetry coefficient and T_L: spectral tilt, in dB at 3000Hz) to the coefficients of $H_1(z)$ and $H_2(z)$. An overview of this work is presented from equations 2.9 to 2.14. We can highlight that expression of b_1 has been corrected in [53], compared to [65] and [52].

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2} \tag{2.9}$$

$$\begin{cases} a_1 &= -2\ e^{-a_p T_e}\ cos(b_p T_e) \\ a_2 &= e^{-2a_p T_e} \\ b_1 &= ET_e \end{cases} \tag{2.10}$$

$$\begin{cases} a_p &= \frac{\pi}{O_q T_0\ tan(\pi \alpha_M)} \\ b_p &= \frac{\pi}{O_q T_0} \end{cases} \tag{2.11}$$

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}} \tag{2.12}$$

$$a_{T_L} = \nu - \sqrt{\nu^2 - 1},\ b_{T_L} = 1 - a_{T_L} \tag{2.13}$$

$$\nu = 1 - \frac{1}{\mu},\ \mu = \frac{\frac{1}{e^{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{F_e}) - 1} \tag{2.14}$$

2.2.6 Minimum of glottal opening (MGO)

Presenting the glottal flow behavior as the concatenation of two separated phases, connected around an event, called the glottal closure instant (GCI), is particularly suitable for modeling. Therefore the assumption of a GCI is often accepted in analysis and synthesis. However we know that the clear closure of vocal folds does not happen sys-

tematically in real phonation. Complete closure is rather limited to the production of low-pitch and low-impedance (open vocal tract) vowels, such as [a] or [o] [178].

On the one hand, the production of a higher fundamental frequency (> 200Hz) progressively reduces the closed phase, for mechanical reasons [100]. This effect is illustrated in Figure 2.9. The relative increase of open and return phases within the fundamental period is achieved, simulating an increase of pitch[3]. The GF is synthesized with CALM in order to avoid the arbitrary synthesis of a GCI. The loss of a clear closure (residual opening) is observed between two maxima of opening.

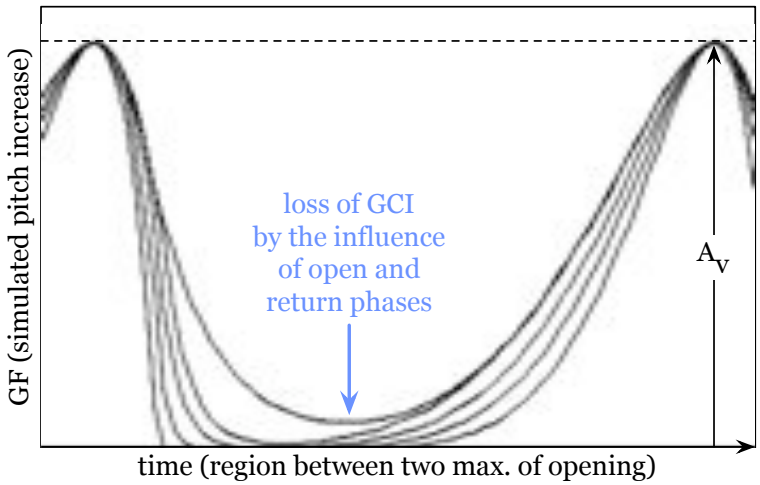

Figure 2.9: Effect of the increase of open and return phases within the fundamental frequency: loss of a clear GCI, visible between two maxima of opening.

On the other hand, if the vocal tract shapes exhibits a high acoustical impedance, the assumption of source/tract decoupling is no more verified. The vibration of vocal folds is influenced by the tract, such as in woodwind or brass instruments [56]. The waveform of the vowel [u], corresponding to a particularly closed tract, is illustrated in Figure 2.10. We can observe a rather sinusoidal behavior, with no precise position of the GCI.

This aspect leads us to consider that in some situations – i.e. some segments of phonation within a large amount of connected speech or singing – it is interesting to take into account that we are no more looking for a GCI, but for a *Minimum of Glottal Opening*. MGO is defined in opposition to the maximum of opening, where A_v is evaluated.

The issue of considering GCI or MGO is discussed in Chapters 3 and 4.

[3] Maintaining durations of open and return phases during the increase of pitch, or maintaing the pitch value during the increase of duration of open and return phases, leads to the same result.

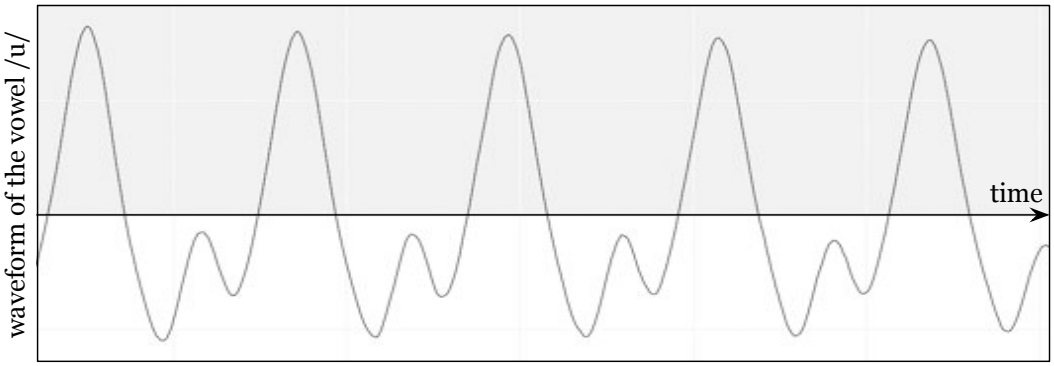

Figure 2.10: Waveform of a /*u*/ vowel, showing a overall sinusoidal behavior.

2.3 Perceptual aspects of the glottal flow

The definition of voice quality (cf. Section 2.2) remains something related to the context. In speech several studies classify voice production in different qualities, meaning perceptually relevant phonation types: modal, pressed, breathy, etc [131]. In singing it is much more related to techniques and nuances: falsetto, clear, opened, etc [23].

There is something flexible and interactive in proposing a sound space with dimensions, instead of classes. Several approaches propose to define the perceptual dimensions of the voice timbre, and moreover to connect them to the voice production level [72].

In 2.3.1 we propose to give an overview of voice quality dimensions. In 2.3.2 we present most commented relations between voice quality dimensions and voice production parameters. Moreover we highlight some inter-dimensional dependencies.

2.3.1 Dimensionality of the voice quality

We propose a list of dimensions that is directly inspired by the state of the art in voice quality perception. It aims at defining a common set of qualities that will be used in our analysis and synthesis work:

- *Pitch* is short-term and long-term inflections in the temporal evolution of the fundamental frequency f_0. [121, 29];
- *Vocal Effort* is a description of the amount of "energy" involved in the production of the vocal sound. Vocal Effort makes the clear difference between a spoken and a screamed voice for example [168, 93, 94];

- *Tenseness* is a description of the muscular pressure over the larynx. Tenseness makes the difference between a lax and a tensed voice [96];
- *Breathiness* is a description of the amount of air turbulence passing through the vocal tract, compared to the amount of harmonic signal [96, 120];
- *Hoarseness* is a description of the stability of sound production parameters (especially fundamental frequency and amplitude of the voice) [123];
- *Mechanisms* (M_i) are voice quality modifications due to type of phonation involved in sound production: mainly the chest or head voices decision [37].

2.3.2 Intra- and inter-dimensional mappings

It is difficult to have an exhaustive picture of relations between voice quality dimensions and voice production parameters, or between voice quality dimensions themselves. Gathering a significant amount of studies in this topic, we realize that a lot of links can be highlighted [120, 94, 9, 101]. We could consider that each voice quality dimension has an influence on each voice production parameter, to some extent. However we try to give a summary of the most relevant influences, as illustrated in Figure 2.11.

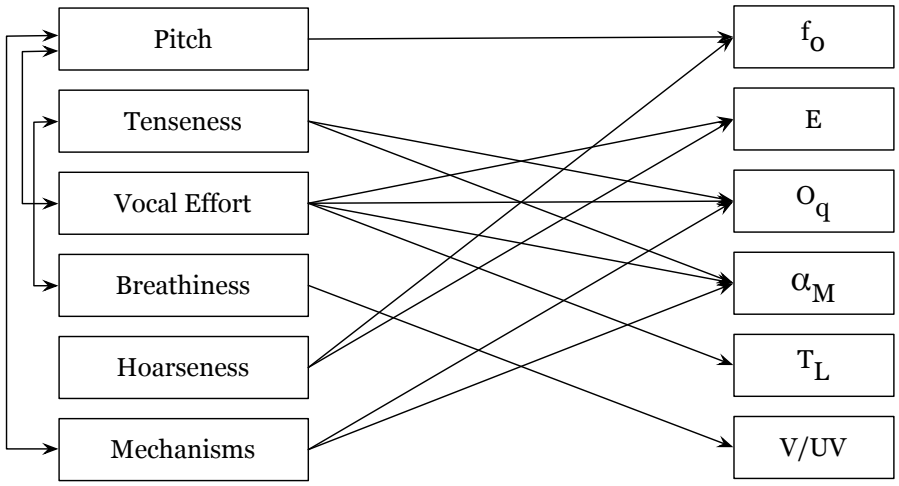

Figure 2.11: Summary of the main (not exhaustive) links that can be found in the literature between percepual dimensions of voice quality and production parameters.

- The relation between Pitch and f_0 is obvious. There are also strong dependencies between Pitch and both Vocal Effort and Mechanisms. These three dimensions are

linked into a diagram called the *phonetogram*[4]. The phonetogram can be observed by computing statistics of f_0 and energy over a large amount of sustained vowels. The dependancy with Mechanisms is more straightforward in singing, as the large range of notes to produce leads the singers to switch from chest (M_1) to head voice (M_2). In this case we use two phonetograms [97].

- There are several studies which highlight that the perception of Tenseness is related to the amount of first harmonic (H_1) in the spectrum. The relative variations between H_1 and H_2 are attributed to the position, amplitude and bandwith of the glottal formant, directly related to O_q and α_M [94].
- The influence of Vocal Effort on the glottal flow is more distributed. Vocal Effort changes intensity and brightness[5] [191] – thus E and T_L – but is also discussed as having an impact on opening and closing phases (O_q and α_M) [96].
- There is a physiological relation between Tenseness and Breathiness, which has been highlighted in some research [120, 9]. Indeed the more the vocal folds are relaxed (low Tenseness) the more they allow the circulation of free air flow through the glottis, thus increasing Breathiness, and vice-versa.
- The main influence of Hoarseness is the introduction of random deviations in fundamental frequency and intensity trajectories. These perturbations of f_0 and E are respectively called *jitter* and *shimmer* [123].
- Mechanisms are related to the way of using the vocal folds in the larynx [100]. M_1 corresponds to a vibration on the whole length, more favorable to low pitch and clear GCI. M_2 corresponds to a shortening of the length of vibration. M_2 leads to higher pitch and smoother opening/closing, thus influencing O_q and α_M.

2.4 Glottal flow analysis and source/tract separation

Changing the glottal source behavior of a recorded voice signal is still an open problem. It has been addressed by many people, in many different ways for the last fifty years [88,180], with various degrees of success. However no clear optimal solution has emerged, which would lead to a high-quality *voice expression codec*.

[4] As it is a important aspect of our realtime implementation, further the phonetogram, as well as relations between Pitch, the Vocal Effort and Mechanisms are described in Chapter 4.

[5] Brightness is a common perceptual measurement of the high frequency energy of a signal.

Most existing glottal flow estimation methods which only use recorded voice signals (i.e. non-intrusive techniques[6]) suffer from significant robustness problems [67]. Moreover these algorithms work only in some limited situations such as low pitch, low impedance (e.g. [a] and not [y]) sustained vowels, $F_1 > F_g$, clear glottal closure, etc.

We start, in Section 2.4.1, with an overview of the main problems related to the wrong manipulation of source components, from the analysis, transformation and synthesis of voice. Then we describe methods for glottal flow estimation and source/tract separation, in Section 2.4.2. Once a GF or GFD signal is accessible, several source parameters (cf. Section 2.2) can be estimated. Various methods exist and are explained in Section 2.4.3.

2.4.1 Drawbacks of source-unaware practices

It can be legitimate to question the need of a high-quality source-based coding/decoding in voice processing applications. This fundamental issue is addressed in the literature, but practical experimentations using voice analysis/synthesis quickly reveal limitations of the voice quality misunderstaning. Here we present a list of what we consider being the main problems.

Discontinuities in unit selection speech synthesizers

Some studies related to the synthesis quality of unit selection speech synthesis systems reveal that after some recording time, significant variations can be observed in the way a speaker pronounces read sentences. This is mainly due to the tiring of the vocal folds and a "relaxation" of the phonation. It results in timbral discontinuities when units have to be concatenated. With high-quality phonation compensation algorithms, such as GMM-based techniques [183], this effect can be alleviated .

Pitch modifications

Unintentional voice quality modifications due to pitch shifting are discussed in some papers. They happen even with formant preservation techniques, such as pitch shifting on LP residual [75] or PSOLA [122]. Indeed, in these techniques, pitch shifting changes the relative impact of open and return phases in the time domain, and thereby affects the voice quality by provoking unwilling laxness [111] or hoarsness [155] in the processed

[6] Best-known intrusive techniques are EGG [99] and videokymography [7].

voice. This problem is illustrated in Figure 2.12, with a pitch doubling on a LF-based glottal flow and with PSOLA.

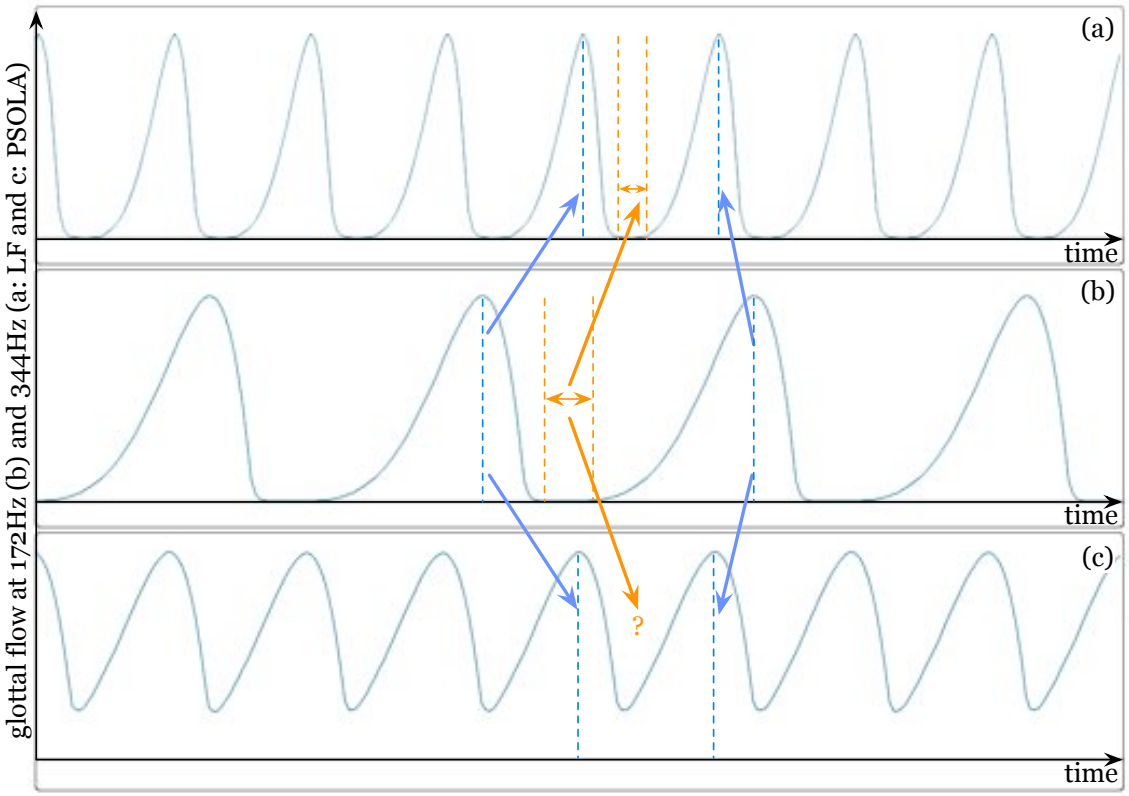

Figure 2.12: GF at 172Hz generated by integrating the LF model (b). Pitch is doubled by changing f_0 on the LF model (a) or by applying the PSOLA algorithm (c). We observe how the closed phase disappears (orange) with pitch shifting.

Errors in formant tracking

As Linear Prediction does not consider the mixed-phase representation of voice, it often happens that the glottal formant "attracts" the resonant poles of the LP estimation. If this LP analysis is used as a basis in order to track formant frequencies, as it can be done in some speech recognition systems, the trajectory of the first formant can be lost, depending on F_1 and F_g values [25]. In Figure 2.13 we can see that estimated F_1 and F_g have a quite common behavior in the syllable [bɔ̃] of the french word "bonjour". It leads us to consider that F_g influences the tracking of F_1 achieved by LP analysis[7].

[7] In this example, F_g is estimated using the algorithm explained in Chapter 3 and [50].

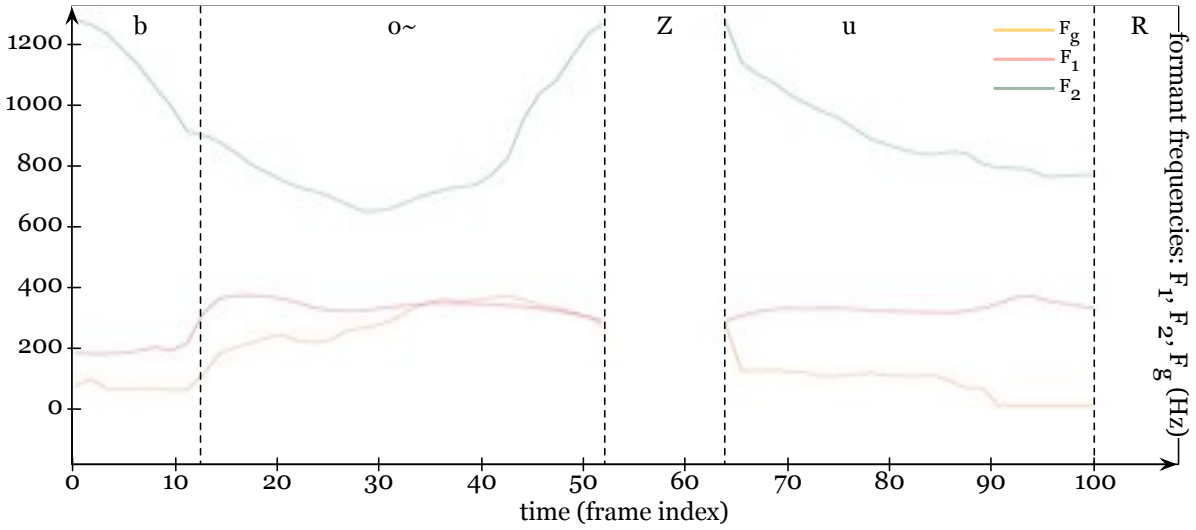

Figure 2.13: Evolution on 100 pitch-synchronous frames of the glottal formant (yellow), and first (red) and second (green) vocal tract resonances. We can see the confused area in the syllable [bɔ̃]. F_g is estimated by the algorithm described in [50].

Buziness in resynthesis

The mixed-phase representation of voice production shows that the time domain evolution of the voice waveform is a subtile overlapping of maximum-phase resonant shapes, return and closing phases, and minimum-phase multi-resonant vocal tract responses. Moreover some hearing tests highlight that we have a significant perception of phase information in transitory segments of speech [150]. When a voice signal is resynthesized, inconsistencies in the time domain sequencing of these acoustic phenomena (meaning phase mismatching) provoke typical "source/filter like" undesired buziness.

2.4.2 Estimation of the GF/GFD waveforms

In this part, we give an overview of GF/GFD estimation techniques. These algorithms aim at retrieving the GF or GFD as a time domain signal, supposing the effect of the vocal tract has been removed. The source/filter theory has a particularly interesting consequence, in the field of glottal flow waveform retrieval. Indeed the source/filter model explains the voice signal as the result of a convolution, as in equation (2.2). Both the well-known Fant model [78] and the recent mixed-phase version make the assumption that the effect of the vocal tract $V(z)$ can be removed by deconvolution[8].

[8] Concerning the mixed-phase model, the return phase of the glottal source is embedded in the minimum-phase component. Consequently only the effect of the glottal formant can be isolated.

In this context, we present two different ways of addressing this deconvolution problem. On the one hand, a large amount of algorithms analyse the voice spectrum with LP for its well-known performance in formant parametrization, and also as a way of removing the periodicity $\uparrow\uparrow\uparrow$ (z). We can highlight two categories of techniques: iterative LP estimation and LP estimation achieved on the closed phase. On the other hand, a new technique has been deduced from the mixed-phase model. It separates voice frames into causal and anticausal components, using the zeros of the z-transform.

Iterative LP estimation

As in Fant's theory [77], the main idea of the LP analysis is that both glottal source and vocal tract spectral envelopes can be approximated by an all-pole filter, as described in equation (2.1). Consequently several methods have been developed in order to estimate iteratively glottal source and vocal tract poles, combining LP and inverse filtering.

In [5] the *Iterative Adaptative Inverse Filtering* (or IAIF) method is used on pitch-synchronous frames. IAIF is a two-pass algorithm sucessively computing LP estimations of the glottal source and the vocal tract. A first pre-processing filter (high-pass) is applied in order to remove low-frequency fluctuations due to the microphone. Then two iterations of the following process are achieved. The output of the first iteration is reused as the input of the second one, with different adjustements of LP orders:

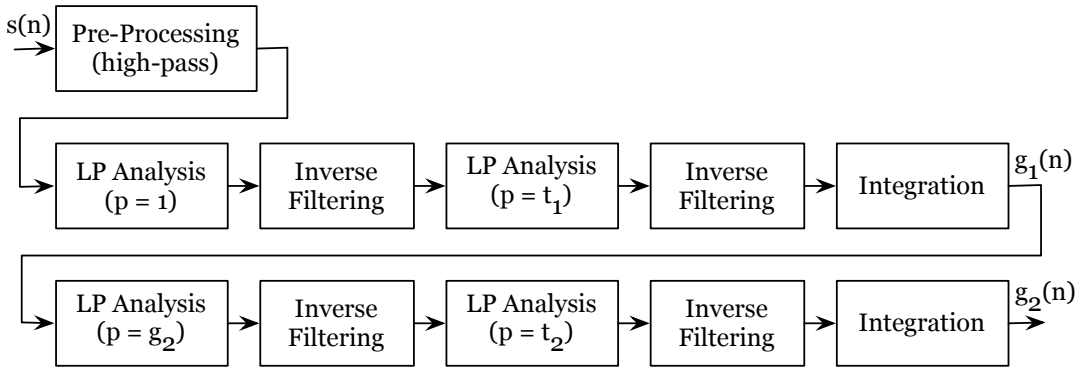

Figure 2.14: Block diagram of the two-pass IAIF algorithm, as a way of estimating the glottal flow iteratively, $g_1(n)$ then $g_2(n)$, from the voice signal $s(n)$.

1. estimation of the source spectrum by low-order LP on the voice signal;
2. inverse filtering in order to remove this source component;
3. high-order LP estimation of the vocal tract on the residual;

4. inverse filtering in order to remove the vocal tract component;
5. integration of the residual signal in order to get the glottal flow.

Figure 2.14 illustrates the two iterations in order to refine estimates of the glottal flow, $g_1(n)$ then $g_2(n)$, from the voice signal $s(n)$. PSIAIF [5] uses $g_2(n)$ as a way to place markers on each GF period. From these markers, IAIF is relaunched pitch-synchronously. This improvement refines the analysis by providing one glottal pulse estimate (through the position of the g_2 poles) by period of length T_0. We can also highlight the updated PSIAIF described in [8], and using Discrete All Pole modeling (DAP) [76] instead of LP. Results are compared to videokymography images in [7].

Arroabarren's method uses a similar idea [12], but using the Klatt's *KLGLOTT88* model [120], instead of LP estimates. The simplicity of *KLGLOTT88* allows a first compensation of the glottal source, by substracting it in the spectral domain. Then the residual is used to get a first spectral model of the vocal tract, by DAP analysis. The spectral tilt effect is evaluated by observing the real poles of the tranfer function, and then removed. Finally this corrected estimation of the vocal tract is used for another inverse filtering, in order to get an estimate of the glottal source derivative. This process is achieved for several values of O_q (open phase of the *KLGLOTT88* model) and the solution which minimize the glottal formant ripple is chosen (cf. Figure 2.15).

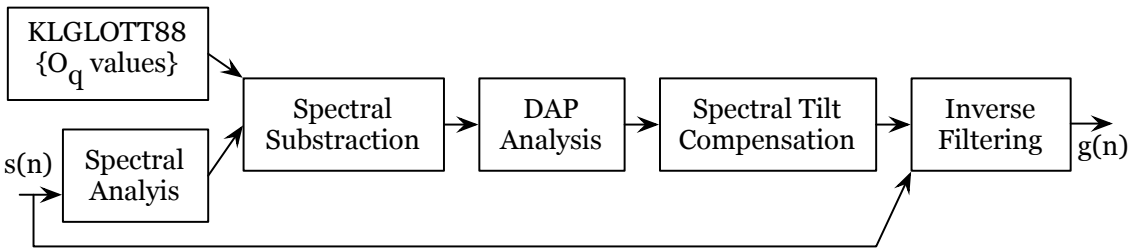

Figure 2.15: Block diagram of the iterative Arroabarren's algorithm, changing the O_q of a KLGLOTT88 model, in order to obtain the best $g(n)$ by inverse filtering.

LP estimation on closed phase

The LP estimation on closed phase makes the assumption that the voice waveform is strictly due to the vocal tract transfer function during the time when vocal folds are closed. Thus it appears to be the appropriate moment to compute LP analysis and evaluate the vocal tract. Then the LP estimation of the vocal tract can be inverted and

used as coefficients for inverse filtering, in order to get an estimation of the GFD. In this Section we notice three main methods going in that direction.

The first one is proposed by Wong *et al.* [201]. After high-pass pre-processing, LP analysis (covariance method) is used over the whole voice signal. The length of the analysis window is fixed and the window is shifted sample by sample. Each frame contributes to a total squared error signal $\alpha_M(n)$. The energy of the prediction error is normalized by the energy of the voice signal $\alpha_0(n)$, giving the *Normalized Error Criterion*:

$$\eta(n) = \frac{\alpha_M(n)}{\alpha_0(n)}$$

Thanks to pitch estimation, a period-by-period observation is achieved and it appears that minimal values of $\eta(n)$ are synchronized with closed phases. Closed phases are located and a second LP analysis is achieved within their boundaries in order to estimate the vocal tract. Real and high-bandwidth poles are removed from the transfer function. Finally the whole voice signal is inverse filtered and an estimation of the GF can be observed (after integration). The process is illustrated in Figure 2.16.

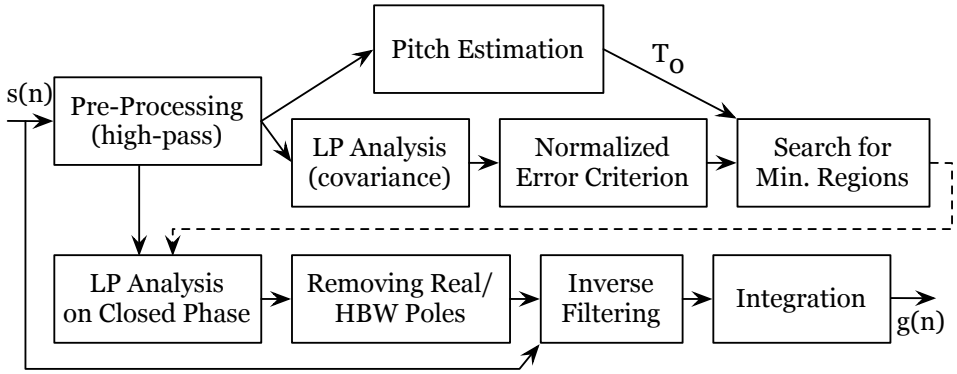

Figure 2.16: Block diagram of the Wong's algorithm, inverse filtering the voice signal after LP estimation of the vocal tract on closed phases, thanks to *min* $\eta(n)$.

Childers follows a similar idea in [43], expect for the localization of the closed phase. This work compares signals coming from electroglottography – exhaustively used in [99,101] – with the LP residual of the corresponding vowel. Some synchronicity between negative peaks of the derivated EGG and spikes in the residual is found. Consequently they decide that the closed phase starts after each peak located on the residual and stops at 35% of the interval between two peaks. The rest of the process is similar to [201].

The third approach that we want to examine is Plumpe's algorithm [151], based on a quite uncommon aspect of phonation. From the physiological point of view, we can consider that the opening and closing of the vocal folds respectively lengthen and shorten the overall length of the vocal tract, by adding a small subglottic section. Such a variation in length has an impact on formant frequencies, which is a coupling effect ignored by the source/filter model. Consequently the closed phase is the only phase in which formant frequencies do not shift.

Plumpe's approach consists in tracking formant frequencies on short-term LP analysis – thus preferably covariance-based[9] – and locating closed phases by targeting the most stable regions. This process is achieved in two steps. A first marking is done by peak picking on the LP residual, achieved pitch-synchronously. A second LP analysis is performed around the resulting peaks, in order to get a favorable estimation of formant frequencies. Formant frequencies are tracked with a Viterbi algorithm. Stable regions are located, then extended. From the final marking, a third LP analysis on the closed phase is performed, followed by inverse filtering, as done in other papers [201, 43]. This process is illustrated in Figure 2.17.

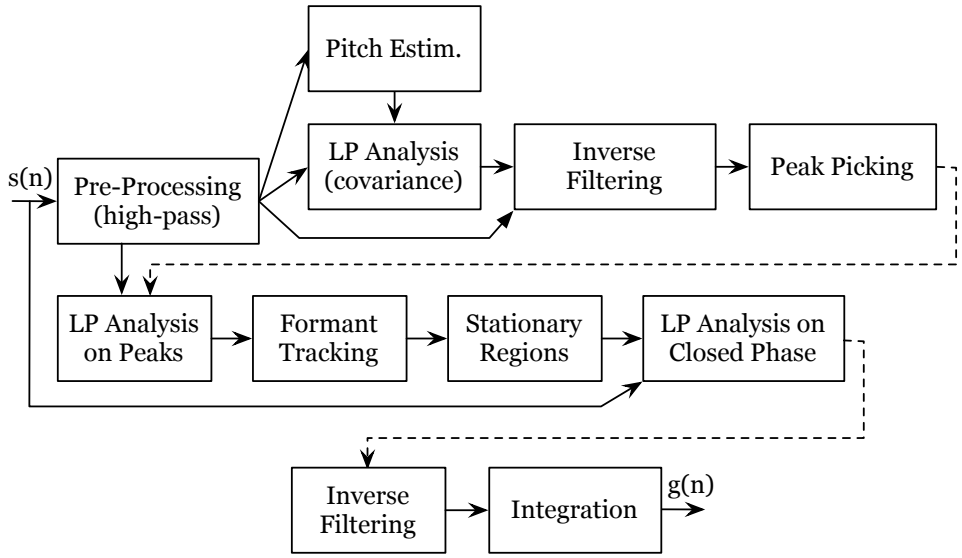

Figure 2.17: Block diagram of the Plumpe's algorithm, inverse filtering the voice signal after an LP estimation of the vocal tract on closed phases. Closed phase are estimated by locating stable regions on formant frequency trajectories.

[9] Indeed it is shown that autocorrelation-based and covariance-based LP have similar performances for long analysis frames. Covariance-based LP analysis becomes more relevant for short frames.

Zeros of the Z-Transform

Considering the mixed-phase model of speech, we can consider that causality is a discriminant factor in order to separate a part of the glottal source signal (the open phase). Here we describe the algorithm used in order to achieve a first separation of anticausal and causal contributions, using zeros of the z-transform (ZZT) [26].

For a series of N samples $(x(0), x(1), ..., x(N-1))$ taken from a discrete signal $x(n)$, the ZZT representation (zeros of the z-transform) is defined as the set of roots (zeros of the polynomial) $\{Z_1, Z_2, ..., Z_m\}$ of its z-transform $X(z)$, as illustrated in equation (2.15).

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} = x(0)z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \tag{2.15}$$

This representation implies to compute roots of polynomials [74] whose degree increases with the sampling frequency. This tends to introduce errors on the estimation of zeros in high frequencies, due to the iterative computation of roots. For this reason, ZZT computation is usually performed at 16kHz. Speech sampled with higher frequency has to be downsampled for ZZT estimation.

The mixed-phase model of speech implies that the ZZT representation of a speech frame contains zeros due to the anticausal component and to the causal component [27]. Consequently zeros due to the anticausal component lie outside the unit circle, and zeros due to the causal component inside the unit circle. Under some conditions about the location, the size and the shape of the analysis window, zeros corresponding to both anticausal and causal contributions can be properly separated by sorting them out according to their radius in the z-plane, as illustrated in Figure 2.18. Bozkurt recommands the use of a Hanning-Poisson window, centered on the GCI, with a length of $2 \times T_0$.

The spectrum of each contribution is obtained by computing the influence of the cloud of zeros on several points distributed on the unit circle, as described in equation (2.16). Time domain waveforms are obtained by applying IFFT on both components.

$$X(e^{j\phi}) = Ge^{j\varphi(-N+1)} \prod_{m=1}^{N-1} (e^{j\varphi} - Z_m) \tag{2.16}$$

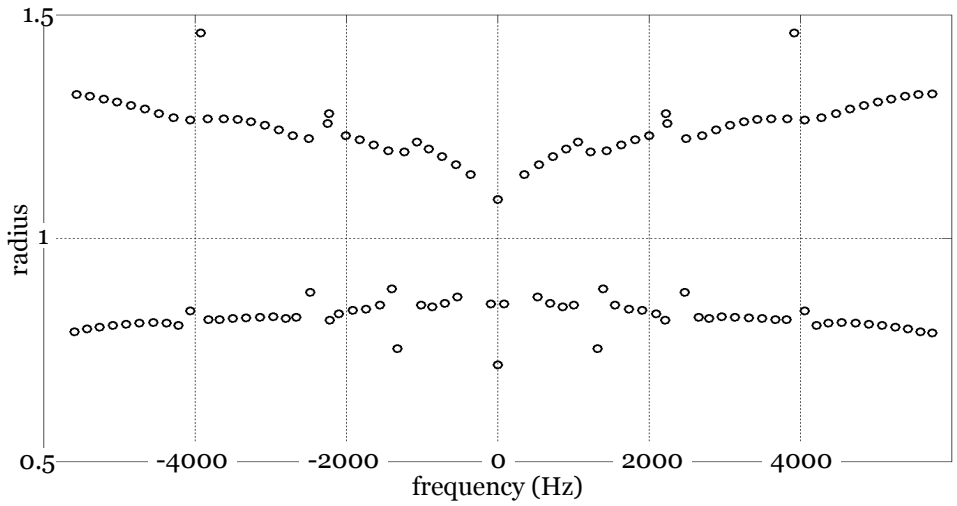

Figure 2.18: Distribution of Z_m in the Z plane in polar coordinates, showing that inner and outer zeros can be sorted out, here on a synthetic speech frame.

Applying this algorithm on typical voice segments, we identify *causal* and *anticausal* frames resulting from the ZZT-based decomposition around the k^{th} GCI, respectively by $x_{C,k}$ and $x_{A,k}$. Examples if these frames are illustrated in Figure 2.19.

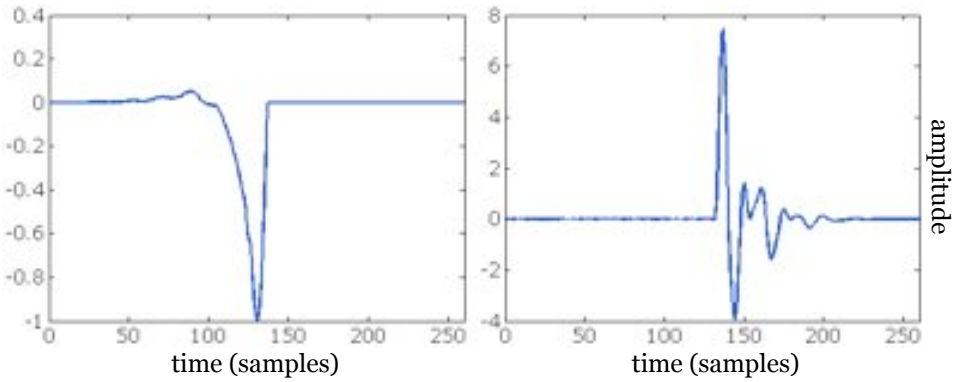

Figure 2.19: ZZT-based decomposition on a real speech frame of a [a]. We see that $x_{C,k}$ is causal (right) and $x_{A,k}$ is anticausal (left).

It has been highighted in several papers that ZZT-based source/tract decomposition is not particularly robust and often presents noisy decomposition results [68, 67]. Along this thesis, different ways of avoiding this noise are tested (cf. Chapter 3).

2.4.3 Estimation of the GF/GFD parameters

The methods presented in the previous Section provides an estimation of the GF or GFD; it is now interesting to extract some parameters from these waveforms, in order to quantify voice quality.

Some techniques are available, based on direct measurement of key points. The idea of fitting a more complex GF/GFD model on estimates will also be addressed below, with both time and frequency domain fitting strategies. We also present a cluster of techniques which jointly estimate source and tract parameters.

Measurement of key points in the time domain

Some algorithms try to estimate GF parameters by directly measuring key points on estimated signals. It is achieved by locating important landmarks like zero crossing, maxima or minima within a fundamental period [4]. More specific algorithms exist, like those based on estimating the *Normalized Amplitude Quotient* (or NAQ) [13, 2]. These techniques, however, are not robust to noise, especially for estimating zero crossing positions. An example of landmark location within PSIAIF is illustrated in Figure 2.20.

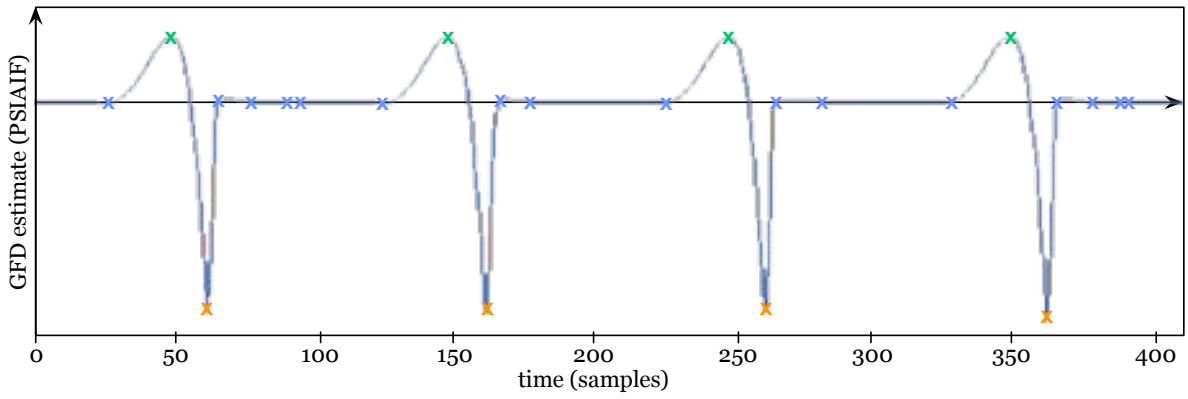

Figure 2.20: Location of maxima (green), minima (orange) and zero crossings (blue) on the GFD estimate corresponding to a normal [a], achieved with PSIAIF [4].

Fitting in the time domain

Fitting a model in the time domain has also been explored, for the last 15 years [157,181], with advantages and drawbacks. Most of them use a non-linear glottal source model, and mainly LF [78]. Curve fitting is performed, with non-linear least squares estimation

techniques. The most popular technique is to follow the *gradient* of the error function $\epsilon = f(model\ parameters)$. Starting with a first approximation (often achieved by direct measurement), it iterates until the error function exhibits a minimum: the gradient gets close to zero.

After a first measurement of some time domain parameters on the signal, Childers minimizes the error separately and interatively on open and return phases [41]. Strik [181] and Lu [131] low-pass the estimate by convolving it with a 7-point Blackman window, in order to remove the noise and ripple. Iseli's method [108] minimizes the least squares error thanks to the Newton-Raphson algorithm. Finally Plumpe's algorithm [151] uses a particular non-linear least squares regression technique, called NL2SOL. It allows the setting of constraints in order to avoid physically unrealistic configurations.

Taking their information from the time domain waveform, these techniques have a lot of problems with phase distorsion. Indeed we know that LP-based inverse filtering on real voice signals is particularly weak in interpreting the phase information.

Fitting in the frequency domain

Considering that most of the GF/GFD estimation methods use LP analysis as a way of removing the vocal tract contribution, some algorithms use an all-pole modeling of the source in order to extract GF/GFD parameters [110, 86, 98]. Indeed the GFD spectrum can be seen as a second-order resonance. A low-order LP analysis on the GFD estimate allows the parametrization (center frequency, amplitude and bandwith) of the glottal formant, and thus the time domain parameters like O_q [64].

LP analysis has also been used with the mixed-phase model of speech [25]. Indeed it has been shown that the glottal formant can be tracked and parametrized by observing anticausal poles that appear in a covariance-based LP analysis of speech frames [28].

Other approaches exist. In [6], Alku defines a parametric parabolic curve and fits it to the low-frequency bands of the estimated source spectrum. Let us mention also Oliveira's spectral fitting [148], based on the spectral representation of the Rosenberg model.

Join estimation of source and tract parameters

In the source/filter model [77], a sample $s(n)$ of speech is modeled by the auto-regressive (AR) equation (2.17), where $u(n)$ and $e(n)$ are samples of the source and the residual and $a_n(i)$ and b_n are the AR filter coefficients representing the vocal tract.

$$s(n) = -\sum_{i=1}^{p} (a_n(i)s(n-i)) + b_n u(n) + e(n) \tag{2.17}$$

This formulation of the predictive equation integrates a model for the source, and is called *Auto-Regressive with eXogenous input* (or ARX), because the input signal $u(n)$ is no more white (impulse train or white noise). However this change prevents the use of Yule-Walker equations. Instead one has to solve the system of linear equations 2.18 obtained by writing equation 2.17 for successive values of n:

$$S = MA + E \tag{2.18}$$

where S is a vector of (possibly windowed) speech samples $s(n)$, M is the concatenation of a matrix of $-s(n-i)$ values and a vector of glottal source samples $u(n)$. A is the vector of unknown values $a_k(i)$ and b_k. E is a vector of residual samples $e(n)$: the vector of modeling errors that we want to minimize when computing A.

There are several ways of implementing ARX-based parametric estimation, based on various glottal source modeling. In [61] the Rosenberg-Klatt model (or RK) [120] of the GF is used, and the joint estimation is achieved by an adaptative procedure, based on Kalman filtering. We also find some work using the LF model in a similar procedure [83]. Fu introduces a two-step algorithm, first using the Rosenberg model as a way of initiating the estimation, and then the LF model, in order to get more precise values of GF/GFD parameters [84].

Vincent *et al.* also use the LF model of the GFD [193, 194]. In their work, finding the unknown values $a_k(i)$ and b_k requires to define a set of glottal sources $\Theta = [u_1 \dots u_W]$ and to choose among these the one which minimizes the modeling error of the ARX model. In other words, it requires to solve the system of equations for each u_w and to

select the one that minimizes $\|E\|^2$. That glottal flow u_w minimizing the modeling error is considered as the most accurate estimate of the actual glottal flow produced by the speaker. Parameters $a_k(i)$ and b_k, as well as position of the GCI, are also refined in the optimization process.

Estimation glottal flow parameters on the voice signal

It is often commented that parameter extraction on the LP residual is not robust to noise and phase distorsion. This is why some techniques target the measure of GF/GFD parameters directly on the voice signal. Indeed glottal formant and spectral tilt have visible effects on the voice spectrum [63]. We know that the glottal formant has a significant impact on relative values of H_1 and H_2 (first and second harmonics of the spectrum. In [94] and [96] these relations are highlighed, and particularly the link with O_q with equation (2.19). Iseli proposes a corrected equation (2.20) in this equation, in order to compensate the effect of the vocal tract, by the evaluation of F_1 [109].

$$(H_1 - H_2) = -6 + 0.27 \times e^{5.5O_q} \tag{2.19}$$

$$H^* = H - 20log_{10}\frac{F_1^2}{F_1^2 - f^2} \tag{2.20}$$

2.5 Background in singing voice synthesis

Being a convergence between voice synthesis technologies and requirements of live performing arts, this thesis naturally addresses the field of singing voice synthesis. Singing voice synthesis research aims at artificially producing novel[10] singing voice utterances.

With the development of computer-based technologies, various digital synthesis techniques have emerged and have been used for the synthesis of the singing voice: source/ filter, harmonic plus noise, digital waveguides, unit selection, etc. In this Section, we give an overview of the most significant systems, from seventies to today.

[10] In this case, "novel" means that does not exist on any recording support. The novelty of the utterance is based on the idea that it results from an arbitrary request from the user.

MUSSE (DIG): formant-based source/filter synthesis

MUSSE is probably the first singing voice synthesizer. It is released in the seventies and is based on an analog parallel formant synthesizer, driven by a set of rules [126]. Later this rule-based paradigm has been transposed on computers, in order to release MUSSE DIG [19]. As every formant-based system, MUSSE (DIG) has this typical robotic sounding, but has a remarkably small footprint in memory and CPU.

CHANT: formant wave functions

CHANT is developed at Ircam, by Rodet *et al.* in the eighties [159]. CHANT also uses the idea of parallel formants, but in a different way. Indeed each formant is represented by its impulse response, and is excited by a pseudo-periodic controlling source. Modifications applied on these Formant Wave Functions (FOF) [158] lead to changing the spectral envelope of the formants. The nice sounding of CHANT is based on refined adjustments of these control parameters, based on singing voice analysis.

SPASM: digital waveguides

In the early nineties, the computer music research starts to adapt the concept of digital waveguides [177] to musical purposes, particularly for creating computationally light physical models: string, plate, etc. With SPASM, Cook extends the idea of physical waveguide to the modeling of the whole vocal tract [45]. SPASM integrates an interesting model for the interaction between nasal/oral cavities and glottal source reflections. This system is particularly efficient for liquid consonants and nasal sounds.

Lyricos: diphone concatenation and sinusoidal modeling

In the nineties, the great success of MBROLA [70] has shown that the combination of diphone concatenation and HNM is particularly efficient for the natural sounding of the voice. Lyricos uses the same idea, applied to singing contents [132]. Added to the phonetic target, a score is used in order to build the prosody of the singing voice. The main drawback of this system is the metallic sounding encountered on long vowels.

Unit selection scheme applied to singing voice contents

In speech as well as in singing synthesis, the increase of computer capacities in the late nineties allowed the use of larger units than diphones. Meron adapted the idea of non-uniform unit selection (that was getting successful in speech synthesis) to the synthesis of the singing voice [139]. Meron's system analyses and segments a large database of singing performance (one singer) and a retrieving algorithm is trained in order to concatenate these units at the synthesis time. One main drawback is the huge size of the database.

SMS: performance sampling and harmonic plus noise modeling

Bonada *et al.* probably propose the state of the art in high quality singing synthesis. The basis is the *Spectral Modeling Synthesis* (SMS) [173]. This technology performs the pitch-synchronous framing and harmonic plus noise modeling of a large amount of singing voice material. The high quality of the synthesis results from the interpolation of the phase (such as phase-locked vocoding techniques) and the representation of source and tract components within the HNM framework [24]. This synthesis technique has been used as the engine of the successful commercial product *Vocaloïd* [195].

STRAIGHT: speech-to-singing conversion

STRAIGHT is a recent speech synthesis system. This technology uses a smoothing between spectral envelopes, a new estimation of the fundamental frequency and harmonics, and measurements on the group delay as a way of estimating the glottal source activity [115]. STRAIGHT has been used for the conversion from speech to singing. Mainly this is based on the pitch shifting of the speech sound into a singing melody, and the modification of spectral envelopes in order to simulate the singing formant [165].

HTS: stochastic control of production models

For the last five years, the HTS technology [189] attracted the interest many researchers. HTS uses a new idea for the production of speech. The algorithm relies on a well-know production model, source/filter or harmonic plus noise, but the production parameters are controlled by a stochastic process: Hidden Markov Models. These HMMs are trained on a huge amount of data. Recently this technique has been adapted to the synthesis

of singing in japanese [164]. One main advantage of HTS is that the database is not needed at the runtime, which significantly reduce the footprint of this system.

Chapter 3

Glottal Waveform Analysis and Source/Tract Separation

" I don't believe in fundamental frequency. "
— Yannis Stylianou

3.1 Introduction

In Section 2.4 we have seen that the analysis of the glottal source behavior on recorded voice segments is an ongoing reseach topic, addressed by various techniques. Most of these approaches expect to estimate GF or GFD waveforms with high precision, and apply parametric estimation on them. However most of current systems propose solutions only assessed on synthetic speech or on sustained vowels [5, 26].

The RAMCESS analysis framework is in the continuity of these approaches. However, as the purpose is different from that of voice quality analysis, the method also differs. Indeed we analyze a database with the only purpose of using extracted voice production features (glottal source parameters and vocal tract impulse response) within an interactive synthesis engine, which is explained in Chapter 4.

Following the definition of expressivity that we proposed in Section 1.1, the expressivity of the synthetic voice rather results from the interactive control of our synthesis engine, than expressive properties of the database itself. Consequently we prefer a database

with limited voice quality variations, and our source/tract separation algorithm takes the benefit of this stability in order to be more assessable and work better.

In this Chapter we propose an evolution of the well known ARX-LF estimation algorithm, explained in [193] and 2.4.3. This evolution introduces the use of the ZZT-based causal/anticausal decomposition – [26] and 2.4.2 – as a pre-processing aspect of the ARX-based source/tract joint estimation. The use of ZZT shapes our RAMCESS analysis framework as a two-step model fitting strategy. First, the glottal formant is estimated by fitting the LF model [78] on the anticausal component, obtained from ZZT. Then a modified version of ARX-LF is used in order to model the whole speech. The whole analysis pipeline, with the framing and the two-step modeling, is illustrated in Figure 3.1.

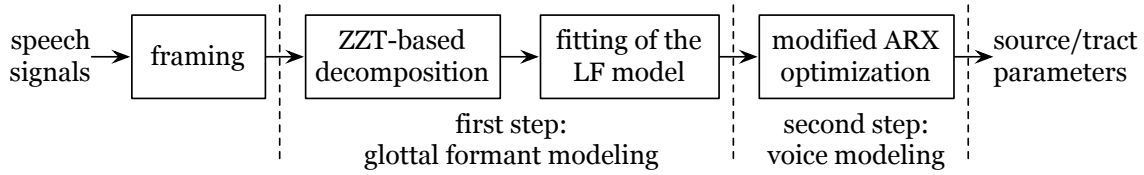

Figure 3.1: Diagram of the RAMCESS analysis pipeline: voice signal framing, ZZT-based causal/anticausal decomposition, fitting of the LF model [78] on the anticausal component, and modified ARX optimization.

This Chapter starts, in Section 3.2, with a explanation about the database we have created, as our speech corpus is recorded and segmented in a particular way. Then we propose a discussion about the validation of source/tract separation results in Section 3.3. Moreover we highlight some results related to ZZT-based analysis. This is also an opportunity to introduce some dedicated validation criteria. Section 3.4 describes our procedure for the estimation of the glottal formant. Section 3.5 presents our modified version of ARX-LF. Finally, the performance of the analysis is discussed in Section 3.6.

3.2 Working with connected speech

In this work we run ZZT-based algorithms on a limited-size connected speech database. Our corpus consists in various sentences in French and English. We use three different corpus sizes in this thesis:

- When we work at the frame level, we consider small segments of vowels and consonants in order to quickly iterate on prototyping.

- In the RAMCESS synthesizer, we work with a database consisting in 8 sentences pronounced by one speaker in English.
- When we compute statistics related to the analysis process, we use 38 sentences, gathered from 3 separate speakers.

Some additional constraints are imposed (flat pitch, constant effort, etc), in order to facilitate the analysis of voice quality. These constraints influence the recording protocol (3.2.1), the database segmentation (3.2.2) and the GCI marking (3.2.3). The database segmentation is also refined by an inter-phoneme segmentation (3.2.4). In the following paragraphs, we describe these specifications, in order to make the setting of a RAMCESS-friendly analysis framework reproducible in further research.

3.2.1 Recording protocol

In a recent paper, Walker discusses the importance of verifying the recording conditions in the making of voice corpus, that will be used for glottal flow analysis [197]. In this part we present the protocol that we have defined in order to verify these recording conditions and facilitate the analysis process, and finally achieve a corpus of limited size, while containing various kinds of phonetic articulations.

Adapting recording conditions to phase processing

Let us mention several recommandations that are used in this work:

- We drop any kind of pre-processing, hardware or software, such as high-pass filtering, compression or noise reduction. Indeed we expect the phase information (easily distored by pre-processing) to be as preserved as possible.
- We reduce the overall amount of electrical and acoustical noise by using a dynamic directional microphone, XLR connections, a high-quality analog-to-digital converter, and achieving the recording in a low-reverberation room.
- The microphone is placed at 80cm (at least) of the mouth in order to reduce the low-frequency bias due to breath burst on the microphone (cf. Figure 3.2).
- The speaker is asked to sit on a chair and look at a given target point. The aim is to stabilize both the directivity of the mouth and the microphone.

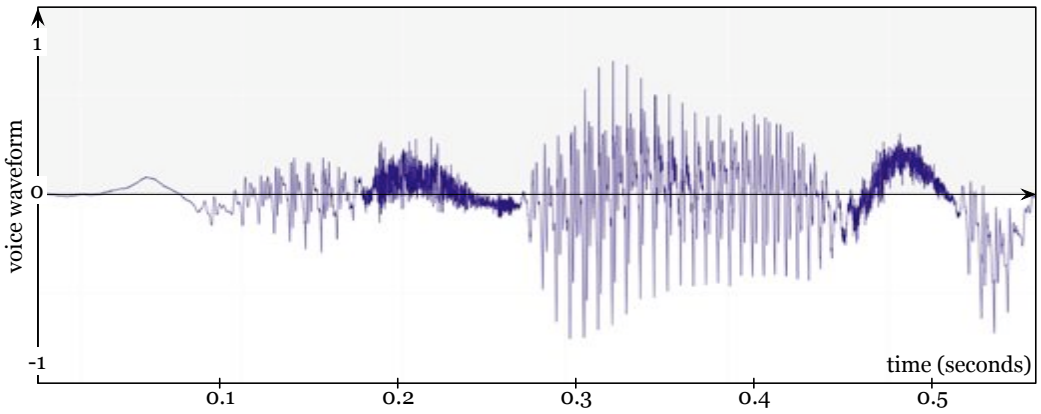

Figure 3.2: Waveform of connected speech with typical offset bursts on unvoiced consonants. Bursts are due to the small distance between the mouth and the microphone.

Leading the speaker by iterative stimulus synthesis

As opposed to usual conventions used in the analysis of expressive voice, our aim is not to maximize the expressivity of the database. Indeed the expressive aspect is brought by the realtime gestural control. Moreover as we want the analysis to work on the largesy possible part of the database, we consider it is more relevant to minimize the variability of the phonation (pitch, glottal flow features, etc) along the whole database.

However, since it is not possible to manually check the phonation quality of a speaker during a whole recording session, we insert him/her in an automatic loop. The loop alternates between playing stimuli and recording the speaker's voice. Indeed we use mimicking capacities of the speaker confronted with a synthetic stimulus in order to maintain his/her phonation inside a given range. Mainly pitch is used as the leading parameter, due to its high correlation with other voice quality dimensions [100].

In Figure 3.3 we give the details of the recording protocol. The aim is mainly to maintain the intonation (and hopefully the overall voice quality) as stable as possible. The following sequence is repeated for a given amount of requested sentences:

1. the expected corpus is converted into the corresponding phonetic and prosodic transcription, with the help of a state-of-the-art phonetizer [17];
2. a flat pitch curve is imposed;
3. the duration of each vowel is multiplied by a factor > 1 (typically 1.5), in order to generate a region of stable voiced phonation in each vowel;

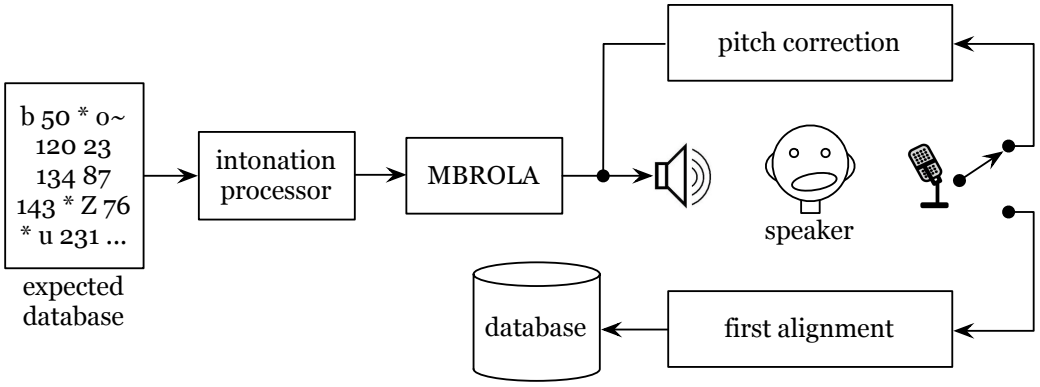

Figure 3.3: Diagram of the recording protocol used for the RAMCESS database. The speaker is inserted in an automatic loop where stimuli are played (synthetic then corrected real voice), and his/her mimicking is recorded right after the playing.

4. the modified target is sent to the MBROLA synthesizer; this operation can be achieved in realtime with the MBROLA external object for Max/MSP [51];
5. the speaker hears this first stimulus and is asked to mimick it;
6. his utterance is recorded, and sent to a pitch correction module which replays the recorded utterance with the same flat pitch line as in the synthetic stimulus;
7. the speaker hears this second stimulus and is asked to mimick it;
8. this final utterance is recorded and stored into the RAMCESS database, with a first approximation of the phonemic alignment.

The use of synthetic voice stimuli as a way of leading the recording session is particularly efficient in the controlling of the fundamental frequency. In Figure 3.4 we highlight that non-assisted recording leads to wide and unfocused gaussian pitch[1] distribution around the natural tone of the speaker. While using the stimulus-based recording session, the pitch distribution is a narrower gaussian around the expected pitch target.

A database that exhibits this kind of narrow pitch distribution is easier to pre-process. Indeed, due to the use of ZZT, most of the analysis that we achieve on the database requires efficient pitch marking. With a narrow range of possible f_0 we can better correct problems such as H_2 detection (pitch doubling), improve the voiced/unvoiced detection, and thus better perform pitch-synchronous operations, such as GCI marking.

[1] In this work, we use YIN [58] as the pitch detection algorithm.

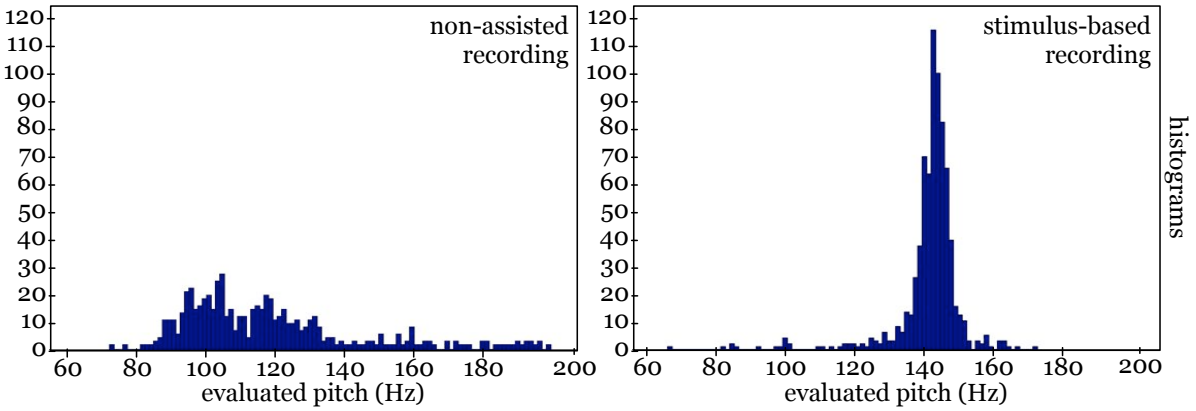

Figure 3.4: The left histogram represents the distribution of the pitch values in a non-assisted recording session. The right one represents the distribution of the pitch values in a stimulus-based recording session with a flat pitch target of $f_0 = 140$Hz.

Finally we can highlight that the recorded sounds rely on a precisely known phonetic/prosodic target (symbols sent to MBROLA). Obviously we can not consider that the speaker is able to exactly reproduce this target, but it gives a first approximation for the alignment of the phonemes on the waveform.

3.2.2 Phoneme alignment

The recording protocol (3.2.1) provides a first approximation of the phonemic segmentation of the recorded utterances. This first proposal is then manually corrected in order to perfectly fit the waveform. The size of our database allows this manual correction.

From this phoneme segmentation, an annotation file is encoded, which associates every phone of the database with a specific phonetic symbol and other linguistic and phonologic information: vowel/consonant, voiced/unvoiced, fricative/plosive, etc.

3.2.3 GCI marking on voiced segments

Within the phoneme-segmented database, voiced parts can be processed separately in order to get a first estimation of the GCI positions. Our approach is to focus on unvoiced/voiced or silence/voiced transitions. For example, respectively [ʃɛ] or [_lɛ].

At these transitions, the onset of the *voiced island*[2] is clearly visible, as we can see in Figure 3.5. This onset is the first GCI of the island. Indeed, due to previous unvoiced contents, this first GF cycle is not yet overlapped with vocal tract impulse responses.

Practically this means that the first GCI of the island – that we propose to call GCI_1 – can be located by a negative peak search after the unvoiced/voiced segmentation point. The searching technique is described in equation (3.1).

$$GCI_1 = \underset{n=[L,L+1,5T_0]}{\operatorname{argmin}} \; x(n) \tag{3.1}$$

where $x(n)$ is the signal, T_0 is the local fundamental period, and L is the segmentation point starting the current voiced island. The searching area is $1,5 \times T_0$ after the segmentation point. This searching area corresponds to a good compromise, considering that the manual segmentation. Indeed it has been observed that manual segmentation points are usually set slightly before GCI_1 locations.

From the position of GCI_1 and the estimation of the fundamental frequency along the signal, other GCI positions – referenced as GCI_k – can be extrapolated. The extrapolation works in two successive steps, achieved for each value of k:

1. GCI_k location is first defined as $GCI_{k-1} + T_0(k)$;
2. GCI_k location is refined by searching the local negative peak (if any).

The pitch-based extrapolation is pursued until we meet a new unvoiced island. There, the GCI_1 searching process restarts. The operation is repeated until the whole database has been processed.

We work in a GCI-synchronous framework. This means that the k^{th} analysed frame is centered on GCI_k. Moreover the window length is set at $2 \times T_0$ (two times the local fundamental period). In the sequel, we denote $x_{V,k}$ is the *voice* frame extracted around the k^{th} GCI. The whole process is illustrated in Figure 3.5.

[2] Voiced island refers to a region of the speech which is continuously voiced.

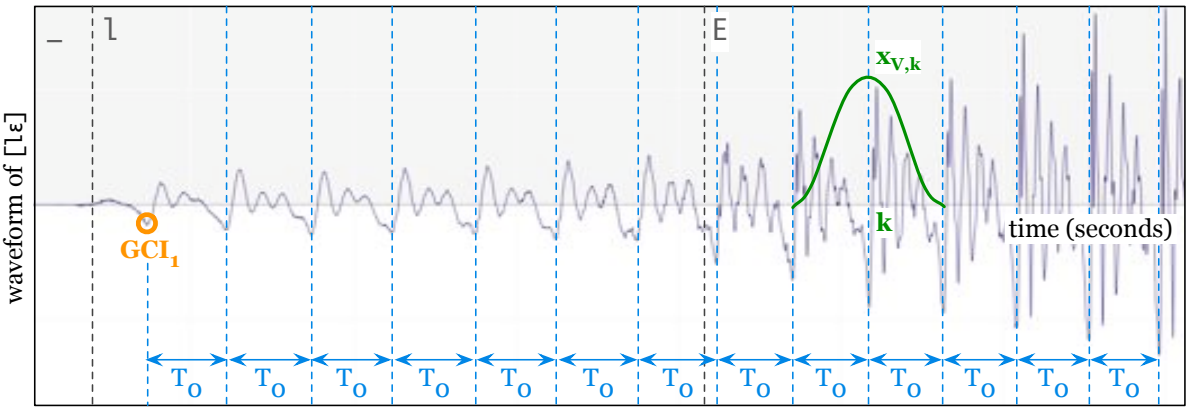

Figure 3.5: Annotated waveform of the syllable [lɛ]. GCI_1 is located slightly after the unvoiced/voiced segmentation point. Other GCI_k locations are extrapolated from locally estimated periods T_0. Then frame $x_{V,k}$ is extracted around GCI_k.

3.2.4 Intra-phoneme segmentation

Inside the phoneme segmentation, we make one more subdivision. Frames $x_{V,k}$ within a vowel are further associated with one of the three following sections: left transient, stable part and right transient. The left and right transients are the regions of the vowel that are coarticulated respectively with the preceding and following vowel, consonant or silence. As the coarticulation is necessary to synthesize an intelligible voice, this segmentation helps the synthesis engine not to alter these regions in further transformations.

Knowing the vowels of the database, this sub-segmentation can be achieved automatically. We use a GCI-synchronous (based on frames $x_{V,k}$) version of the spectral flux $F_S(k)$ (one value for each frame $x_{V,k}$), through equation (3.2).

$$F_S(k) = \sqrt{\sum_{n=0}^{N_d-1} (X_{V,k}(n\frac{\pi}{N_d-1}) - X_{V,k-1}(n\frac{\pi}{N_d-1}))^2} \tag{3.2}$$

where $X_{V,k}(\omega)$ is the Fourier Transform of the frame $x_{V,k}$, discretized on N_d points along the interval $[0, \pi]$. The squared DFT magnitude is used here as an estimator of the PSD.

Computing the value of $F_S(k)$ for each frame $x_{V,k}$ within a vowel, we can observe that coarticulated regions (left and right) correspond to higher values than the central stable part. This convex evolution is due to the quick spectral movements that happen within these coarticulated parts of the vowel, and is illustrated in Figure 3.6.

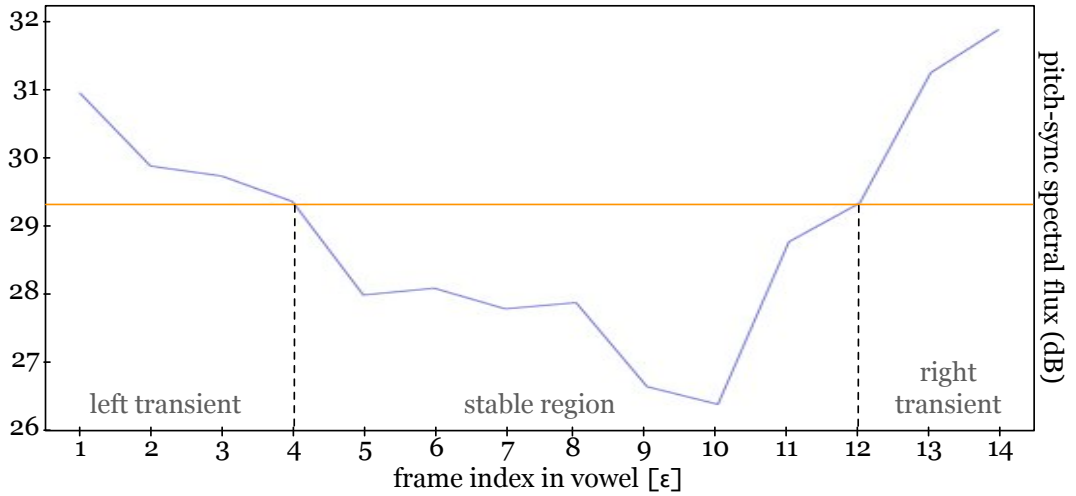

Figure 3.6: Evolution of $F_S(k)$ along the frame index of a vowel [ɛ]. The function decreases, stabilizes and increases. The threshold (orange) defines the three subdivisions.

In the processing of the database, we decide to threshold the values of $F_S(k)$ in each vowel island with a parameter T_f, in order to separate frames into the three groups:

- left transient: $F_S(k) > T_f$, from the beginning;
- stable region: $F_S(k) < T_f$, from the middle;
- right transient: $F_S(k) > T_f$, from the end.

The segmentation obtained with a given T_f can be observed, and T_f can be adjusted in consequence. The value of T_f has to be chosen as a compromise between several aspects: the amount of frames, the max/min values or the mean of $F_S(k)$ in each voice segment. The value of T_f is relative to each voice segment and aims at keeping a reasonable size for transient regions, typically more than 4-5 periods of the waveform.

3.3 Validation of glottal flow analysis on real voice

We have seen in Chapter 2 that voice quality analysis is a research topic that has been addressed in many different ways. Extracting glottal flow parameters from the voice waveform is still a widely open problem.

In this Section, we first discuss the underlying problem of glottal waveform estimation. This problem is the non-accessibility of the sub-glottal pressure (3.3.1). Then we sum-

marize the possible validation techniques that can be applied on ZZT-based algorithms (3.3.2). In the following Sections, these techniques are detailed and indicators of decomposition efficiency are evaluated: separability of ZZT patterns (3.3.3), noisiness of the anticausal component (3.3.4), and model-based validation (3.3.5).

3.3.1 Non-accessibility of the sub-glottal pressure

One of the main underlying problems is the impossible access to the real sub-glottal pressure waveform, from the physiological point of view. We can highlight some intrusive techniques, such as electroglottography (EGG) [99] or videokymography [7] but they merely provide an interpretation – respectively larynx impedance and glottis area – of the GF behavior. Moreover the intrusion limits or biases the phonation process.

Being able to directly observe the sub-glottal pressure would give an absolute reference with which every non-intrusive algorithm – based on digital waveform analysis – could be compared. In this context, existing research proposes two different approaches:

- Analysis methods may be validated with synthetic stimuli. In this case, the "real" GF parameters are set at the input, and some estimation error can be computed. This approach is called *analysis by synthesis* (AbS). In a recent study, Drugman *et al.* proposed two AbS validation factors in the estimation of the glottal formant [67]:

  1 – the *distance* between the magnitude spectrum of the synthetic glottal source and the magnitude spectrum of the estimated glottal source;

  2 – the *determination rate* as the ratio between the amount of frames where F_g has been correctly estimated and the total amount of tested frames.

- When real speech is used for the testing, glottal source analysis algorithms rely on their own validation protocol. They are mainly based on the comparison between the estimated glottal source signal with a model [6], or on measuring noisiness and ripple of the magnitude spectrum of the estimated glottal source signal [12].

3.3.2 Validation techniques used in the improvement of ZZT-based results

As described in 2.4.2, the ZZT-based causal/anticausal decomposition is a recently developed analysis paradigm. The question of *decomposition efficiency* has not been extensively addressed yet. Knowing that absolute validation is not possible – as mentioned in 3.3.1 – we use a more pragmatic approach. We evaluate various indicators of decomposition efficiency (existing and new ones) at different steps of the RAMCESS analysis pipeline, compute their statistics, and discuss correlation between them.

The aim of this study is to provide a common validation strategy for the various approaches that have been used in ZZT-based algorithms. One main interest of this comparison is the use of the same real connected speech database. This choice gives a new feedback on previous and ongoing research with the ZZT.

The first category of indicators concerns ZZT-based algorithms that are already referenced in the literature. Indeed these methods propose some improvement strategies. We explain these improvements and formulate them as quantitative criteria:

- the separability of ZZT patterns, in Section 3.3.3;
- the noisiness of the anticausal component, in Section 3.3.4.

The second category of indicators concerns RAMCESS-specific measurements. The RAMCESS framework is based on frame extraction and two steps of modeling, as described in Section 3.1. We think that the efficiency of ZZT-based decomposition can also be validated at the modeling level, based on the behavior of parameters used in our models. In Section 3.3.5, we formulate quantitative criteria related to:

- the relevance of extracted source parameters;
- the mean modeling error.

3.3.3 Separability of ZZT patterns

Right from the prototyping of the ZZT method, Bozkurt proposes an extensive study of the influence of windowing conditions on the decomposition results. The assumption is made that the wrong representation of the source component essentially comes from truncation problems, as encountered e.g. in asynchronous windowing [25].

The study implicitly proposes the *separability* of zero patterns around the unit circle as the most influent correlate of ZZT efficiency, and more generally of phase processing efficiency. Indeed zeros close to the unit circle provoke a significant phase shift, resulting in spikes in the group delay [26], which lead to noisy decomposition. Consequently, increasing separability leads to a more efficient ZZT-based decomposition.

Bozkurt proposes three guidelines in order to minimize these problems: centering the window on the GCI, a length of $2 \times T_0$ and the use of the Hanning-Poisson window shape. The efficiency of this solution is illustrated in Figure 3.7 [25]. However this study does not explicitly formulate a ZZT separability factor.

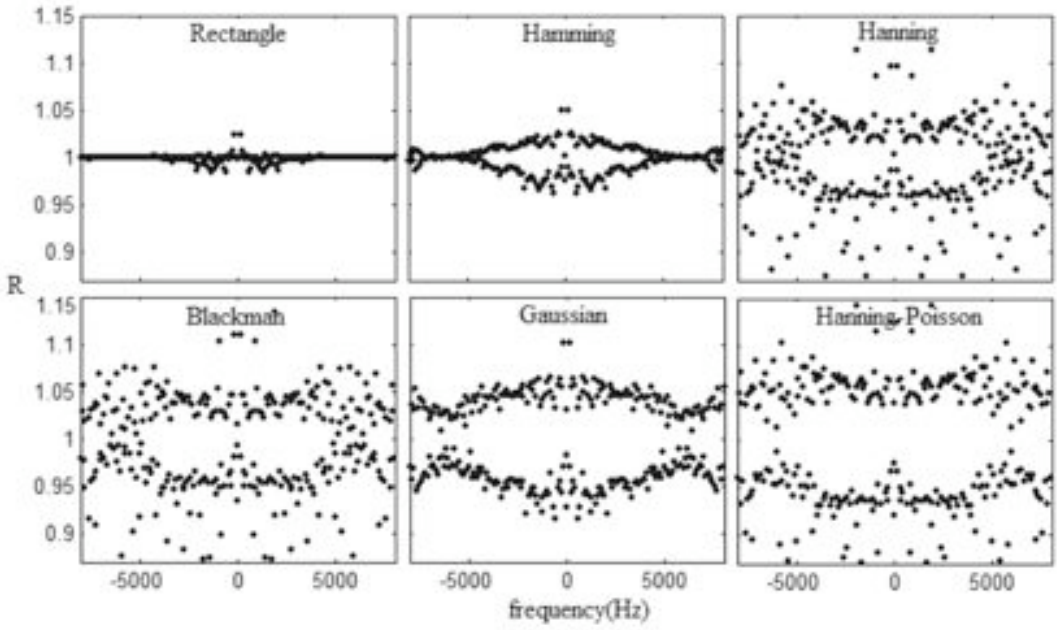

Figure 3.7: Influence of the window type on the separability of ZZT patterns [25].

Within the RAMCESS framework – meaning we work with frames $x_{V,k}$, centered on a given GCI_k – we define a numerical criterion for separability. The separability factor S_k for the k^{th} frame is described in equation (3.3).

$$S_k = \min_{m=[0,N_o]} |Z_o^k(m)| \; - \; \max_{n=[0,N_i]} |Z_i^k(n)| \tag{3.3}$$

where Z_o^k and Z_i^k are the zeros of the Z-transform of $x_{V,k}$ respectively outside and inside the unit circle; N_o and N_i respectively the number of zeros in Z_o^k and Z_i^k.

We propose that S_k has to be maximized in order to improve the ZZT-based decomposition. This factor relies on a recent work, where a jump in the sorted ZZT moduli is discussed and used in order to find a better separation radius than $R = 1$ [66].

Discussing S_k statistics

We can observe the properties of our separability factor S_k over the whole RAMCESS database. In Figure 3.8 and 3.9 we illustrate that Bozkurt's assumptions on GCI centering and window type are verified for a large corpus of real speech. The verification is done by comparing S_k histograms in different conditions.

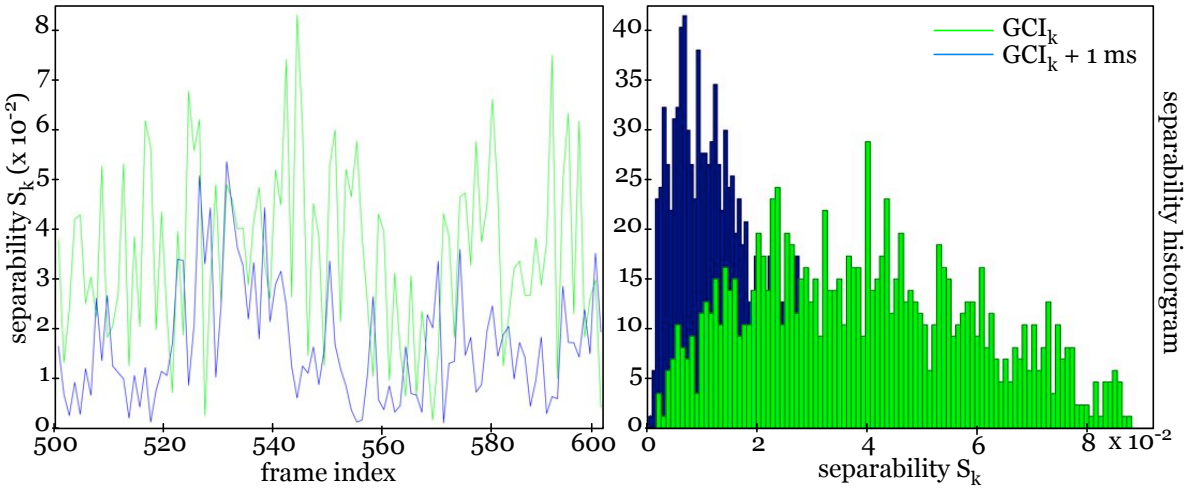

Figure 3.8: Evolution of separability S_k along 100 frames of the database, and corresponding histogram (for the whole database). Comparison between decompositions at GCI_k (green) and $GCI_k + 1$ms (blue) locations.

Histograms in Figure 3.8 show two distributions of S_k. The green distribution (mean = 0.039) corresponds to S_k statistics for frames centered on GCI_k, as determined by the pitch-based extrapolation method (3.2.3). The blue distribution (mean = 0.015) corresponds to S_k statistics for frames centered on $GCI_k + 1$ms.

Comparing those two distributions gives a quantitative impact of the GCI centering, in the context of a large amount of real connected speech. We can observe that the mean of S_k significantly decreases, with the shift of 1ms forward: 0.024 for a range of $[0, 0.09]$.

Another interesting measurement is the *degradation factor*: the amount of frames that encounter a degradation of S_k. This value is 87.4% for a shift of 1ms forward.

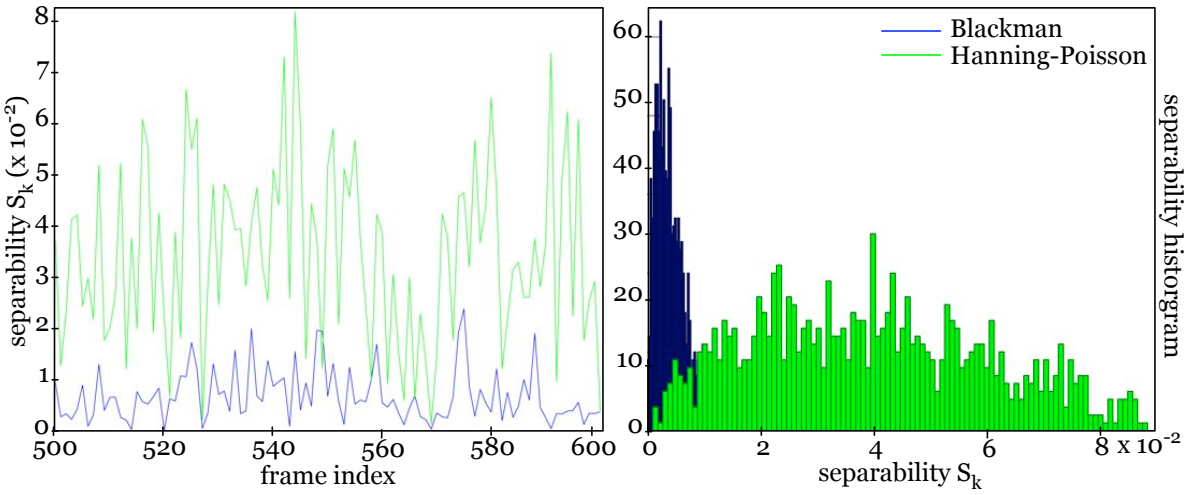

Figure 3.9: Evolution of the separability S_k along 100 frames of the database and the corresponding histogram (for the whole database). Comparison between the decomposition with Blackman (blue) and Hanning-Poisson (green) windowing.

We can also verify the assumptions made by Bozkurt on the window shape. Histograms in Figure 3.9 also show two distributions of S_k, measured along the whole RAMCESS database. The green distribution (mean = 0.039) corresponds to S_k statistics for Blackman-windowed frames. The blue distribution (mean = 0.005) corresponds to S_k statistics for Hanning-Poisson-windowed frames. The *improvement factor*[3] (from Blackman to Hanning-Poisson) is 98.3%. As it could be expected, Hanning-Poisson windowing improves the separability of ZZT patterns.

Studying the statistics of S_k on a large corpus gives the opportunity to evaluate quantitatively the impact of several pre-processing assumptions (such as window types, pitch estimators, GCI tracking algorithms, etc) on further ZZT manipulation.

3.3.4 Noisiness of the anticausal component

Dubuisson introduces the idea of evaluating the efficiency of ZZT decomposition directly on causal and anticausal time-domain waveforms [68], as obtained from equation 2.16. Using the notation introduced in 2.4.2, $x_{C,k}$ and $x_{A,k}$ are respectively the *Causal* and *Anticausal* time-domain signals extracted from the frame $x_{V,k}$ (centered on GCI_k).

[3] We can also evaluate the opposite of the degradation factor: the amount of frames that encounter an improvement due to a given manipulation. In this case we evaluate the *improvement factor*.

Choosing the appropriate type and length for windowing is obvious. The open problem underlying ZZT analysis is the centering of each window on the GCI. Dubuisson's approach is to consider that the best GCI_k location (in a given region) corresponds to the one which minimizes the noisiness of the anticausal component $x_{A,k}$.

The improvement of GCI_k locations is obtained by combining two mechanisms:

- Systematic shifts are realized around each GCI_k, estimated by the pitch-based extrapolation, as described in 3.2.3. If the maximum shift range is set to 4 samples, 9 $x_{A,k}$ candidates (thus 9 ZZT) are computed around each GCI_k.
- The noisiness of each $x_{A,k}$ candidate is evaluated. This measurement is made in the spectral domain. Indeed by comparing the magnitude spectrum of a correct $x_{A,k}$ and a noisy $x_{A,k}$, we can observe that their behaviour is quite similar below 2kHz, but significantly different in higher frequencies, as in Figure 3.10.

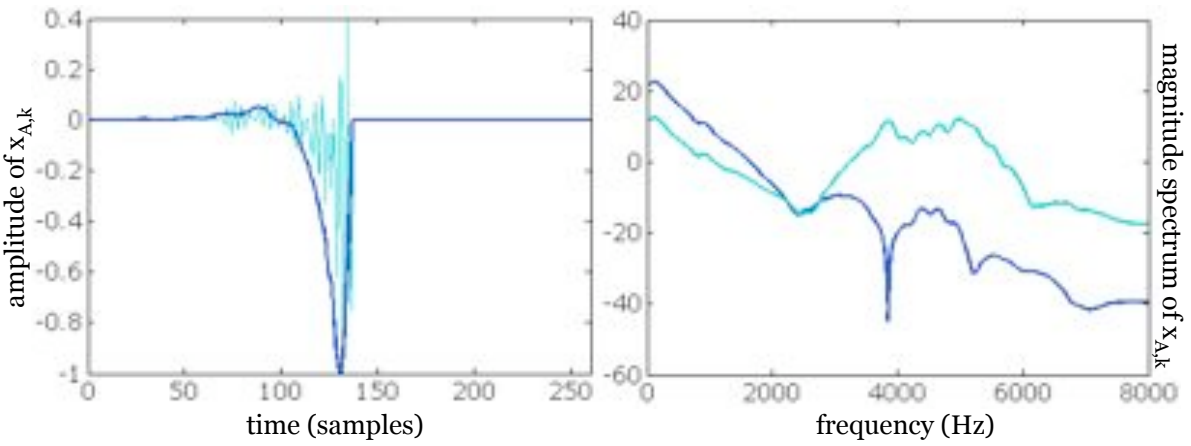

Figure 3.10: Correct $x_{A,k}$ (dark) vs. noisy $x_{A,k}$ (light): the time-domain noisiness is due to the increasing of high frequencies when the ZZT decomposition fails.

In order to choose the best $x_{A,k}$ among all candidates (for a given k), the *smoothness* D_k is defined as the ratio between the energy in the $[0-2\text{kHz}]$ frequency band and the energy in the whole spectrum $[0, F_s/2]$, as in equation (3.4).

$$D_k = \frac{1}{\alpha} \times \frac{\sum_{m=0}^{\alpha N_d - 1} |X_{A,k}(m \frac{\pi}{N_d - 1})|}{\sum_{n=0}^{N_d - 1} |X_{A,k}(n \frac{\pi}{N_d - 1})|} \tag{3.4}$$

where $X_{A,k}(\omega)$ is the Fourier Transform of the frame $x_{A,k}$; $\alpha = \frac{2000}{F_s/2}$ is the ratio between the two frequency bands. Thus αN_d and N_d are respectively the number of frequency bins corresponding to $[0, 2\text{kHz}]$ and $[0, F_s/2]$ frequency bands.

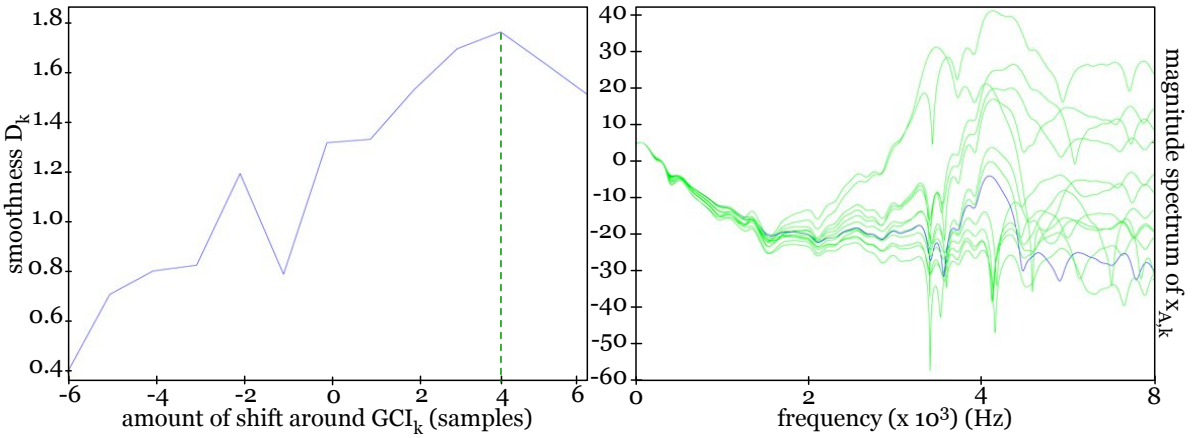

Figure 3.11: Left: Computation of D_k for 13 shifts around GCI_k: $GCI_k + [-6, 6]$ samples. The maximum of D_k is in $GCI_k + 4$ samples. Right: $GCI_k + 4$ samples gives the $|X_{A,k}|$ spectrum with a minimum of high-frequency noise (blue).

In Figure 3.11 (left part) we observe that a forward shift of 4 samples – among a total searching zone of $[-6, 6]$[4] – gives the maximum value of the smoothness D_k. We verify (right part) that this shift – corresponding to the blue curve – provides the $|X_{A,k}|$ spectrum with the minimum amount of high-frequency noise.

Discussing D_k statistics and impact on S_k

As done for S_k in the previous section, we now evaluate the statistics of D_k, so as to check if the systematic maximization of D_k over all attempted shifts improve the separability of ZZT patterns. Moreover, as we shall see, observing D_k statistics gives us a criterion for excluding some unanalysable frames.

As shown in Figure 3.12 (panel A), there is an improvement of D_k over the whole database, due to the shifting algorithm within the interval $[-6, 6]$ samples. Indeed the mean of the D_k distribution of goes from 1.60 to 2.15, with quite an equivalent variance.

Another interesting aspect of this distribution is that we can set a threshold in order to decide to reject some frames, supposed too noisy from a macroscopic point of view. For

[4] The samplerate of the database used in [68] is also 16kHz, a common value for speech processing. A 13-sample searching zone corresponds to less than 1 millisecond.

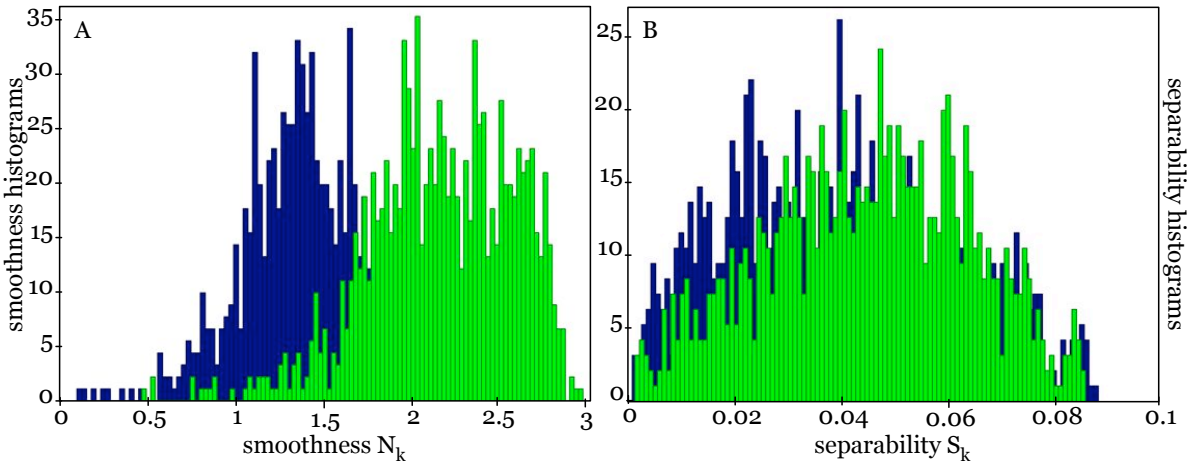

Figure 3.12: A: histograms of D_k without (blue) and with (green) the optimization by shifting frames around GCI_k. B: histograms of S_k without (blue) and with (green) the optimization by shifting frames around GCI_k.

example, the threshold $D_k > 1.5$ (after shifting) rejects 6% of the frames included in the database. This gives some information on "undecomposable" regions.

We also study the relation between D_k and S_k. It is interesting to highlight that the optimization of GCI_k locations by D_k maximization does not significantly improve the separability of ZZT patterns for the whole database. Figure 3.12B shows S_k distributions before (blue) and after (green) D_k optimization as mainly overlapped. The difference of the means is small (5×10^{-3}) and the improvement factor is only 56%.

Consequently, improvements of ZZT decomposition from S_k and D_k points of view correspond to two different approaches, with no significant correlation. It highlights that there is not one best way of optimizing ZZT-based decomposition algorithms. Until now, the understanding of correlation between the analysis protocol and the resulting ZZT pattern is still at an early step. We expect that further investigation on this topic would lead to a more coherent set of optimization rules.

3.3.5 Model-based validation criteria

As explained in Section 3.1, RAMCESS analysis is achieved in two main steps: ZZT decomposition, fitting of decomposition results with two models: a LF-based model of the glottal formant (cf. Section 3.4) and ARX modeling of the frame (cf. Section 3.5).

Validation strategies based on ZZT pattern separability (based on S_k) and smoothness (based on D_k) propose a spectral-based approach for the efficiency of the ZZT decomposition. In this Section, we take the advantage of working within a full modeling pipeline, in order to propose model-based validation criteria:

- the relevance of extracted source parameters;
- the mean modeling error.

Applying a model on a raw result consists in finding the set of parameters which verify the best a given criterion. In our analysis we aim at finding relevant model parameters $P_k : \{p_1, p_2, ..., p_N\}_k$ for each GCI-centered frame $x_{V,k}$. Then the behavior of each parameter $p_{i,k}$ is observed and commented for the whole database.

First we define a subset of modeling parameters that are specific to the modeling of the glottal source, and we denote them as $g_{i,k}$ $(i = 1...G)$. By extension, we can thus refine the description of P_k by assuming that it is made of two subsets: one for the source parameters $g_{i,k}$ and another for the vocal tract parameters $t_{i,k}$, containing respectively G and T parameters:

$$P_k : \{g_1, g_2, ..., g_G, t_1, t_2, ..., t_T\}_k$$

Stability of $g_{i,k}$

Due to the special recording conditions that have been described in 3.2.1, we expect $g_{i,k}$ to exhibit some specific statistics. In particular, source parameters (e.g. f_0, O_q, α_M, T_L) should be significantly stable (as we were expecting to stabilize the voice quality).

For a given voiced island v we evaluate $m_{i,v}$ and $\sigma^2_{i,v}$, respectively the mean and the variance of the parameter $g_{i,k}$. These indicators are described in equations 3.5 and 3.6.

$$m_{i,v} = \frac{1}{N_v} \sum_{k=0}^{N_v-1} g_{i,k+b_v} \tag{3.5}$$

$$\sigma^2_{i,v} = \frac{1}{N_v} \sum_{k=0}^{N_v-1} (g_{i,k+b_v} - m_{i,n})^2 \tag{3.6}$$

where N_v and b_v are respectively the number of frames and the index of the first frame (considering that each frame has an unique index) in the v^{th} voiced island.

Thus the flatness of the fundamental frequency presented in 3.2.1 can now be studied through these two parameters. For example, the values for a [ɛ] in the beginning of the database ($v = 3$) and considering f_0 is the one parameter of the model ($i = f_0$):

$$m_{f_0,3} = 136.18 \text{ (in Hz)}$$
$$\sigma^2_{f_0,3} = 26.79 \text{ (in Hz)}$$

At a more macroscopic level, we want to verify three different aspects:

- Obviously we expect the overall mean of a given parameter over the whole database to center around expected values. Thus we compute M_i, the mean of all the $m_{i,v}$ values, in equation (3.7), and the results will be compared with usual values encountered in normal male voices (as our database is made of male speakers).
- $m_{i,v}$ should not to jump from one value to a totally different one in successive voiced islands. Indeed it would mean that the computation of source parameters $g_{i,k}$ is influenced by the phonetic context, which is the main disturbing aspect of voice quality analysis (formant perturbation). This property is evaluated by computing F_i, the mean of $m_{i,v}$ fluctuations, in equation (3.8).
- We also want to verify that the average variance V_i^2 remains low for the whole database, as described in equation (3.9).

$$M_i = \frac{1}{N_t} \sum_{v=0}^{Nt\_1} m_{i,v} \tag{3.7}$$

$$F_i = \frac{1}{N_t} \sum_{v=1}^{N_t-1} |m_{i,v} - m_{i,v-1}| \tag{3.8}$$

$$V_i^2 = \frac{1}{N_t} \sum_{v=0}^{Nt\_1} \sigma^2_{i,v} \tag{3.9}$$

where N_t is the total number of voiced islands in the database.

Mean modeling error

Once the parameters of one model have been estimated, the k^{th} original $x_{V,k}$ and resynthesized $x_{R,k}$ signals can be compared. At this level, a modeling error e_k can be evaluated, as described in equation (3.10)[5]. The mean modeling error E for the whole database, is computed by the equation (3.11), and is one aspect of the performance of the analysis.

$$e_k = \sqrt{\frac{1}{N_s}\sum_{i=0}^{N_s-1}\left(x_{R,k}(i) - x_{V,k}(i)\right)^2} \tag{3.10}$$

$$E = \frac{1}{N_f}\sum_{k=0}^{N_f-1} e_k \tag{3.11}$$

where N_s is the number of samples in the frame $x_{V,k}$ or $x_{R,k}$, and N_f is the total number of frames in the whole database.

3.4 Estimation of the glottal formant

The main issue of this Section is the fitting of a glottal formant model on the raw anticausal component coming from the ZZT decomposition algorithm. As explained in Section 2.2.3, the glottal formant can be parametrized by its spectral attributes (F_g, A_g). Accordingly to the mixed-phase model, the glottal formant is due to the anticausal part of the GFD [65]. Thus it can also be represented by time-domain parameters, such as (O_q, α_M). In this work, we associate a (O_q, α_M) coordinate to each frame $x_{A,k}$.

This attribution is performed in three steps:

- shifting the analysis frame around GCI_k to find the best decomposition (3.4.1);
- evaluating the frequency of the glottal formant F_g (3.4.2);
- minimizing the error between $x_{A,k}$ and the model-based candidate (3.4.3).

[5] As we essentially work on phase estimation, we prefer using a time-domain measurement of the error, instead of a spectral magnitude measurement.

3.4.1 Shifting the analysis frame around GCI_k

In [68] the idea of shifting frame $x_{V,k}$, around the GCI_k (i.e., the value proposed by the GCI tracking technique given in Section 3.2.3) is introduced as a way of decreasing the noisiness of the anticausal component. In Section 3.3.4 this was formulated by maximizing the smoothness criterion D_k, and the interest of the approach was confirmed.

Considering that frame $x_{V,k}$ is selected in order to maximize D_k, another problem can degrade the resulting anticausal frame $x_{A,k}$ and make the glottal formant less observable. This problem is due to the nature of ZZT patterns and their implication in the shape of the anticausal magnitude spectrum $|X_{A,k}(\omega)|$.

Occurrence of the glottal formant in ZZT patterns

According to ZZT-based causal/anticausal decomposition (2.4.2), the anticausal magnitude spectrum $|X_{A,k}(\omega)|$ is computed with equation (2.16), using zeros outside the unit circle ($R > 1$). However we know that the glottal formant is observable on $|X_{A,k}(\omega)|$ only if the resulting ZZT pattern exhibits a "gap" in the low frequencies [27].

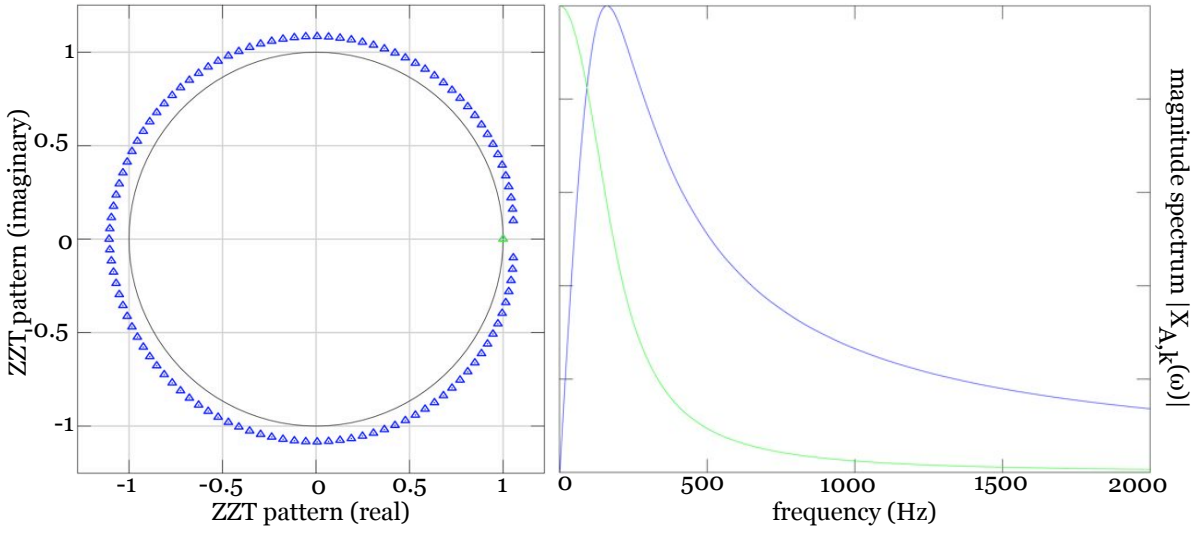

Figure 3.13: Influence of the presence/absence of zero in $(1, 0)$. When all the zeros are present (left: blue + green triangles), the magnitude spectrum $|X_{A,k}(\omega)|$ has a formantic shape (right: blue curve). When $(1, 0)$ is removed (left: blue triangles only), $|X_{A,k}(\omega)|$ has a decreasing shape (right: green).

In Figure 3.13, we observe the ZZT pattern of a typical LF glottal pulse. We can see that the regular ZZT pattern outside the unit circle (left: green and blue triangles)

leads to a formantic shape for the anticausal magnitude spectrum $|X_{A,k}(\omega)|$ (right: blue curve). This resonnance results from the gap which is observable between the zero in $(1, 0)$ (left: green triangle) and other zeros in high frequencies (left: blue triangles). If we intentionally remove the zero in $(1, 0)$, the gap now occurs in 0 Hz. This altered ZZT pattern leads to a constantly decreasing magnitude spectrum (right: green curve).

This simple example based on the LF model reveals that, in the context of real speech, the presence of a resonance in the magnitude spectrum of the anticausal component $|X_{A,k}(\omega)|$ relies on how this single zero in $(1, 0)$ is classified. If this zero in $(1, 0)$ is considered as part of the anticausal ZZT pattern, the resulting $|X_{A,k}(\omega)|$ exhibits a resonance, otherwise the anticausal ZZT pattern leads to a non-relevant shape.

Bozkurt has shown that, as a result of windowing and truncation of the speech signal, the zero expected in $(1, 0)$ can sometimes be retrieved inside the unit circle [25]. Some techniques have been implemented in order to look for this zero within a larger area than strictly $R > 1$, for example by also searching around the axis $(R_s, 0)$ with $0.9 < R_s < 1$.

However, informal observation of $|X_{A,k}(\omega)|$ over the RAMCESS database reveals that a significant amount of frames – particularly in consonants and coarticulated parts of vowels – remain "non-formantic". This problem in the shape of the magnitude spectrum $|X_{A,k}(\omega)|$ is encountered, despite the extended searching of the zero that is expected in $(1, 0)$, and despite the optimization strategy based on D_k (cf. 3.3.4).

Formanticity criterion

In this work, we extend the idea of maximizing D_k (smoothness criterion) with another aspect. We introduce the *formanticity criterion* F_k that aims at measuring the formantic shape of a given $|X_{A,k}(\omega)|$, among several candidates. The criterion F_k is a combination of two measurements, F_k^1 and F_k^2. It is defined in equations (3.12) to (3.14).

$$F_k^1 = \max_{\omega=[\omega_L,\omega_H]} |X_{A,k}(\omega)| - |X_{A,k}(\omega_L)| \tag{3.12}$$

$$F_k^2 = \max_{\omega=[\omega_L,\omega_H]} |X_{A,k}(\omega)| - |X_{A,k}(\omega_H)| \tag{3.13}$$

$$F_k = F_k^1 \times F_k^2 \tag{3.14}$$

where ω_L and ω_H are respectively the low and high frequency boundaries where the F_k magnitude differentiation is evaluated. These boundaries are tuned in order to define the frequency range $F_g \in [10, 500]$ (in Hz), where we expect to find the glottal formant. This frequency range is large enough to detect any acceptable value of F_g ($\pm$ 200 Hz for male voices), but small enough not to be perturbed by higher frequency noise.

The measure of formanticity F_k can even be combined with the evaluation of the smoothness D_k already used in [68]. Actually expecting a formant-shape magnitude spectrum $|X_{A,k}(\omega)|$ corresponds to the maximization of F_k for the whole shift range s. As Dubuisson has shown that the best shift rarely exceeds the value 8, we work with $s = [-8, 8]$. Both criteria have to be maximized. We can thus consider to maximize C_k:

$$C_k = \frac{D_k^N + F_k^N}{2} \tag{3.15}$$

where D_k^N and F_k^N are respectively the normalized versions of D_k and F_k among the whole shift range $s = [-8, 8]$. D_k and F_k are normalized in amplitude, between 0 and 1. Figure 3.14 shows values of D_k, F_k and C_k for the shift range $s = [-8, 8]$, and for two voiced sounds: [ɛ] and [l].

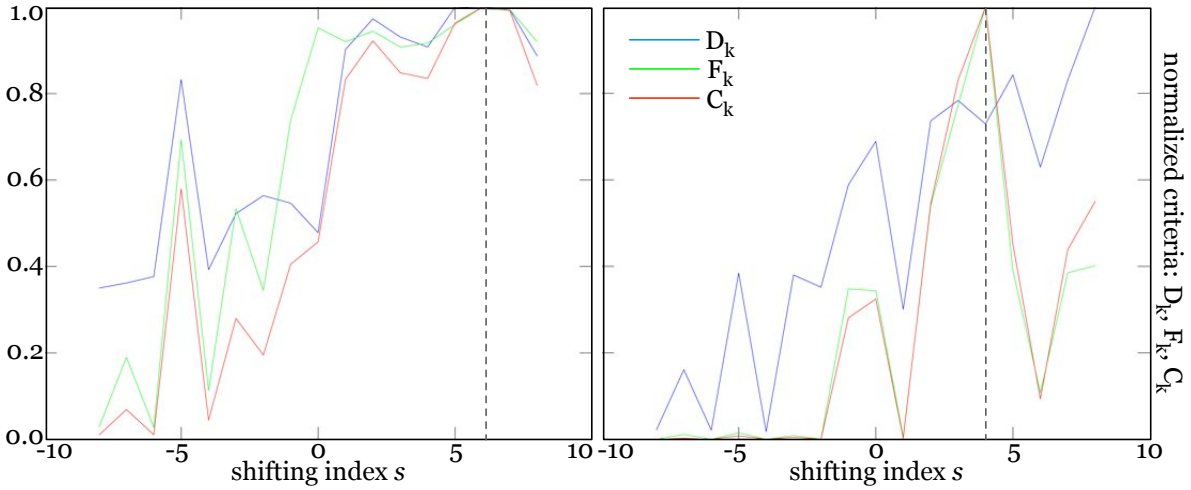

Figure 3.14: Evolution of normalized D_k (blue), F_k (green) and C_k (red) criteria among different $|X_{A,k}(\omega)|$ candidates, for the shift range $s = [-8, 8]$, and for two voiced sounds: [ɛ] (left) and [l] (right).

It appears that D_k (blue) and F_k (green) are maximized around the same value of shift (around $s = 5$). However the combination of both – through the value of C_k (red) – provides a clearer peak in problematic sounds, such as [l] (right).

3.4.2 Evaluation of glottal formant frequency

The maximization of C_k allows us to pick the glottal formant frequency on particularly suitable magnitude spectra. As the glottal formant is closely correlated to some aspects of the voice quality perception [49], it is really important to rely further developments on a good estimation of F_g. Due to our recording condition, we expect F_g to be flat.

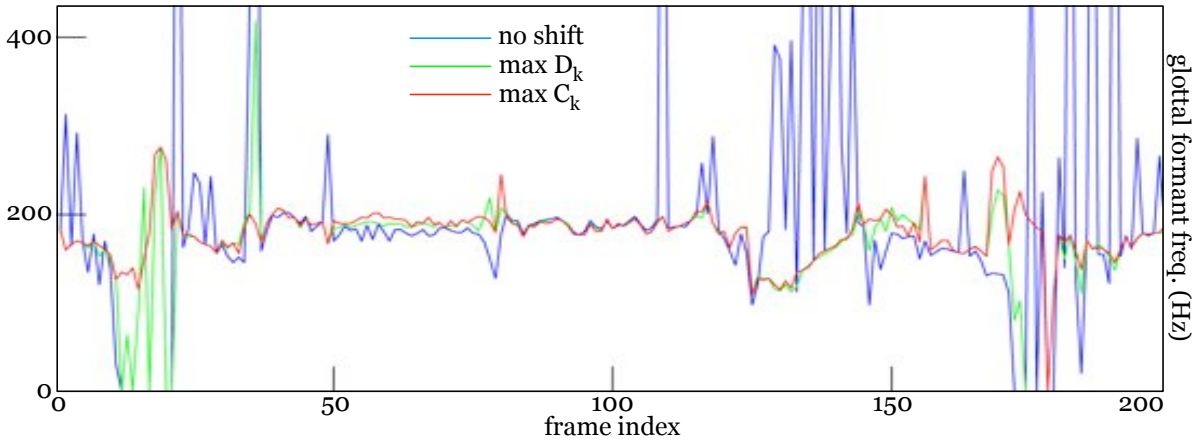

Figure 3.15: Comparing three F_g tracking methods on several frames of the sequence [lɛtmi]: no shift (blue), maximization of D_k (green) and maximization of C_k (red).

In Figure 3.15 we see trackings of F_g for several frames of the sequence [lɛtmi][6], a particularly difficult situation for ZZT analysis. Indeed this sequence is made of coarticulated regions and consonants. We compare the value of F_g picked from the magnitude spectrum $|X_{A,k}(\omega)|$ in three different situations:

- without any shifting strategy;
- with the maximization of D_k alone;
- with the maximization of C_k.

We can see that there is a significant improvement in using the shifting strategy, as explained in [68]. Moreover we observe that that the combined criterion C_k can be more reliable in difficult cases, such as in transitions between vowels and voiced consonants.

[6] As [t] is an unvoiced sound, it is not part of the plotting of F_g in Figure 3.15.

3.4.3 Fitting of the LF model

Once the best possible $x_{A,k}$ have been obtained for each k, glottal flow parameters can be estimated from these frames, such as the open quotient (O_q) and the asymmetry coefficient (α_M). These parameters result from a particular fitting between $x_{A,k}$ and the anticausal part of the Liljencrants-Fant (LF) GFD model [78]. This synthetic frame – denoted x_G – is aligned with $x_{A,k}$ and also multiplied by a Hanning-Poisson window.

The fitting strategy that we use is a spectral-domain error minimization, the error being computed both on the magnitude spectrum and on the group delay[7] of the frames. This *spectral distance* between frames $x_{A,k}$ and x_G is presented in equations 3.16 to 3.18.

$$E_{m,k} = \sqrt{\frac{1}{\pi}\int_0^{\pi} (|X_G(\omega)| - |X_{A,k}(\omega)|)^2 \, d\omega} \tag{3.16}$$

$$E_{p,k} = \sqrt{\frac{1}{\pi}\int_0^{\pi} (\frac{\partial \Phi_G(\omega)}{\partial \omega} - \frac{\partial \Phi_{A,k}(\omega)}{\partial \omega})^2 \, d\omega} \tag{3.17}$$

$$E_k = E_{m,k} + E_{p,k} \tag{3.18}$$

This error is computed in an iterative process:

- The local period of frame $x_{A,k}$ from the pitch estimation, previously achieved in 3.2.3. We denote it T_0^k. The amplitude of the negative peak of $x_{A,k}$ is also measured. Figure 3.10 (left panel, dark) shows that $x_{A,k}$ exhibits a clear negative peak. The amplitude of this peak, denoted E^k, is obtained by locating the minimum of $x_{A,k}$.
- Ranges of variation and resolutions of O_q and α_M are set. For example, O_q can vary in $[0.3 - 0.9]$ and α_M in $[0.6 - 0.9]$, both by steps of 0.05; These values are represented by $\vec{O_q}$ and $\vec{\alpha_M}$ vectors in the sequel. $\vec{O_q}(m)$ $(m = 1...M)$ corresponds to the m^{th} value of $\vec{O_q}$ over M; $\vec{\alpha_M}(n)$ $(n = 1...N)$ to the n^{th} value of $\vec{\alpha_M}$ over N.
- A codebook $\Theta_{F,k}$, containing a matrix of LF-based GFD periods (anticausal component only) x_G, is computed with the same period and amplitude as frame $x_{A,k}$

[7] We would like to highlight that the notation for the phase of $X_{A,k}(\omega)$ is $\Phi_{A,k}(\omega) = \arg\{X_{A,k}(\omega)\}$. Moreover, the phase is unwrapped before being used for the computation of the distance.

and for all the values of $\vec{O_q}$ and $\vec{\alpha_M}$. We denote $x_G^{m,n}$, the frame computed with parameters T_0^k, E^k, $\vec{O_q}(m)$ and $\vec{\alpha_M}(n)$. This frame is stored in $\Theta_{F,k}(m,n)$.

$$\Theta_{F,k}(m,n) \Rightarrow x_G^{m,n} = f(T_0^k, E^k, \vec{O_q}(m), \vec{\alpha_M}(n)) \tag{3.19}$$

- The glottal formant frequency of each frame $x_G^{m,n}$ is computed. This glottal formant frequency, that we denote $F_g^{m,n}$, is obtained with the estimation technique described in 3.4.2. For each entry of the codebook $\Theta_{F,k}(m,n)$, $F_g^{m,n}$ is compared with the glottal formant frequency of frame $x_{A,k}$, that we denote F_g^k. If the distance between $F_g^{m,n}$ and F_g^k is greater than a given ΔF_g, the entry $\Theta_{F,k}(m,n)$ is removed from the codebook $\Theta_{F,k}$. This process finally results in a reduced codebook, denoted $\Theta_{R,k}$.

$$\Theta_{R,k} = \Theta_{F,k} \mid_{F_g^{m,n} \in [F_g^k \pm \Delta F_g]} \tag{3.20}$$

- The spectral distance E_k between each instance of the reduced codebook $\Theta_{R,k}$ and $x_{A,k}$ is computed, resulting in a matrix of error values $\vec{E_k}$. $\vec{E_k}(m,n)$ corresponds to the spectral distance between $x_G^{m,n}$ and $x_{A,k}$. As the codebook has been reduced, not every value of m and n leads to a value of E_k. $\vec{E_k}$ contains some gaps.
- The smallest value in $\vec{E_k}$ indicates which entry of $\Theta_{R,k}$ best fits $x_{A,k}$ (in the sense of the spectral distance), and provides values for O_q and α_M. We denote a and b, the indexes in $\vec{O_q}$ and $\vec{\alpha_M}$ respectively, which lead to the minimal value in $\vec{E_k}$.

$$\underset{m \in [1,M], n \in [1,N]}{\operatorname{argmin}} \vec{E_k}(m,n) = \{a,b\} \tag{3.21}$$

$$O_q = \vec{O_q}(a); \ \alpha_M = \vec{\alpha_M}(b) \tag{3.22}$$

- Finally the fitting algorithm provides the *fitted*, that we denote $x_{F,k}$. This frame is the LF modeling of frame $x_{A,k}$ is stored in $\Theta_{R,k}(a,b)$.

$$x_{F,k} = \Theta_{R,k}(a, b) \tag{3.23}$$

In Figure 3.16 we present results of the fitting algorithm on several $x_{A,k}$ periods. It shows that estimated O_q and α_M lead to anticausal components of the GFD (green) which particularly fit the frames $x_{A,k}$, obtained from ZZT-based decomposition (blue).

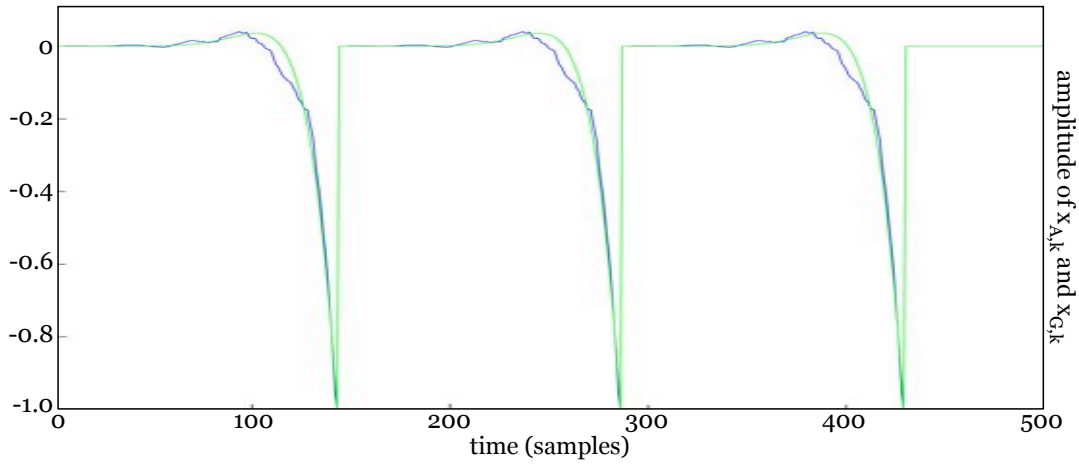

Figure 3.16: Result of the fitting between the anticausal component coming from ZZT-based decomposition $x_{A,k}$ (blue) and the fitted LF-based GFD $x_{F,k}$ (green).

Coherence and stability of estimated glottal source parameters can be verified by observing O_q and α_M statitics, as illustrated in Figure 3.17. O_q (left) and α_M (right) distributions have means of respectively $O_q = 0.54$ and $\alpha_M = 0.84$, and both a rather limited variance. Considering the values that we can expect for these parameters (normal male voice) [96] – $O_q \in [0.5, 0.7]$ and $\alpha_M \in [0.7, 0.9]$ – our estimations seem relevant.

3.5 Joint estimation of source/filter parameters

In order to determine causal components (i.e., the vocal tract parameters and the return phase of the glottal source, through the spectral tilt value T_L) of each frame $x_{V,k}$ we use a modified version of the ARX-LF method, as described in 2.4.3. The modified ARX algorithm is based on two ideas: reducing the size of the ARX-LF codebook [193], thanks to previous ZZT-based (O_q, α_M) results (cf. Section 3.5.1), and re-shifting around GCI_k in order to find the best solution (cf. Section 3.5.2).

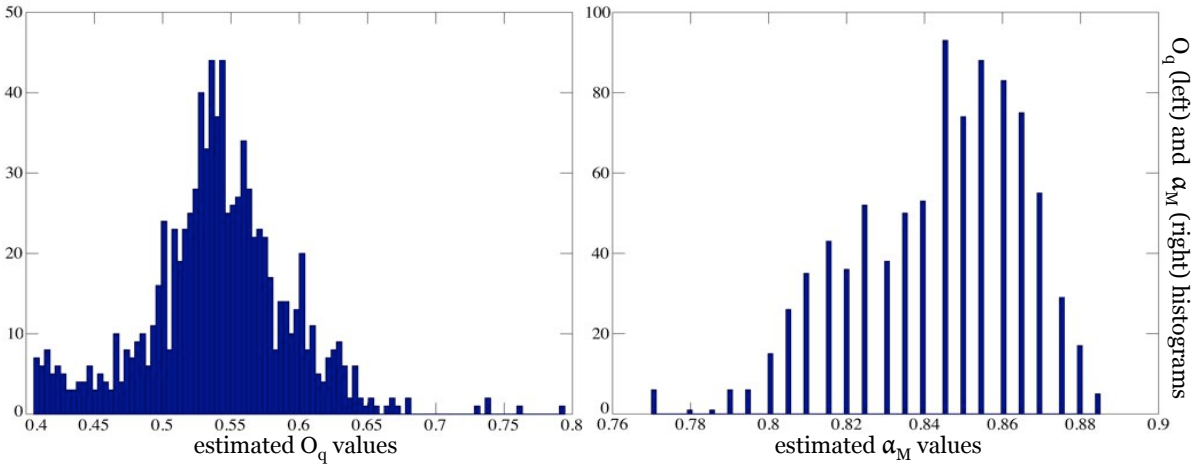

Figure 3.17: Histograms of estimated O_q (left) and α_M (right) resulting from fitting of LF model on ZZT-based anticausal frames $x_{A,k}$.

3.5.1 Error estimation on a sub-codebook

A complete codebook Θ_J of GFD periods, based on the possible variations of their parameters (O_q, α_M and T_L) would be rather bulky and solving (2.18) for all the entries of that codebook would be computationally expensive. Moreover it has been highlighted that ARX-LF could sometimes converge to very unprobable consecutive values [50].

Fortunately O_q and α_M estimations are already known for each GCI_k, thanks to ZZT analysis and LF-based fitting, as applied in Section 3.4. This allows us to reduce the codebook Θ_J to a frame-dependant sub-codebook, which we will denote as $\Theta_{S,k}$.

The basic way of operating consists in taking T_L as the only varying parameter of that sub-codebook $\Theta_{S,k}$. However, although we are confident in the estimate of O_q and α_M, we could refine these results by selecting a somehow larger sub-codebook, allowing slight variations of O_q and α_M around their initial estimations.

Let us assume that, for each GCI_k, the corresponding sub-codebook $\Theta_{S,k}$ contains a number of W glottal flows. We compute the LP coefficients – $a_k(i)$ and b_k – for every entry in $\Theta_{S,k}$ and we resynthesize an approximation $x^w_{R,k}$ of the frame of speech $x_{V,k}$ by applying equation 2.17. At GCI_k, the error for the w^{th} frame $x^w_{R,k}$ is then measured as its Euclidean distance to the original frame $x_{V,k}$:

$$E_w = \sqrt{\sum_{n=1}^{N_t} (x_{V,k}(n) - x^w_{R,k}(n)\ w(n))^2} \tag{3.24}$$

where $w(n)$ is a Hanning window, and N_t is the number of samples in frames $x_{V,k}$ and $x^w_{R,k}$ i.e., in two periods $(2 \times T_0)$.

3.5.2 Error-based re-shifting

Before actually computing error values, two important points remain: the position of GCI_k and the stabilization of the AR filter. Indeed, the estimate of each GCI position is provided by ZZT analysis. Although that position fits very well for ZZT decomposition, it's not necessarily the best one for ARX optimization. For that reason one more step is added to the algorithm explained above: just like during ZZT analysis we do not consider only the analysis window $x_{V,k}$ centered on GCI_k but also windows centered a few points on the left and on the right of that location.

In our implementation we move the frame three samples before and after the position of GCI_k. Henceforth we have $7 \times W$ $x^w_{R,k}$ and their corresponding error measurements. Then, the minimum error gives us $x_{S,k}$ (the best guess for the glottal flow, with O_q, α_M and T_L as its parameters), as well as the optimal position of GCI_k.

Finally, although LP analysis guarantees that the AR filter has all of its poles inside the unit circle, this is no longer the case when solving equation 2.18. Consequently, the last step before synthesizing any of the $x^w_{R,k}$ is to mirror the outside poles of a_k inside the unit circle and to adapt the value of parameter b_k.

3.5.3 Frame-by-frame resynthesis

The last step of the RAMCESS analysis process is the resynthesis each frame of the database, by generating the GFD period with the best O_q, α_M and T_L candidates, and then filtering this source with the all-pole filter $H(z) = \frac{b_k}{A_k(z)}$. This resynthesis process leads to frame $x_{R,k}$, which is the best model-based estimation of frame $x_{V,k}$. The whole RAMCESS database is resynthesized using this technique.

Figure 3.18 shows an example of the difference between the original signal from the database (blue) and the resynthesis from source and filter parameters (green). $|X_{V,k}|$ and $|X_{R,k}|$ are illustrated in Figure 3.19. We can see that the fitting is efficient in low frequencies but has a great difficulty in modeling high frequency components. This problem leads to quite a big difference between the two waveforms in the time domain, and the resynthesized signal sounds artificial and low-pass filtered.

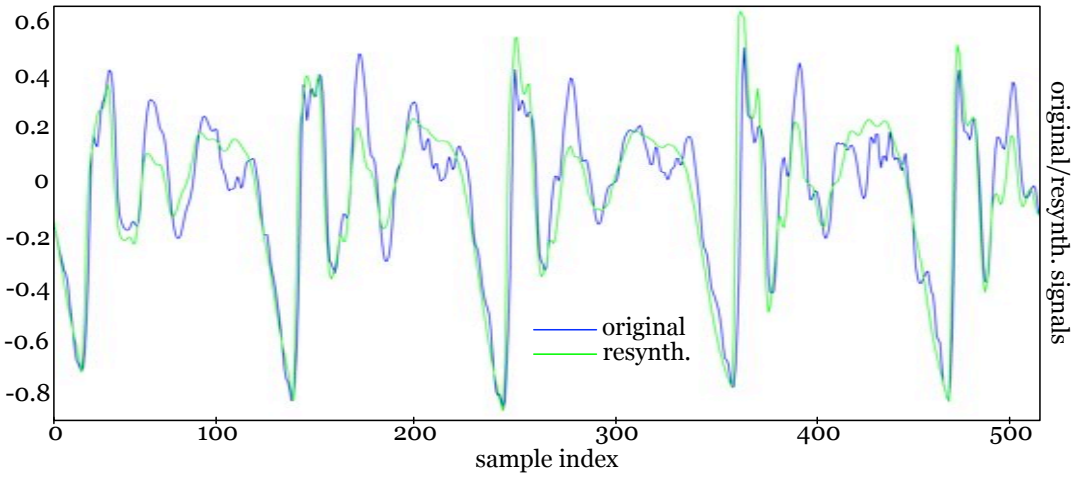

Figure 3.18: Superposition of original (blue) and resynthesized (green) signals, after the computation of ARX-LF on a sub-codebook dened by ZZT-based parameters.

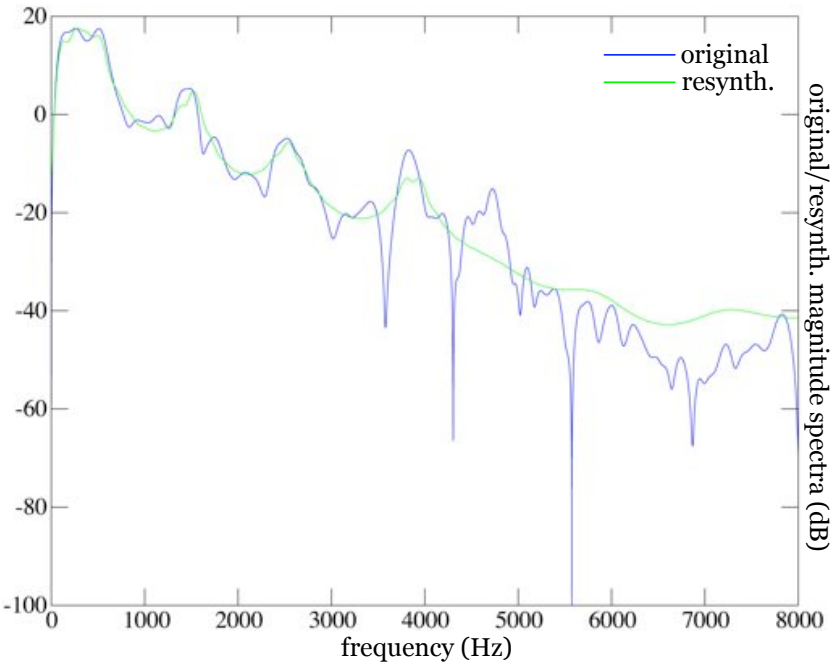

Figure 3.19: Original (blue) and resynthesized (green) magnitude spectra, after the computation of ARX-LF on a sub-codebook dened by ZZT-based parameters.

The problem of high frequency modeling has already been addressed in the existing ARX-LF algorithm, and several solutions have been proposed in order to compensate this difference [194]. Our system does not integrate these refinements yet.

3.6 Evaluation of the analysis process

The evaluation of the analysis framework exposed in Sections 3.4 to 3.5 is done in comparison with ARX-LF alone. We process a classical ARX-LF estimation of the source parameters on the RAMCESS database. Then we compare M_i, F_i, V_i^2 values and the mean modeling error, criteria that have been studied in Section 3.3.5.

3.6.1 Relevance and stability of source parameters

As mentioned in Section 3.3.5, i refers to the parameter on which the statistics are computed[8]. Thus we now consider $i = \{O_q,\ \alpha_M,\ T_L\}$ for the following discussion.

O_q statistics

Table 3.1 shows that the three quality indicators for the open quotient O_q are better within the RAMCESS analysis framework than in ARX-LF. We can see that the mean M_{O_q} with RAMCESS stands around 0.55, which is an expected value for a normal male voice ($O_q \in [0.5, 0.7]$), as the value with ARX-LF is significantly higher. The variance $V_{O_q}^2$ and the fluctuation rate F_{O_q} are clearly lower than for ARX-LF.

method	M_{O_q}	F_{O_q}	$V_{O_q}^2$
ARX-LF	0.90241	0.071698	0.018388
RAMCESS	0.53638	0.042736	0.007324

Table 3.1: Comparison of O_q statistics with ARX-LF and RAMCESS analysis.

[8] Let us also remember that the statistics are evaluated on the whole RAMCESS database.

A_m statistics

Table 3.2 shows that similar conclusions for the asymmetry coefficient α_M. $M_{\alpha_M} = 0.83597$ is an expected value for a normal male voice ($\alpha_M \in [0.7, 0.9]$).

method	M_{α_M}	F_{α_M}	$V^2_{\alpha_M}$
ARX-LF	0.68275	0.032285	0.013499
RAMCESS	0.83597	0.017054	0.002362

Table 3.2: Comparison of α_M statistics with ARX-LF and RAMCESS analysis.

T_L statistics

There is an interesting aspect in the statistics of estimated spectral tilt T_L. We observe that 100% of the frames in the RAMCESS database lead to the minimal modeling error for the case $T_L = 0$. Consequently, when the whole database is processed, the resulting glottal source signal always exhibits an abrupt return phase, which is physiologically impossible. We do not have any further explanation for instance, but it suggests to work deeper on the analysis of the return phase.

3.6.2 Mean modeling error

In Figure 3.20 we can see the distribution of the error e_k for the whole database. The error is evaluated with equation (3.10) between the original frame $x_{V,k}$ and the resynthesized frame $x_{R,k}$. The mean error E is 0.05.

It is interesting to notice that the error with ARX-LF alone is 0.016. It shows that the minimization of the error can not be considered as the only aspect which has to be reduced. Our method slightly increases the mean error, but clearly improves the stability of the extracted parameters (cf. F_i and V_i^2 values for O_q and α_M).

This characteristic of our analysis pipeline explains why the overall synthesis quality is lower than the ARX-LF algorithm (cf. Figure 3.18). However, our approach is more focused on the glottal source modeling, and does not contain the refinements yet for significantly reducing this mean modeling error. We can argue that our analysis/resynthesis

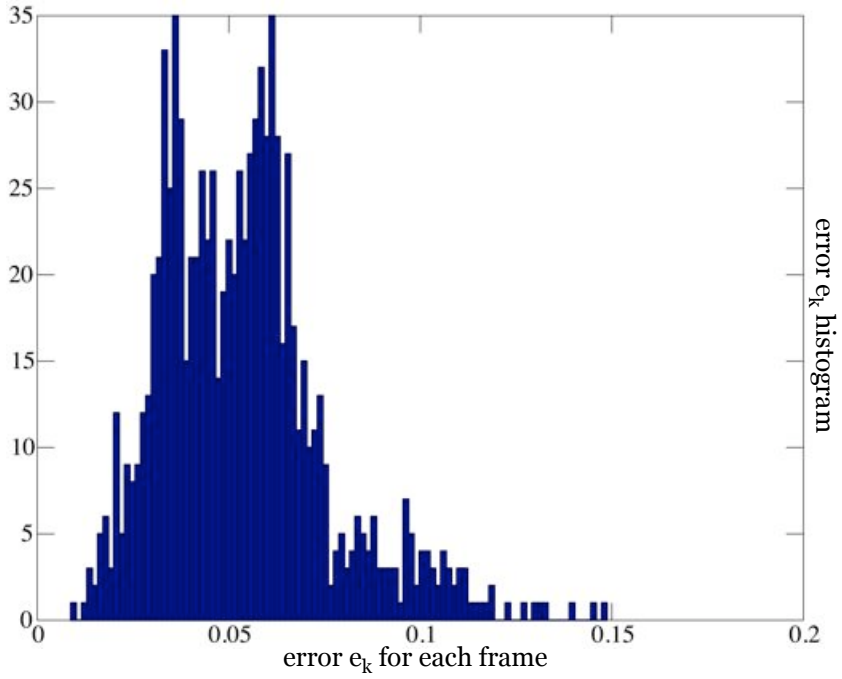

Figure 3.20: Histogram of the error e_k along the whole database.

process is efficient for the glottal source estimation. Drastic voice quality modifications have been achieved convincingly with sentences of the RAMCESS database: chest to head voice conversion, creaky voice, excessive pressure or laxness of vocal folds, etc.

3.7 Conclusions

In this Chapter, we presented the main contributions of the RAMCESS analysis algorithm. This framework is based on the ZZT-based decomposition of causal and anticausal components of voice. Here we summarize the important axes:

Quantification of decomposition efficiency for ZZT-based algorithms

The ZZT-based decomposition of causal and anticausal components of voice is quite a new algorithm. This Chapter has been the opportunity to gather the most significant attempts for optimizing this decomposition, and compare them in one formalism and using the same connected speech database. Finally we propose a new indicator at the ZZT level, C_k, which more focuses the optimization on the glottal formant detection.

Integration of ZZT-based decomposition in the ARX-LF algorithm

The RAMCESS analysis framework is not based on a new algorithm for the estimation of the glottal source parameters. However we have used a pragmatic approach, by combining two promising glottal source analysis algorithms, ZZT and ARX-LF, in order to reinforce the efficiency of the whole analysis. First, ZZT decomposition has been used to extract the anticausal component. Then the glottal formant frequency has been measured and glottal formant parameters (O_q and α_M) have been estimated by fitting of the LF model. Finally, these two parameters have been used in order to reduce the size of the ARX-LF codebook, and constraint the error minimization algorithm.

New indicators for the validation of model-based GF/GFD estimation algorithms: mean values and overall stability of extracted model parameters

We have proposed three new indicators for evaluating the RAMCESS analysis pipeline. Knowing that the ZZT-based decomposition is followed by the fitting of LF and LP (through ARX-LF) models, the statistics of estimated glottal source parameters have been computed. These statistics have been gathered in three indicators, in order to verify the mean values of extracted parameters and their stability over the whole database.

Chapter 4

Realtime Synthesis of Expressive Voice

" More effort results in greater intensity and spectral complexity. "
— John M. Chowning

4.1 Introduction

This Chapter describes the architecture of the RAMCESS synthesizer. This synthesis software is the step following the whole analysis process that is described in Chapter 3. Indeed the database source/filter decomposition is motivated by the aim of manipulating this pre-recorded voice material within a realtime and expressive sound generator.

The RAMCESS synthesis engine aims at respecting our definition of *expressivity*, as introduced in Section 1.1. Thus we need units of the spoken language and a way of delivering these units with subtle degrees of freedom. On the one hand, the spoken language has to be natural and intelligible, and we know that the use of databases is an efficient solution. On the other hand, we aim at giving a refined control on parameters which have a significant impact on vocal expression: prosody and laryngeal voice quality.

Our analysis pipeline allows this combination between prerecorded voice material and subtle control, through the separation of source and filter components, and the approximation of them with controllable models. As the processing of the database is a source/filter deconvolution, the realtime synthesis achieved in RAMCESS is obviously

a source/filter convolution. Although this synthesis technique is well-known and quite simple, we propose several significant improvements:

- the vocal tract LP coefficients are evaluated from the causal component of the ZZT-based decomposition, and are thus a better representation of the real minimum-phase component of the voice coming from the database;
- the glottal source generator is rewritten – based on the LF model [78] – in order to provide a real and flexible period-by-period access to voice source parameters, in realtime and on a large range of the voice quality dimensions;
- the voice source is manipulated through a complex and physiology-related mapping layer, integrating a significant amount of voice quality mechanisms that have been studied in the literature: registers, phonetogram relations, etc.

Chapter starts with an overview of the RAMCESS synthesis architecture and its relation with the analysis process, in Section 4.2. Then we focus on the realtime glottal source generator, in Section 4.3. Section 4.4 describes several mapping layers that aim at linking voice production parameters to voice quality dimensions. Finally a description of the data-driven vocal tract filter is done in Section 4.5.

4.2 Overview of the RAMCESS synthesizer

The main idea of the RAMCESS synthesizer is to convolve vocal tract impulse responses coming from the database, with an interactive model of the glottal source. The interaction with the realtime glottal flow is achieved through a dimensional mapping. The key aspect of this process is already highlighted in the main diagram of the whole work – in Figure 1.4 – but a close-up on the synthesis process is now presented in Figure 4.1.

We work with the database that have been analyzed and decomposed in source and tract components during the analysis process (Chapter 3). We use vocal tract impulse responses from the database by requesting phonetic contents[1]. Then they are converted into geometrical coefficients, like *log area ratio* (LAR_i) and relative areas A_i.

[1] The problem of realtime selection of phonetic contents – such as solutions proposed in the GRASSP system [80] – has only been marginally addressed in this thesis. For instance, we browse the database by sequencing multiphones, going from one stable part of a vowel to the next one, e.g. [astɾo].

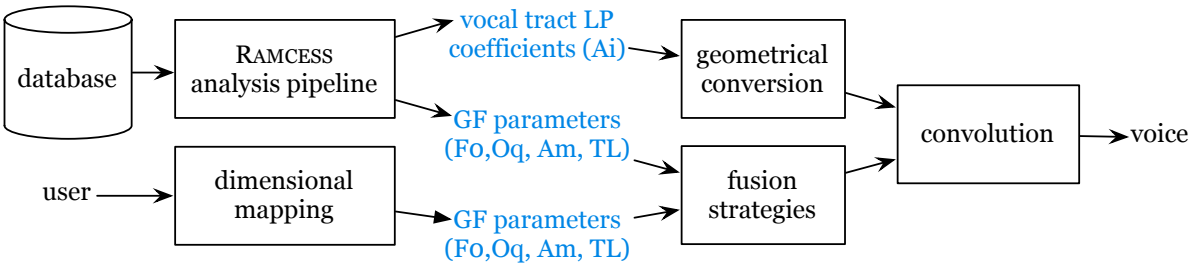

Figure 4.1: Overview of data processing through the RAMCESS synthesizer: using the decomposed database, generating GF parameters through dimensional control, fusing it with database information, and finally convolving with vocal tract impulse responses (converted into geometry-related coefficients).

Vocal tract impulse responses are convolved at synthesis time with a realtime generated (period-by-period) GFD signal. The user of the synthesizer does not interact directly with source parameters. The interaction is achieved through different kinds of dimensional mappings, relying on research in voice quality perception.

The synthesis engine allows GFD parameters to be controlled in two different ways by the user. Indeed two separate streams of GFD parameters can be fused: one from the database and another from the dimensional mapping. In expressive speech it can be interesting for the user to control a deviation on the estimated parameters – e.g. $O_q + \delta O_q$ – in order to alter the recorded expression. But in singing, it is more relevant to provide the user with a whole control of GFD parameters, like a musical instrument.

4.3 SELF: spectrally-enhanced LF model

Source modeling is an old problem, yet it is still currently being studied. Four or five really interesting GF models have emerged in the literature. In Chapter 2 we present two of them: LF [78] and CALM [65]. These two GF models give the most flexible control over the behavior of vocal folds, as they propose five parameters for the shaping of the GF period. However none of them are suitable for interactive use. If a model can produce one period of GF in a reasonable time, it is realtime-friendly. But being interactive also rely on the consistency of the model (stability and interpolability) over a wide range of its parameters. LF and CALM exhibit some consistency problems.

In this Section we start by highlighting these consistency problems for both the LF and CALM models. It gives us the opportunity to introduce our own generator, as a hybrid of LF and CALM. We aim at keeping the best of each model, and solve their respective

problems by combining them. We also discuss the issue of the independent control of energy and spectral tilt.

4.3.1 Inconsistencies in LF and CALM transient behaviors

The problems that are encountered in LF-based and CALM-based glottal flow synthesis are really different. The LF model proposes a non-interpolable parameter space, isolating the smooth sinusoidal phonation from any other phonation type. The CALM model exhibits some over-resonating configurations, due to spectral-based processing.

Non-interpolable LF-based sinusoidal phonation

In equation (2.4) we can see that the LF model is based on the assumption that there is always a GCI in the vibration. This inflection point in the waveform is used as a way of connecting the two sides of the model. This approach is quite efficient if ranges of O_q and α_M are maintained around average male chest voice values [120, 186]:

$$Oq \in [0.4, 0.8] \text{ and } \alpha_M \in [0.7, 0.9]$$

Limiting the LF model to such a narrow use can be somehow confusing. Indeed it is theoretically possible to produce a smooth and "GCI-free" vibration by setting $O_q = 1$ and $\alpha_M = 0.5$. In that case, the LF model produces a sinusoidal waveform. Figure 4.2 shows a sinusoidal GF and its corresponding GFD, as produced by the LF model.

Considering all the possible movements of vocal folds, the sinusoidal phonation is obviously the smoothest[2]. We know that the voice production never leads to a perfectly sinusoidal signal, but the sinusoidal configuration appears as the theoretical boundary of smoothness, useful for approaching the case of quasi-sinusoidal voice.

If we consider that the LF model is able to produce such a smooth phonation type, we could think that the whole range of phonation is accessible. However the sinusoidal configuration is not usable in an interactive context, because this behavior of vocal folds is not interpolable. Indeed it corresponds to a very specific case of the LF-based

[2] If we accept the behavior of the open phase presented in [78], any configuration that proposes $Oq < 1$ and $\alpha_M > 0.5$ is less smooth (spectrally richer) than the sinusoidal phonation.

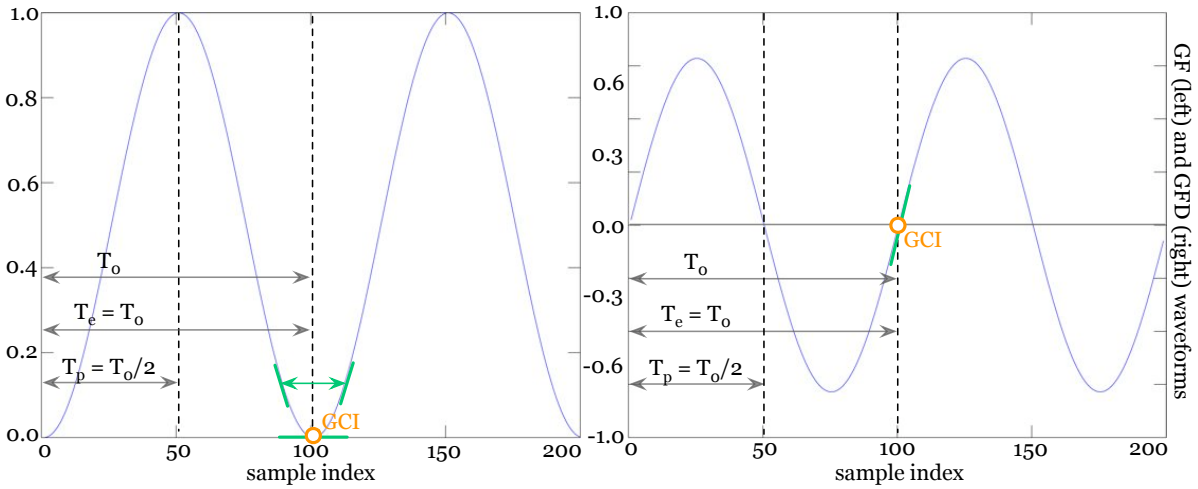

Figure 4.2: Two periods of GF (left) and GFD (right) computed with the LF model for $O_q = 1$ and $\alpha_M = 0.5$. f_0 is 160Hz, with $F_s = 16$kHz. We observe the location of the GCI (orange) and the symmetry of the GF/GFD (green) around it.

GF/GFD synthesis, where some features are synchronous. In Figure 4.2 we also highlight three aspects of this particular alignment:

- The GCI is not missing, but occurs at the end of the period (orange circle). This location results from the value of O_q which leads to $T_e = O_q \times T_0 = T_0$.
- The GCI location on the GF corresponds to a smooth return to zero. This smooth return results from the perfect symmetry within the period: $T_p = \alpha_M \times T_e = T_0/2$. Thus the GFD is truncated exactly on zero crossing, with no discontinuity.
- This perfect symmetry of the open phase also makes two consecutive GFD periods to perfectly match their slope at the junction (green lines), with no breakpoint.

If we observe the GF from the physiological point of view, it appears that this particular configuration emulates a soft and long return phase[3], that starts when the waveform concavity goes from negative to positive. On the GFD, its means that the "disappeared" GCI stands around the negative minimum of the sinusoidal period. In Figure 4.3 this expected GCI is highlighted (orange circle) on sinusoidal GF and GFD (blue curves).

If we target to interpolate this configuration with a more tensed phonation, we expect the vibration to progressively go out of symmetry, with the return phase decreasing at the same time. Thus the GCI would "reappear" on the GFD by the narrowing of

[3] In this case, we mention the return phase from the physiological point of view, because the model-based return phase T_a is zero for this configuration of the LF model.

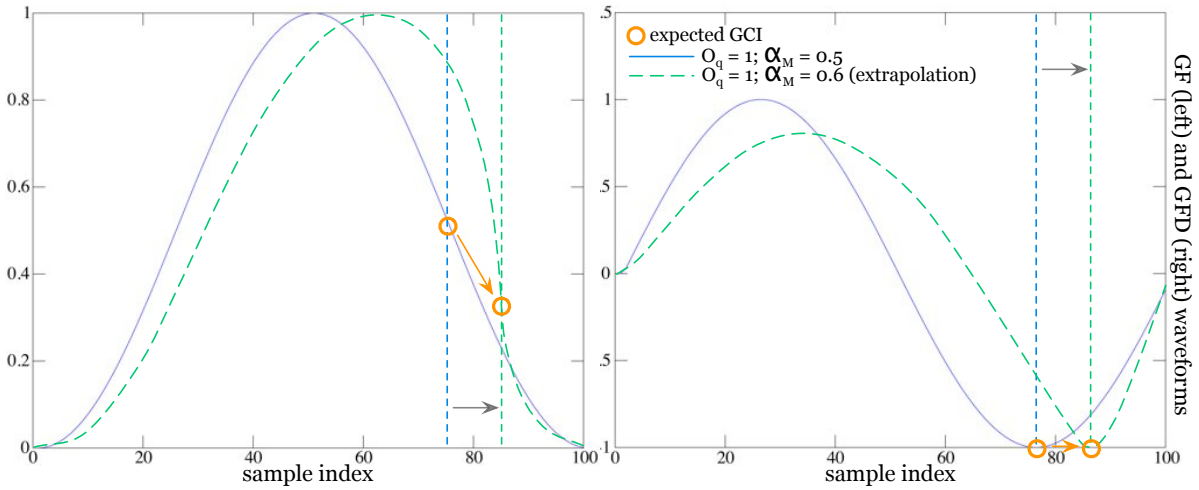

Figure 4.3: One period of GF (left) and GFD (right). The expected GCI is highlighted (orange circle) on the sinusoidal pulse (blue), and the ideal evolution to a more tensed pulse (dashed green) is suggested: asymmetry increases and return phase decreases on the GF; narrowing of the GCI happens on the GFD.

the negative minimum of the sinusoidal period into the well known inflection point. Figure 4.3 represents the ideal evolution (not LF-based) of GF and GFD when moving to the configuration $O_q = 1$ and $\alpha_M = 0.6$ (green dashed). We observe how the GCI location would be consistent, moving progressively with the increasing of the asymmetry.

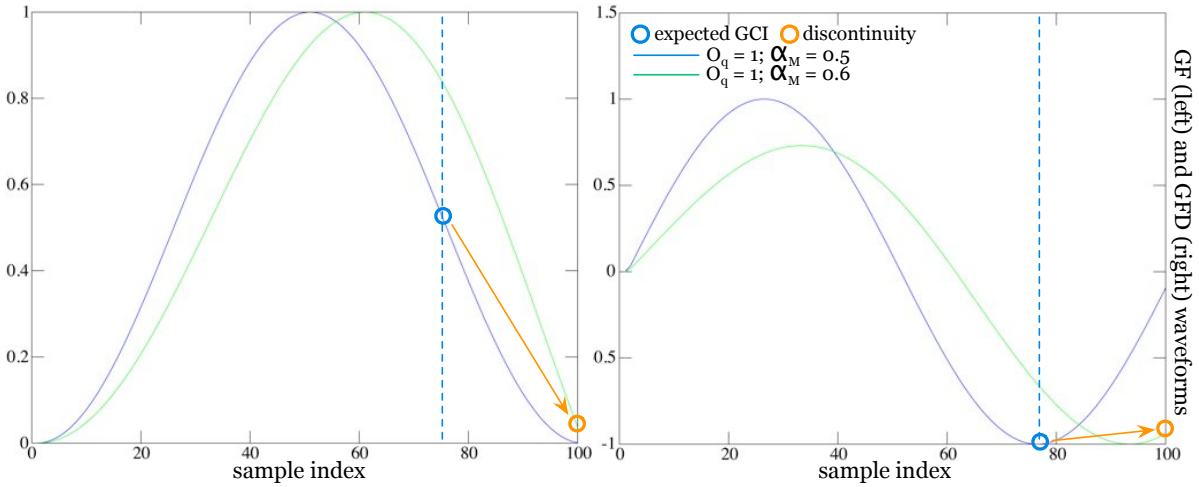

Figure 4.4: One period of GF (left) and GFD (right) computed with the LF model for two situation: always $O_q = 1$, but $\alpha_M = 0.5$ (blue) and $\alpha_M = 0.6$ (green). We observe the inconsistent shift from the expected GCI (blue circle) in the sinusoidal pulse to real appearing discontinuity (orange circle) in the more tensed pulse.

However this ideal evolution is not compatible with the LF model because, in the LF-based sinusoidal phonation, the GCI is theoretically at the end of the period. Thus when

the waveform goes out of symmetry, it directly creates a truncation of the sinusoid in the GF – and thus a discontinuity in the GFD – at the end of the period. Figure 4.4 compares the configuration $O_q = 1$ and $\alpha_M = 0.6$ (green) with the sinusoidal source (blue), both achieved with the LF model. The truncation is highlighted (orange).

In Figure 4.5 we observe the spectral effect of a small variation of O_q and α_M, from the position $O_q = 1$ and $\alpha_M = 0.5$. The magnitude spectra of several GFD periods are compared. We observe the sudden increase of high frequencies due to the discontinuity in the GFD, appearing at the end of the period. The value $\alpha_M = 0.51$ is chosen.

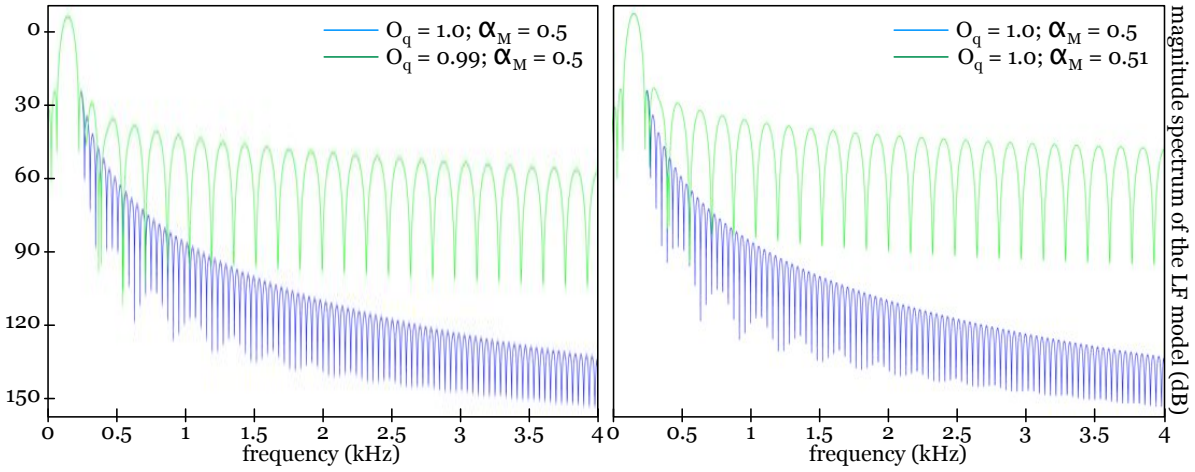

Figure 4.5: Comparison between the sinusoidal phonation (blue) and two close configuration: $O_q = 0.99$ (left) and $\alpha_M = 0.51$ (right). The modified configurations (green) contain more high frequency.

Normally the return phase aims at smoothing transitions between open and closed phases. But the LF model computes the return phase in the time domain, as the decreasing exponential connection between the GCI and the closed phase. Consequently the smoothing ability of the LF-based return phase is useless for O_q values close to 1.

CALM-based over-resonating configurations

The CALM glottal flow synthesis is based on spectral processing. The waveform of the the open phase is computed as the impulse response of the second-order anticausal filter $H_1(z)$, described in equations (2.9) to (2.11), and deeply addressed in [65].

Depending on filter parameters, a second-order impulse response can be damped or oscillating. In the computation of $H_1(z)$ coefficients from O_q and α_M targets, some configurations lead to oscillations when the anticausal processing is achieved [52].

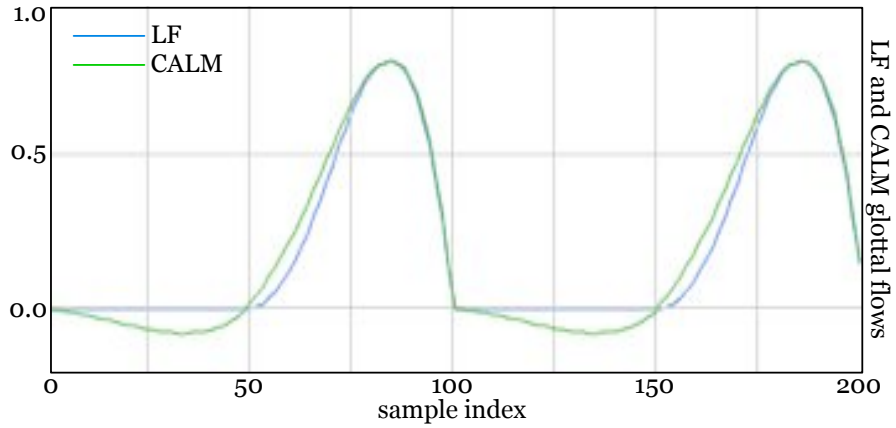

Figure 4.6: Comparison between open phases of the LF and the CALM models, with the configuration $O_q = 0.5$ and $\alpha_M = 0.7$. The CALM model exhibits oscillations.

In Figure 4.6 we compare open phases of the LF and the CALM models, for a particular configuration: $O_q = 0.5$ and $\alpha_M = 0.7$. We can see that the CALM model exhibits an oscillating behavior, as the open phase is not totally damped. As the GF is always positive or null, this waveform is not acceptable from the physiological point of view.

4.3.2 LF with spectrally-generated return phase

The GF/GFD production model that we propose keeps the best of both LF and CALM approches. Indeed it keeps the time domain computation of the open phase, based on an adaptation of LF equations, but it manages the return phase through its spectral effect, the spectral tilt, using a digital filter. Consequently we call this new generator SELF, as it is based on *Spectrally-Enhanced LF* modeling.

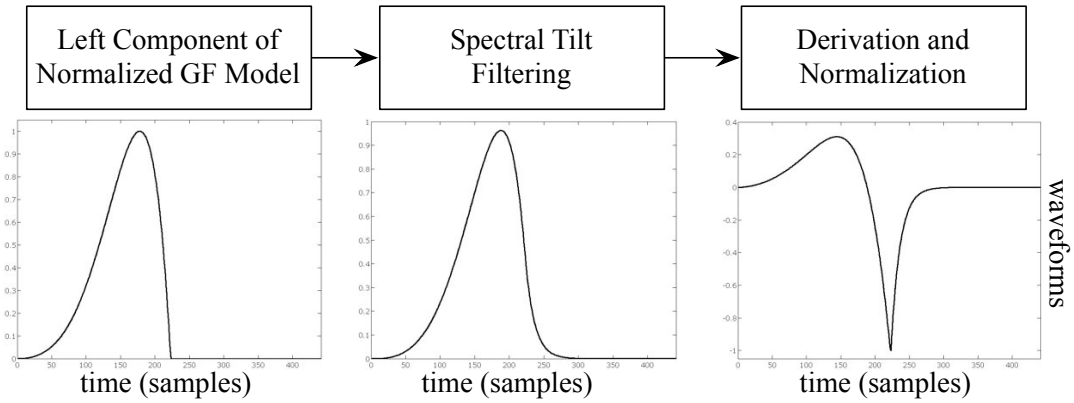

Figure 4.7: The three main steps of the SELF-based synthesis: generating the left component of the integrated LF model, the spectral tilt filter, derivating and normalizing.

In Figure 4.7 we see the basic steps of the synthesis procedure. The open phase of the GF is computed in the time domain as the left component of the integrated LF model (before the GCI). Then this signal with an abrupt closure is processed by the first-order spectral tilt filter $H_2(z)$, as described in [65]. Finally the signal is derivated in order to get the GFD, and the waveform is normalized so that the negative peak equals $-E$.

SELF solves the problems of both LF and CALM used separately:

- The synthesis of the anticausal part in the time domain (based on the LF model) avoids over-resonating configurations that are encountered in CALM processing.
- High spectral tilt values ($T_L > 20$dB) lead to a long return phase. This return phase is convolved with all the samples of the LF period, resulting in the smoothing of the discontinuities that have been highlighted in Section 4.3.1. From the spectral point of view, the spectral tilt filter (first order low-pass) is used to manage the transition between the two separate spectra illustrated in Figure 4.5.

Realtime synthesis of LF anticausal component

Producing an LF-based GF pulse without any return phase (abrupt closure) rather simplifies the problem. Indeed, as described in Section 2.2.2, the complexity in the LF model comes from the time domain adjustment of the two curves. This adjustment is characterized by a system of two implicit equations to be solved, for parameters a and ϵ. If we don't use a return phase, it corresponds to the theoretical case $\epsilon = \infty$. The system of equations (2.5) and (2.6) can be simplified to one single implicit equation:

$$\frac{1}{a^2 + (\frac{\pi}{T_p})^2} \left(e^{-aT_e} \left(\frac{\frac{\pi}{T_p}}{sin\frac{\pi T_e}{T_p}}\right) + a - \frac{\pi}{T_p} cot\frac{\pi T_e}{T_p}\right) = 0 \tag{4.1}$$

Moreover we normalize both the period ($T_0 = 1$) and the open phase ($T_e = T_0$) of the GF period. It gives the relation $T_p = \alpha_M$ which highlights the relation $a = f(\alpha_M)$:

$$\frac{1}{a^2 + (\frac{\pi}{\alpha_M})^2} \left(e^{-a} \left(\frac{\frac{\pi}{\alpha_M}}{sin\frac{\pi}{\alpha_M}}\right) + a - \frac{\pi}{\alpha_M} cot\frac{\pi}{\alpha_M}\right) = 0 \tag{4.2}$$

In SELF, the solution of $a = f(\alpha_M)$ is obtained offline for 100 values of α_M, and saved in some table A. It gives a non-linear relation between the two parameters (cf. Figure 4.8). During realtime synthesis, the exact value of a is obtained from the current value of α_M by a linear interpolation between the two closest entries in A.

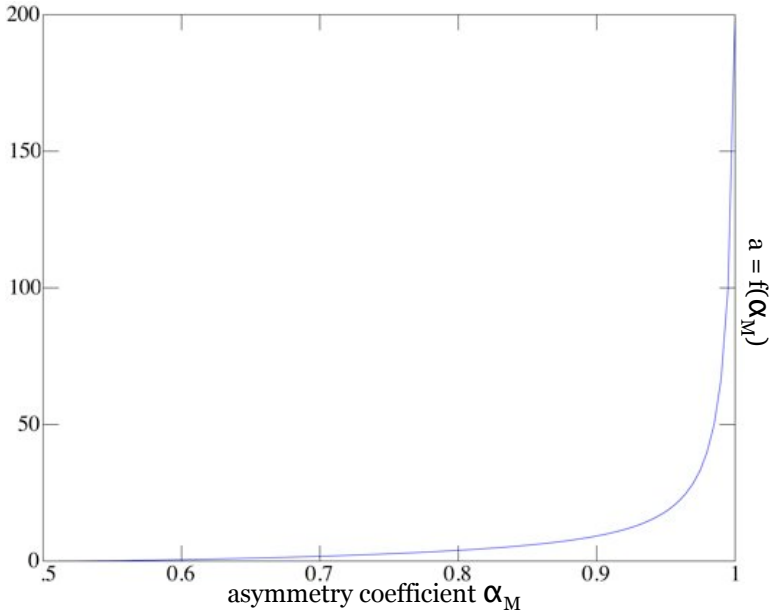

Figure 4.8: Solution of equation (4.2) for 100 values of α_M.

Equation 4.3 gives the normalized GF period, as presented in [64]. Figure 4.9 illustrates a typical example of the function $n_g(t)$ with $\alpha_M = 0.7$ ($a = 1.7595$).

$$n_g(t) = \frac{1 + e^{at}(a\frac{\alpha_M}{\pi} sin(\pi t/\alpha_M) - cos(\pi t/\alpha_M))}{1 + e^{a\alpha_M}} \tag{4.3}$$

This continuous function of time t computes the normalized open phase in the interval $t = [0, 1[$. Then the scaling of the waveform to the requested open phase $T_e = O_q \times T_0$ is achieved by sampling the continuous signal $n_g(t)$ with the appropriate sampling step t_s:

$$t_s = \frac{F_0}{O_q \times F_s}$$

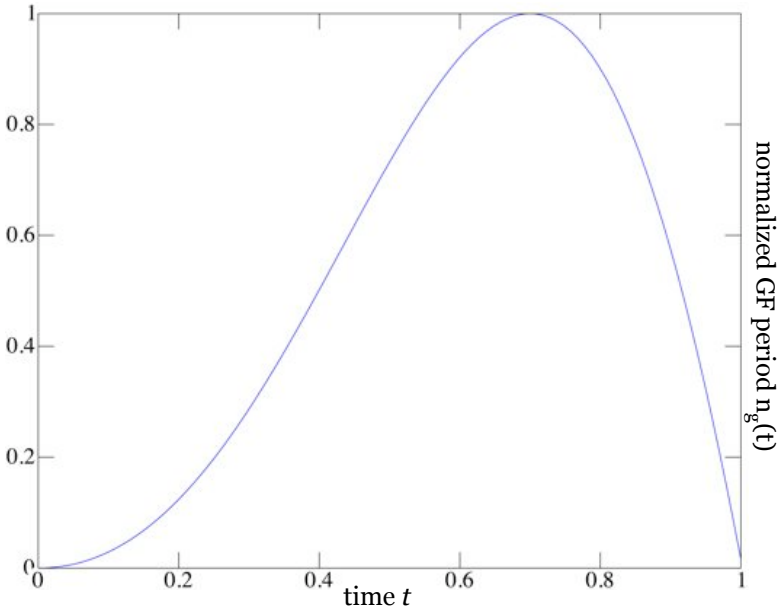

Figure 4.9: Normalized GF period, as described in [64]. $T_0 = 1$ and $O_q = 1$. The choice of α_M defines a and gives the asymmetry of the waveform.

Independant control of energy and spectral tilt

As illustrated in Figure 4.7, the SELF model computes the derivation and the normalization of each GF period. The need for the normalization comes from the use of the spectral tilt filter. Indeed the low-pass filtering significantly reduces the overall energy of the signal. If we want to control independently the energy and the spectral coloring due to the tilt, we have to achieve a post-normalization.

There are various approaches in the normalization of a periodic signal, but we use a really simple idea. Indeed the literature highlights that the negative peak of the GFD has a huge impact on the perception of vocal intensity. Thus we work with a simple rescaling of the negative peak, in order to have a controllable E factor.

However it has been noticed by experiment[4] that the realtime normalization can not be achieved after the spectral tilt operation. It creates hearable clicks, due to inappropriate amplitude modifications on ongoing $H_2(z)$ impulse responses. Indeed if one impulse response goes across two consecutive frames, and these frames are normalized independently, this impulse response can be discontinued, resulting in a click.

[4] The highlighting of these kind of realtime synthesis problems from the Analysis-by-Interaction (AbI) methodology that is used in this thesis. AbI is presented in Chapters 6 to 8.

In SELF, these steps are interchanged. First the derivation is achieved directly on the open phase. It defines a first value for the negative peak value: E_o. Then the impact of the spectral tilt filtering on E_o is evaluated. We measure the value of the negative peak after the spectral tilt filtering: E_a. We define α, the correction factor:

$$\alpha = \frac{E_o}{E_a}$$

The normalization process is achieved directly on the open phase. We expect to set the negative peak of the GFD to a target value: E_t. Without any spectral tilt filtering, the normalization factor β is simply the ratio of E_t and E_o:

$$\beta = \frac{E_t}{E_o}$$

Knowing that the spectral tilt filtering modifies the amplitude of the negative peak by a factor δE, we compute the corrected normalization factor β':

$$\beta' = \alpha \frac{E_t}{E_o} = \frac{E_t}{E_a}$$

The whole GFD synthesis process is illustrated in Figure 4.10. The final implementation of SELF contains several other options, such as the amplitude correction strategy applied on the GF maximum, in order to control A_v instead of E.

4.4 Voice quality control

Modifying production parameters directly, such as O_q or α_M, does not have a strong perceptive effect on listeners. These modifications rather corresponds to synthetic-like transformations of the sound. Indeed production parameters change synchronously along several perceptive dimensions, as they have been described in Section 2.3.1. Perceiving effects like tenseness or vocal effort depends on these synchronous and interdependent movements. This Section aims at discussing this *voice quality control* issue.

As the dependency between production and perception is really complex, strategies have to be decided in order to implement the voice quality control layer. In this work, we present two different approaches:

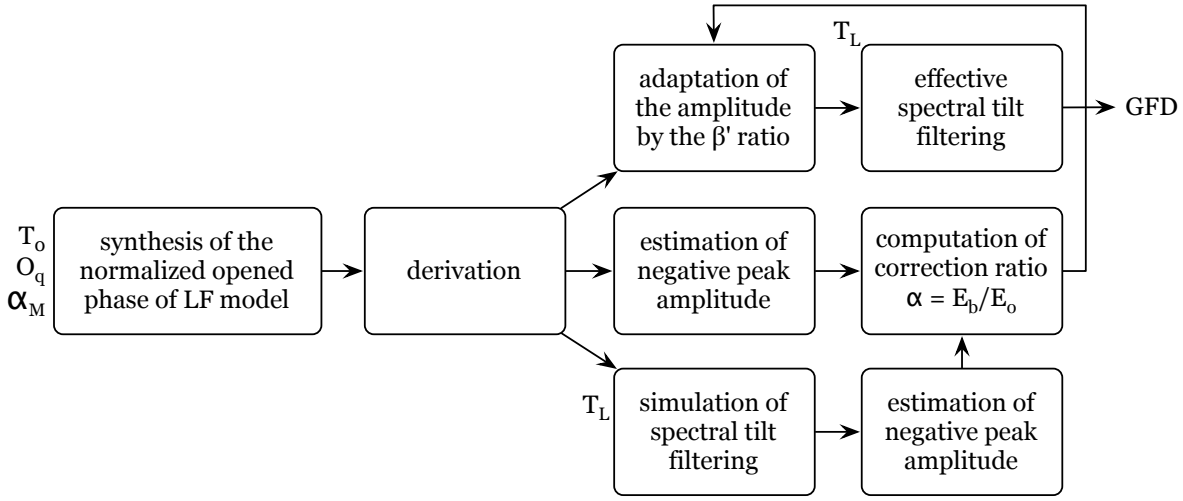

Figure 4.10: Synthesis of the GFD in the SELF engine: a first simulation of the spectral tilt filtering is performed in order to compute the amplitude correction factor α and then apply it to the normalization factor β'.

- The first strategy is simple and mono-dimensional: we gather all the voice quality variations in one single "presfort" axis, in Section 4.4.1.
- The second idea implements more voice quality mechanisms, as in the literature: tension, vocal effort, registers, the phonetogram, in Sections 4.4.2 and 4.4.3.

4.4.1 Mono-dimensional mapping: the "presfort" approach

Many studies show that the voice quality is multi-dimensional. Two perceptual effects are widely discussed: effort and tension (or lax/pressed dimension). However it appears to be quite interesting to test a mono-dimensional mapping. Indeed multi-dimensional mappings require the user to be initiated to voice quality. One single "spectral coloring" axis, with a rather caricatural behavior, can be handled more intuitively.

We call this axis *presfort*, because we gather both the idea of tenseness and effort. The mapping is made by using an interpolation factor y between two configurations of parameters O_q , α_M and T_L. The interpolation is achieved between a "soft and lax" ($y = 0$) to a "loud and tensed" voice ($y = 1$). Values for O_q , α_M and T_L are chosen empirically. The "soft and lax" extremum corresponds to the sinusoidal pulse $\{O_q = 1;\ \alpha_M = 0.5;\ T_L = 30\}$, and the "loud and tensed" extremum corresponds to a asymmetric and bright pulse $\{O_q = 0.6;\ \alpha_M = 0.75;\ T_L = 2\}$. Figure 4.11 illustrates

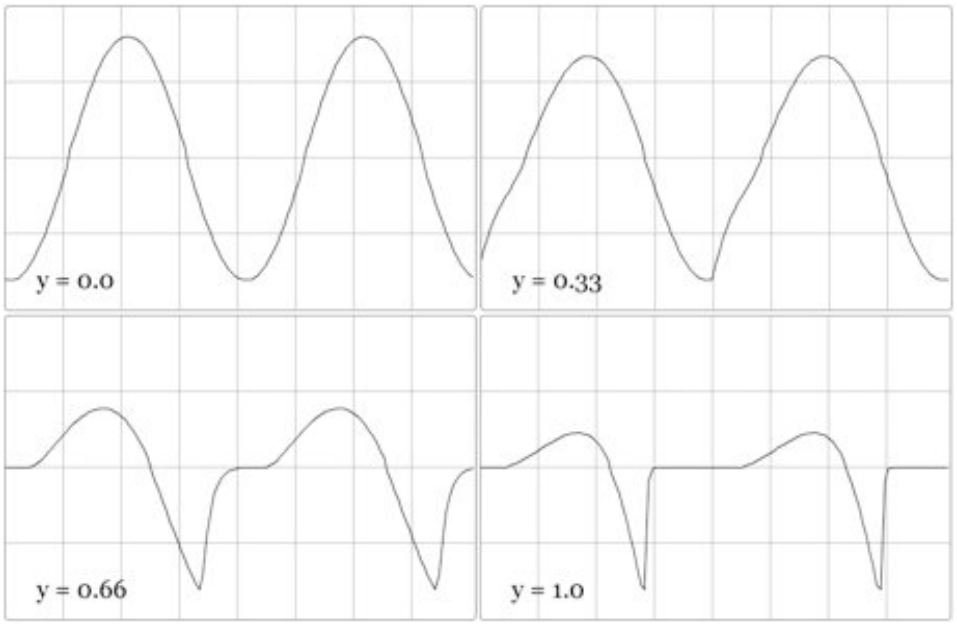

Figure 4.11: Four snapshots of the glottal pulses, with different values for the y interpolation factor. From a soft quasi-sinusoidal vibration ($y = 0$) to an creaky voice ($y = 1$).

the evolution of the SELF-based GFD along values of the interpolation factor y, and equations in (4.4) give the interpolation coefficients.

$$\begin{cases} O_q &= 1 - 0.4 \times y \\ \alpha_M &= 0.5 + 0.25 \times y \\ T_L &= 30 - 28 \times y \end{cases} \tag{4.4}$$

In Figure 4.11 we also see that the SELF synthesizer has the expected behavior, even with a rather simple mapping. Indeed we can see that, while the voice is tensing/getting louder, the negative oscillation of the sinusoidal phonation is progressively converted into a GCI, without any discontinuity or inconsistency.

The perceptual effect of this mapping on listener is the hearing of a clear and strong effort effect. It leads the user to manipulate this axis really carefully. If this axis is added to F_0 and E controls, the whole voice production can be expressively manipulated with only a 3-axes controller, such as joystick, faderbox, glove, camera tracking, etc.

4.4.2 Realtime implementation of the phonetogram effect

From the point of view of voice analysis, the phonetogram is the shape that can be observed when every frame of a database is plotted on the *(pitch, intensity)* map. This shape is speaker-dependant because it relies on the properties of the larynx [100]. As illustrated in Figure 4.12 the shape of the phonetogram highlights that low (white) and high (black) boundaries of the vocal effort[5] are pitch-dependent.

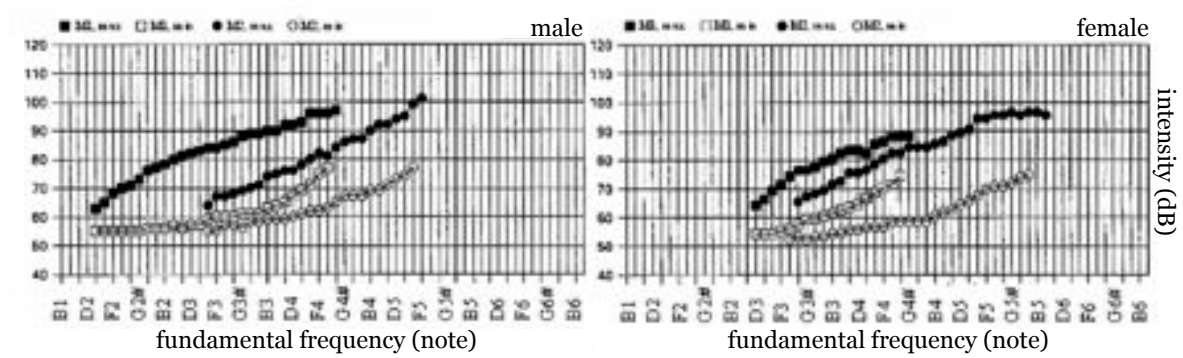

Figure 4.12: Male (left) and female (right) phonetograms: low (white) and high (black) intensity boundaries are illustrated depending on fundamental frequency. Modal (M_1) and head (M_2) register phonetograms are represented [100].

The curves in Figure 4.12 illustrate that the vocal effort range is limited, for a given fundamental frequency. Particularly this range decreases when a low pitch sound is produced. Moreover we observe two different and overlapped drawings. They correspond to chest (M_1) and head (M_2) phonations, meaning that the mechanism M_i influences the relation between fundamental frequency and vocal effort.

In this work we aim at reproducing this property of the larynx at the synthesis time. Boundaries of the phonetogram are estimated from the recording of a given speaker or singer. When all the frames of this recording are plotted on the *(pitch, intensity)* map, we can highlight low and high boundaries of both chest (M_1) and head (M_2) voice.

Once these boundaries have been determined, the phonetogram can be encoded as breakpoint functions[6]. There are four breakpoint functions $V_{k,M_i} = f(F_0)$. Indeed we have low and high boundaries, for both chest and head phonetograms. Each breakpoint func-

[5] The vocal effort is a perceptual dimension. Thus we can only measure the closest physical property which is the intensity of the phonation. The vocal effort influences intensity and spectral tilt [120].

[6] A *breakpoint function* is a function that is defined by a limited sequence of (x, y) points. Every intermediate value is evaluated by interpolating from the two closest (x, y) entries.

tion allows us to summarize one boundary with 5 or 6 points. Intermediate values are linearly interpolated. This breakpoint modeling is illustrated in Figure 4.13.

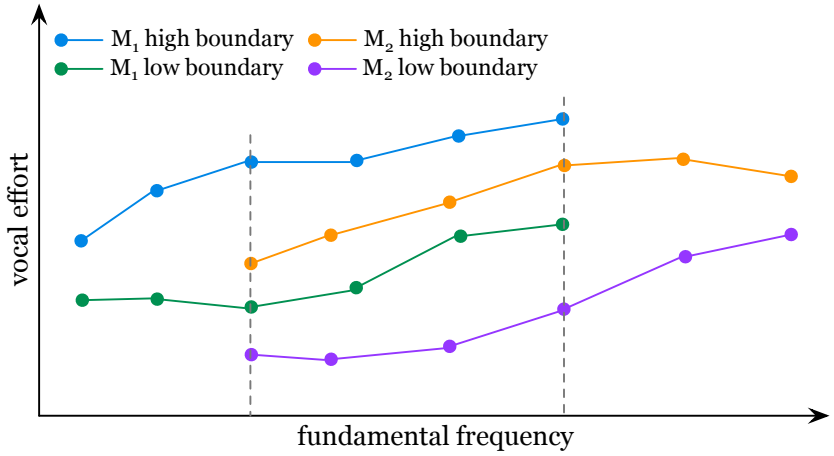

Figure 4.13: Modeling of M_1/M_2 phonetograms with four breakpoint functions: low (green) and high (blue) boundaries in chest voice, low (purple) and high (orange) boundaries in head voice. Dashed lines highlight the overlapping region.

The use of the phonetogram influences the value of the vocal effort. Indeed an absolute vocal effort V_{A,M_i} is computed by the linear interpolation between two boundaries V_{low,M_i} and V_{high,M_i}, for a given mechanism M_i. The interpolation index between these two boundaries can be seen as the relative vocal effort, $V_R \in [0, 1]$:

$$V_{A,M_i} = V_{high,M_i} - (V_{high,M_i} - V_{low,M_i}) \times V_R \tag{4.5}$$

The phonetogram is the first voice quality mechanism that is implemented in our voice quality control layer. In the following development, the absolute vocal effort V_{A,M_i} is simply denoted as the vocal effort V. Indeed the influence of the mechanism M_i leads to separate equations. Consequently the mechanism M_i is explicitly mentioned: V if M_i. We think that combining current and following notations leads to difficult reading.

4.4.3 Vocal effort and tension

In this part, we propose a set of relations between the source parameters $\{O_q\ \alpha_M\ T_L\}$ and two important voice quality dimensions: the vocal effort V and the tension T. We also highlight the dependency between these dimensions and the mechanism M_i that

is used in the phonation. This set of relations aims at being a summary and a first proposition, regarding theories that are explained in Section 2.3.

The issue is the interdependency between source parameters and voice quality dimensions. Indeed both T and V have an impact on O_q, α_M and T_L values. In this work, we combine effects of T and V by considering two "orthogonal" processes in the control:

- On the one hand, the vocal effort V and mecanisms M_i control "offset" values: O_{q_0}, α_{M_0} and T_{l_0}. The vocal effort is considered as the main spectral modification that influence the perception of voice quality.
- On the other hand, the tenseness T controls "delta" values of O_q and α_M around their offset configuration; ΔO_q, $\Delta\alpha_M$. The tenseness is considered as a extra phenomenon, that happens on the top of the main vocal effort situation.

Consequently, synthesis parameters can be described as:

$$\begin{cases} O_q &= O_{q,0} &+ \Delta O_q \\ \alpha_M &= \alpha_{M,0} &+ \Delta\alpha_M \\ T_L &= T_{L,0} & \end{cases} \tag{4.6}$$

In the following development, V and T are normalized between 0 and 1. $V = 0$ is the softest phonation, $V = 1$ the loudest. $T = 0$ is the laxest configuration, $T = 1$ the most pressed. M_i can be M_1 (chest) or M_2 (head).

Vocal effort mapping

In this Section we present the mapping for the vocal effort dimension. This mapping (particularly the boundaries that are chosen) results from a compromise between what is mentioned in the literature and empirical adjustment. This leads to equations for offset values $O_{q,0}$, $\alpha_{M,0}$ and $T_{L,0}$. Theses equations are presented for both M_1 and M_2.

$\boldsymbol{O_{q,0} = f(V, M_i)}$ – The vocal effort V linearly modifies the value of $O_{q,0}$ between in $[0.8, 0.4]$ for chest voice M_1, and $[1.0, 0.6]$ for head voice M_2 [120, 186]:

$$O_{q,0} = \begin{cases} 0.8 - 0.4 \times V & \text{if} \quad M_1 \\ 1.0 - 0.4 \times V & \text{if} \quad M_2 \end{cases} \tag{4.7}$$

$\boldsymbol{\alpha_{M,0} = f(M_i)}$ – The vocal effort V does not influence continuously the value of $\alpha_{M,0}$. Only the mechanism M_i sets $\alpha_{M,0}$ to 0.8 for M_1, and 0.6 for M_2 [102]:

$$\alpha_{M,0} = \begin{cases} 0.8 & \text{if} \quad M_1 \\ 0.6 & \text{if} \quad M_2 \end{cases} \tag{4.8}$$

$\boldsymbol{T_{L,0} = f(V)}$ – The vocal effort V linearly modifies the value of $T_{L,0}$ in the range $[6, 55]$ (in dB). This particularly high values for the spectral tilt aim at achieving the smoothing of LF discontinuities, as described in 4.3.2:

$$T_{L,0} = 55 - 49 \times V \text{ (dB)}$$

Tenseness mapping

The strategy used in the tenseness mapping is based on centered deviations. Actually the configuration $T = 0.5$ does not modifies $O_{q,0}$ and $\alpha_{M,0}$ offset values. If T goes out of this center, ΔO_q and $\Delta\alpha_M$ are progressively applied within determined boundaries.

$\boldsymbol{\Delta O_q = f(T)}$ – The tenseness T creates a deviation ΔO_q that also depends on $O_{q,0}$. Indeed if $T = 0.5$, $\Delta O_q = 0$. If $T = 0$, we want $O_q = 0.4$. Thus $\Delta O_q = -O_{q,0} + 0.4$. The same for $T = 1$, we want $O_q = 1$ and thus $\Delta O_q = -O_{q,0} + 1$. These deviations lead to the same extreme boundaries $O_q \in [0.4, 1]$ that we chose in the vocal effort mapping:

$$\Delta O_q = \begin{cases} (2T - 1) \times O_{q,0} - 0.8T + 0.4 & \text{if} \quad T \leq 0.5 \\ (1 - 2T) \times O_{q,0} + 2T - 1 & \text{if} \quad T > 0.5 \end{cases} \tag{4.9}$$

$\boldsymbol{\Delta\alpha_M = f(T)}$ – The same process is applied for adapting the $\Delta\alpha_M$ value. In this case, $T = 0.5$ also lead to $\Delta\alpha_M = 0$. $T = 0$ gives $\alpha_M = 0.8$, and $T = 1$ gives $\alpha_M = 0.6$.

$$\Delta\alpha_M = \begin{cases} (1-2T) \times \alpha_{M,0} + 1.2T - 0.6 & \text{if} \quad T \geq 0.5 \\ (2T-1) \times \alpha_{M,0} - 1.6T + 0.8 & \text{if} \quad T < 0.5 \end{cases} \tag{4.10}$$

4.5 Data-driven geometry-based vocal tract

In this section, we describe the implementation of a vocal tract model. This module is based on a physical "tube-based" representation of vocal tract filter, which is simultaneously controllable with geometrical (areas) and spectral (formants) parameters.

LP coefficients a_i are not linearly interpolable. This means that, for two configurations $[a_1 a_2 ... a_n]$ and $[b_1 b_2 ... b_n]$ corresponding to two vowels, a linear interpolation between both of these vectors does not correspond to a linear interpolation between the two magnitude spectra, and could even lead to unstable combinations.

Consequently, we use another implementation of the all-pole filter, called the *lattice filter*. The control parameters of such a filter are called *reflection coefficients* and commonly named k_i. Such a filter is represented in Figure 4.14. It is composed of successive sections. Each of them is characterized by a k_i parameter [134].

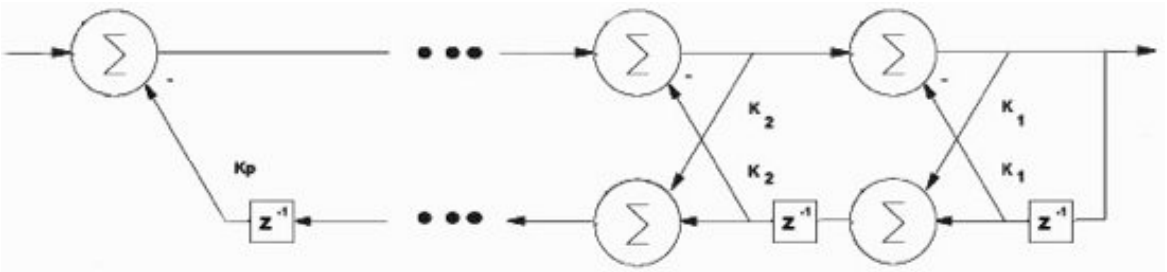

Figure 4.14: Representation of p cells of a lattice filter.

Reflection coefficients correspond to physical characteristics of the vocal tract, which may be represented by the concatenation of cylindrical acoustic resonators, forming a lossless tube. This physical model of the lattice filter is represented in Figure 4.15.

Each filter section represents one section of the tube. The forward wave entering the tube is partially reflected backwards, and the backward wave is partially reflected forwards.

The reflection parameter k_i can then be interpreted as the ratio of acoustic reflection in the i^{th} cylindrical cavity, caused by the junction impedance with the adjacent cavity. This value varies from 1 (total reflection) to -1 (total reflection with phase inversion), and is equal to 0 when there is no reflection.

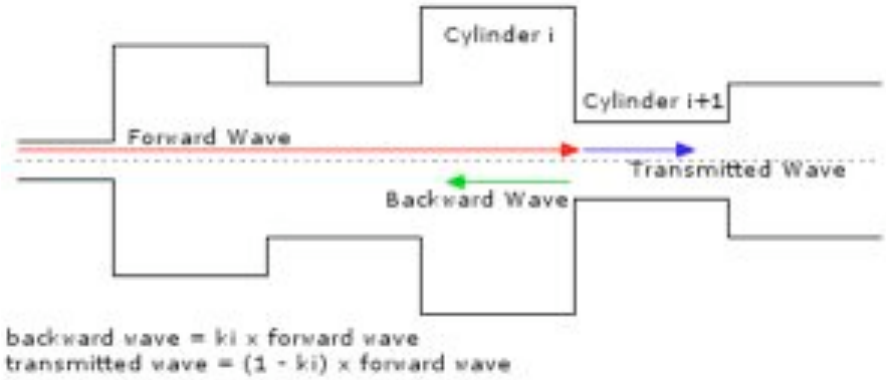

Figure 4.15: Geometrical interpretation of the lattice filter: transmitted and backwards waves at each cell junction.

The filter is stable if $k_i \in]-1, 1[$. However there is no direct relation between these parameters and the sound: a small modification of k_i does not lead to a small modification of the spectrum. Therefore, instead of using the reflection coefficients, we manipulate relative areas A_i, which can be computed from reflection coefficients:

$$\frac{A_i}{A_{i+1}} = \frac{1 + k_i}{1 - k_i}$$

We use A_i coefficients in order to interpolate vocal tract impulse responses that come from the database. In the realtime processing, an interpolating window of 30ms smoothes the transition between consecutive frames. The combination of A_i interpolation and lattice filter structure provides a clean and flexible vocal tract.

4.6 Conclusions

In this Chapter we presented the main axes that underlie the RAMCESS synthesis engine. This synthesizer focuses on the realtime interaction with expressive voice material. Here we present several important aspects of this part of the thesis work:

Source/filter synthesizer based on causal/auticausal decomposition

The RAMCESS synthesis engine achieves a convolution in realtime. Components used in the convolution result from the ZZT-based causal/anticausal decomposition of the RAMCESS database, and the LF modeling of the glottal source. Due to the database, the analysis/resynthesis process leads to a natural and intelligible voice, but the modeling of the glottal source gives the possibility of deeply modifying the phonation properties.

New model for the synthesis of the glottal source signal

A significant work has been done in order to propose a realtime, flexible and consistent glottal source synthesizer. SELF (*Spectrally-Enhanced LF*) is the combination the building of a waveform segment in the time domain (the anticausal part of the LF waveform) and the processing of the return phase (the causal part of the LF waveform) in the spectral domain, by using the spectral tilt parameterization of CALM.

New mapping strategies for the control of voice quality

Several mapping strategies have been presented in order to connect voice quality dimensions – such as tension or vocal effort – to voice production parameters. Particularly, one mapping called the "presfort" approach, aims at being appropriate for controlling the voice quality with a limited amount of dimensions, typically 3-axis controllers. The other proposed strategy aims at combining vocal effort, tension and the effect of the phonetogram in one control space, available for realtime interaction.

Chapter 5

Extending the Causal/Anticausal Description

“ Imagination is more important than knowledge. ”
— Albert Einstein

5.1 Introduction

Sometimes searching in a given direction gives the opportunity to reveal interesting ideas in different topics. In this Chapter, we describe how it is possible to extend the mixed-phase approach (cf. Section 2.2.4) and the analysis framework of Chapter 3 to some typical continuous interaction instruments (CII): brass and bowed string instruments. The aim of this sidetrack is to better characterize CII waveforms through a meaningful representation of magnitude and phase spectra, with the hope of using these models in the realtime expressive transformation of CII sounds [73].

This Chapter is clearly a preliminary exploration and does not lead to any formalization yet. We propose this causal/anticausal representation of CII sounds as a milestone for further work. In Section 5.2, we discuss some causality issues in various sustained sounds, produced by brass and bowed string instruments. Then we describe the causal/anticusal decomposition applied to several CII sounds, in Section 5.3. Finally Section 5.4 presents the first results in the modeling and resynthesis of these CII sounds.

5.2 Causality of sustained sounds

Vocal folds movements can be seen as sequences of two generalized phenomena [64]. On the one hand, an *opening phase*: progressive displacement of the position of the system, from its initial state, resulting from a combination of continuous external forces and inertia reaction. On the other hand, a *closing phase*: sudden return movement, appearing when the previously constrained system reaches its elastic displacement limit. In this Section, we show that similar opening/closing sequences can be found in typical CII excitation mechanisms, like in brass or bowed string instruments.

Causality in brass instruments

Analogies between vocal folds and lips at a mouthpiece are particularly clear. High pressure inside the mouth creates constrained displacements and quick returns of the lips [47]. Modulations are achieved by the embouchure of the musician. Moreover pressure measured at the mouthpiece shows similar anticausal aspects as ones observed in glottal flow [16], as can be seen in Figure 5.1a. Indeed we can observe something similar to a "closure instant" with a negative peak (blue dashed), and the waveform preceding (following) this negative peak exhibits some divergent (convergent) characteristics.

Causality in bowed-string instruments

Literature related to bowed string modeling assumes that the bow-string interaction follows a periodic and distributed stick-slip scheme. Serafin proposed a dynamic elasto-plastic friction model [171]. This approach gives a distributed interpretation of the friction mechanism, which is itself locally represented by a "spring and damper" system. From the mechanical point of view, this "spring and damper" system has a lot in common with the myeloelastic-aerodynamic behavior of vocal folds. Thus, resulting stick-slip sequences should imply anticausal oscillations. This assumption can also be verified by measuring relative bow-string velocity[1], as in Figure 5.1b.

[1] The relative bow-string is the speed of the string measured with the bow as the reference. It can be seen as the derivative of the displacement between the bow and the string [171].

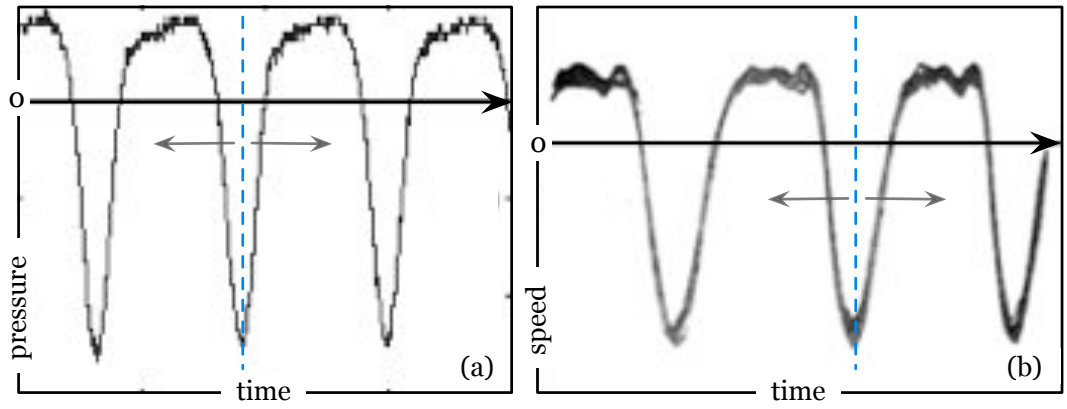

Figure 5.1: Pressure at the mouthpiece of a trombone (a) and relative string-bow speed for violin (b) [47], revealing some causal (right arrows) and anticausal (left arrows) components around a kind of "closure instant" (blue dashed).

5.3 Mixed-phase analysis of instrumental sounds

In order to evaluate decomposition possibilities on typical CII sounds, a large amount of recordings have been collected, targeting two instruments: trumpet and violin. Trumpet sounds were recorded in TCTS Lab. Recording equipment and conditions were formalized. Sound production techniques (e.g. pressing, rounding relaxing the mouth) were commented by the player to allow us to emphasize eventual correlations. The diversity of embouchure techniques was the target. Violin sounds are part of the database from Iowa University Electronic Music Studios [57]. This database contains 89 sounds (single note), all recorded and analysed in CD quality: 16bits/44100Hz.

For each of these sound files, the ZZT-based decomposition module of the RAMCESS analysis framework were used. Anticausal and causal components were computed, and magnitude spectra were correlated with playing techniques, as described by the player.

5.3.1 Trumpet: effect of embouchure

Two kinds of trumpet sounds were analysed. The first one is identified by the player as a *lax* (also *opened*, *round*) production mechanism, the second one as *pressed* (also *closed*, *thin*). A frame is selected in each sound and results of the decomposition are presented in Figure 5.2.

Anticausal and causal contributions show similarities with typical speech decomposition results. Indeed anticausal waveforms look like truncated unstable oscillations, as de-

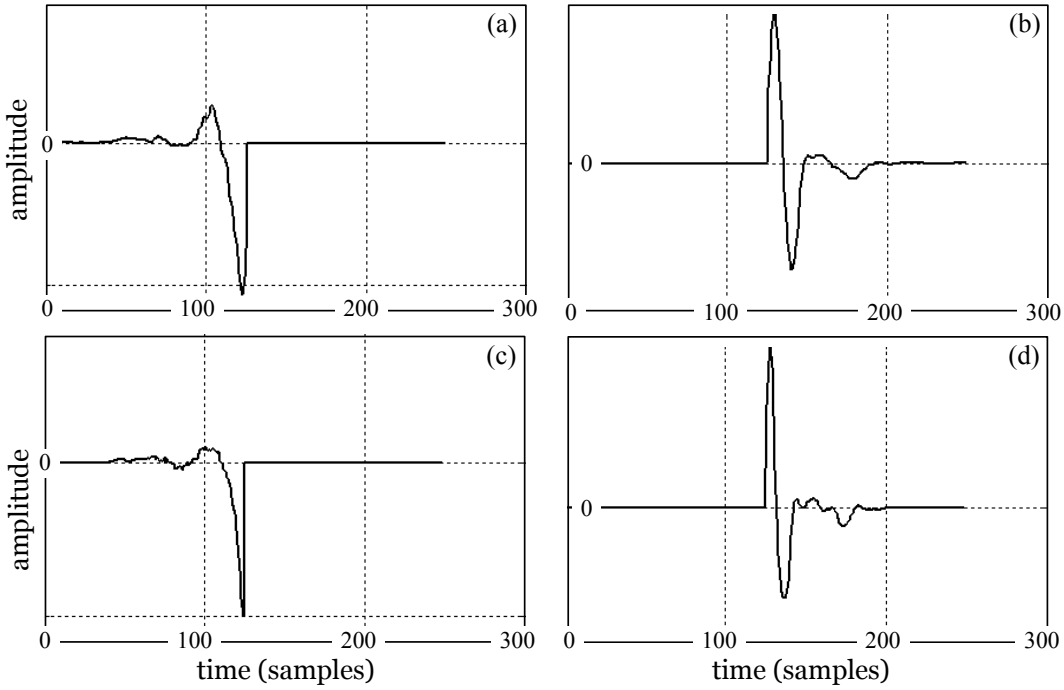

Figure 5.2: Diagrams (a) and (c) show anticausal parts, diagrams (b) and (d) show causal parts obtained by ZZT decomposition of two trumpet sounds: *lax* (top) and *pressed* (bottom) sounds.

scribed in [65]. The same way, causal parts can be interpreted as the impulse response of a linear minimum-phase filter, at least as a first approximation.

The difference between the two kinds of production is more obvious in the spectral domain. In Figure 5.3, spectral envelopes of the above mentioned signals (lax/pressed decompositions) are presented. We can see that stressed production is characterized by a shift of the anticausal formant[2] to higher frequencies. In the causal part, we can see more energy in high frequencies for the stressed sound, while the causal formant remains at the same position as for the lax sound.

5.3.2 Trumpet: effect of intensity

A longer sound of trumpet, corresponding to a continuous timbre modification, has also been analysed. The player was asked to produce an increasing-decreasing intensity. In order to emphasize the spectral impact of this performance, two pitch-synchronous

[2] Using "formant" in this context, we are generalizing terms coming from speech processing: glottal formant and vocal tract formants. The anticausal formant mentioned in this Chapter does not result from the same acoustical phenomena than glottal or vocal tract formants.

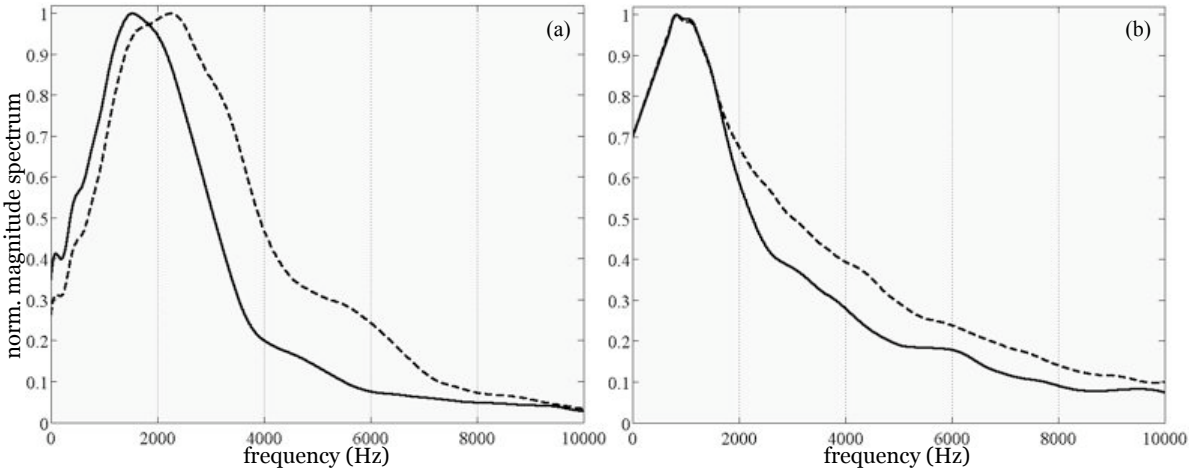

Figure 5.3: Spectral envelopes of anticausal (a) and causal (b) contributions, for trumpet sound production with lax (solid) and pressed (dashed) embouchure.

spectrograms are computed. They show the evolution of the magnitude spectrum of both anticausal (Figure 5.4a) and causal parts (Figure 5.4b) of the sound.

These spectrograms illustrate that the increasing intensity performed by the player provokes a displacement of both anticausal and causal formants to the higher frequencies. In the context of our mixed-phase approach, the typical *brassy* effect – clearly remarkable when trumpet is played loud – can be precisely characterized by movement of anticausal and causal resonances. These spectral movements should obviously be further examined, in relation with perception.

5.3.3 Violin: proof of concept

As for the trumpet, the sound of a violin can be decomposed by the ZZT-based processing. It also shows some similarities with speech decomposition. Anyway, as we could not collect a large and adapted expressive database for this instrument, we only validate the method. Correlations between decomposed signals and bowing techniques are planned as further work, with the target of finding a player able to comment his/her bowing techniques. Results of the decomposition of a violin frame are presented in Figure 5.5.

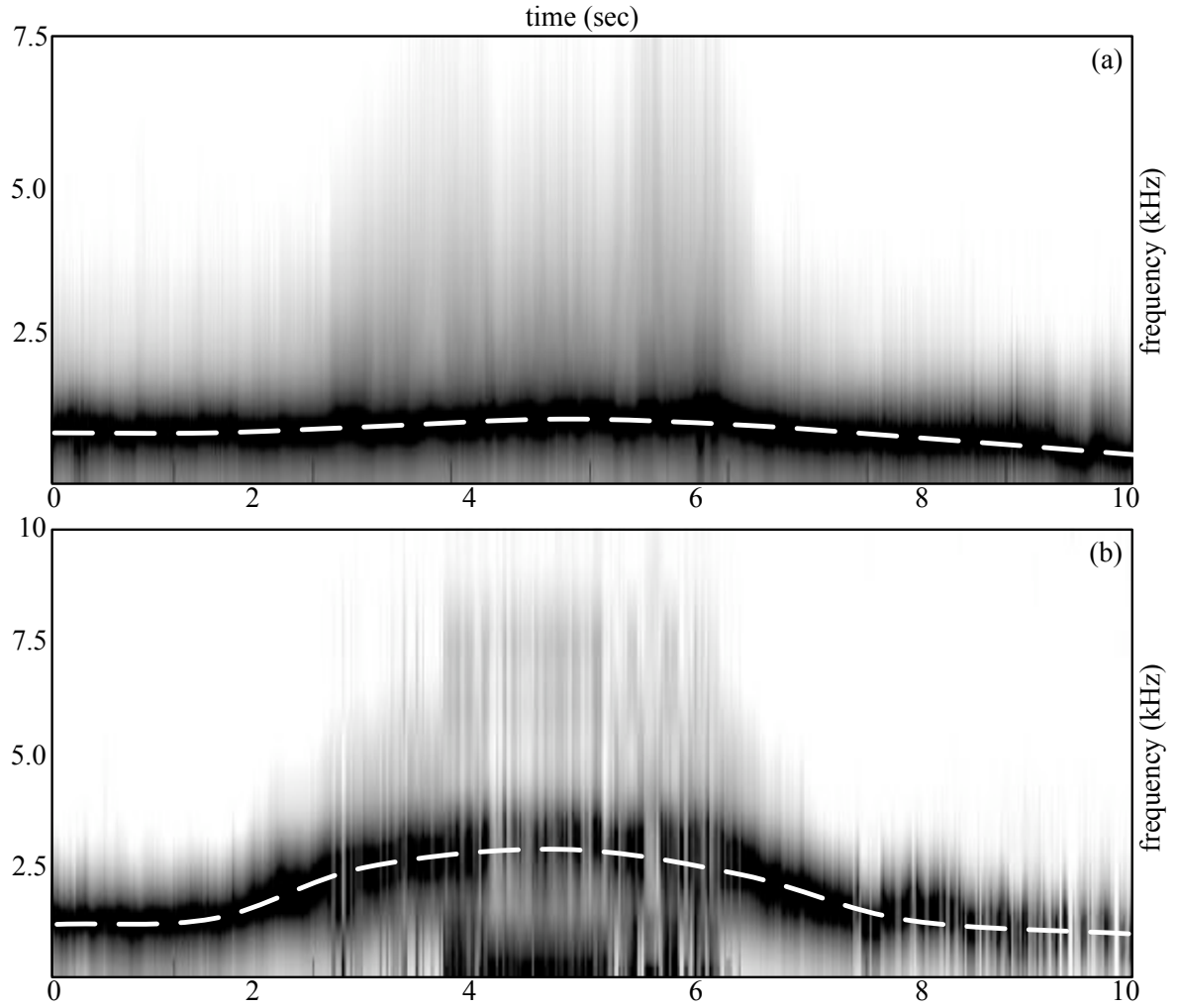

Figure 5.4: Normalized spectrograms of anticausal (a) and causal (b) contributions of a trumpet sound corresponding to an increasing-decreasing intensity.

5.4 Mixed-phase synthesis of instrumental sounds

ZZT-based decomposition demonstrated that typical CII sounds could be represented as the convolution of anticausal and causal contributions. Moreover, correlations with embouchure techniques and intensity have been highlighted for trumpet. Aside with a use in music information retrieval (MIR), these results also lead us to consider that mixed-phase representation of CII sounds is particularly relevant for expressive synthesis.

We propose a new subtractive synthesis technique based on mixed-phase representation of CII waveforms. The original idea is to consider the anticausal signal as the source, and the causal signal as the filter impulse response. We can show that the convolution of anticausal and causal components brings back the original signal [26]. This convolution

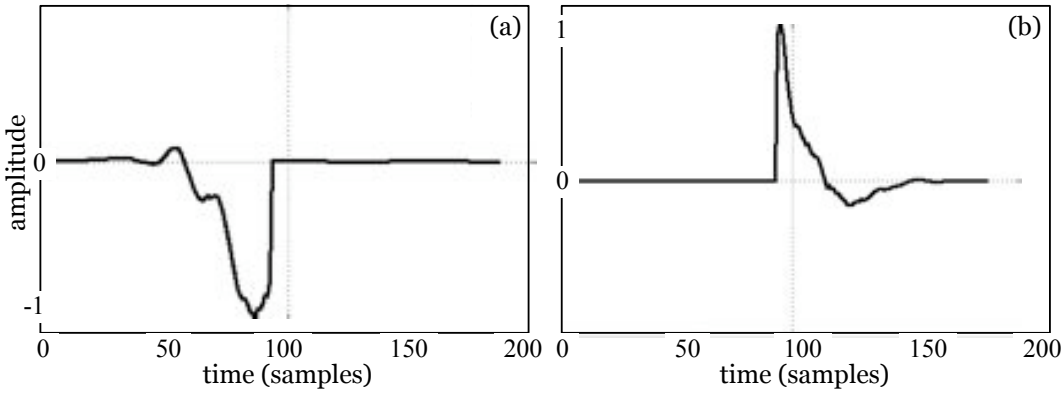

Figure 5.5: Decomposition of a violin sound into its anticausal (a) and causal (b) components.

is illustrated in Figure 5.6a. The difference is due to errors in the computation of the roots of a large polynomial, for computing the ZZT of the signal.

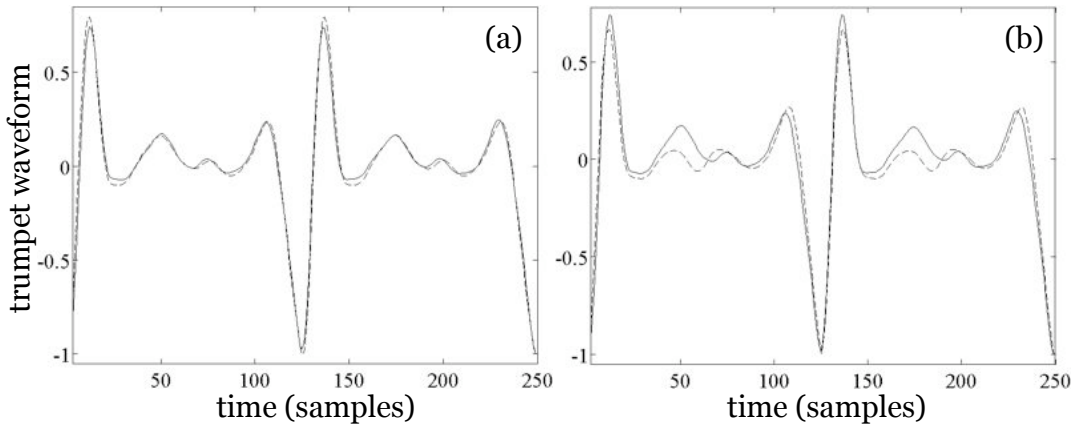

Figure 5.6: Comparison of the original trumpet sound (solid) with (a) the convolution of decomposed components, and (b) the resynthesis based on all-pole spectral models of both anticausal and causal parts (dashed).

Based on this assumption, both anticausal and causal components can be approximated by linear filter impulse responses, introducing two spectral models: $H_a(z)$ for the anticausal part, $H_c(z)$ for the causal part. In order to preserve phase information, the filter representing anticausal component has to be anticausal itself [65]. In this case, we use the causal version of $H_a(z)$ and reverse its impulse response. Figure 5.6b compares the original trumpet signal with results of this process, where filter coefficients have been estimated by LP analysis of both anticausal and causal parts, with $p = 46$.

5.5 Conclusions

In this Chapter, we have presented an efficient framework in order to analyse causal and anticausal components of typical continuous interaction instrument (CII) waveforms.

First we causality of sounds produced by woodwinds and bowed string instruments has been discussed, showing that these acoustical mechanisms exhibit some anticausal components: movement of the lips at the mouthpiece, and bow-string interaction.

Then the main analysis algorithm has been described: the separation of causal and anticausal contributions based on zeros of the Z-Transform (ZZT) of CII signal pitch-synchronous frames. We have shown that a "closure instant" could be found, and that the decomposition led to similar results than with voice analysis.

Decomposition results for trumpet and violin sounds have been discussed. They allowed us to establish interesting correlations between embouchure techniques (pressed, round, open, etc) playing intensity for trumpet (the so-called brassy effect) and the movement of causal and anticausal resonances on the spectrograms.

Finally the decomposition results led us to propose a generalized causal/anticausal linear model for synthesis of CII waveforms in spectral domain. We have shown that with typical LP order, the waveform of the trumpet sound could be resynthesized with an good quality, leading to new possibilities for mixed-phase synthesis of CII sounds.

Chapter 6

Analysis-by-Interaction: Context and Motivations

> *" People who learn to control their inner experience will be able to determine the quality of their lives. "*
> — Mihály Csikszentmihályi

6.1 Introduction

Chapters 3 and 4 highlight the main axes that define the technical structure of this thesis. These ideas are structured in the common pipeline of analysis/resynthesis: proposing segmentation and analysis techniques, extracting model-based parameters and finally building realtime synthesis. This mind map is widely used in sound synthesis research.

In the context of digital musical instrument making, it would appear straightforward to pursue this step-by-step process by presenting an appropriate controller and describing mapping strategies, i.e. how controller dimensions are connected to sound synthesis parameters. This task would appear as one more block in the whole process.

This typical data processing pipeline (inputs, outputs, blocks) particularly fits the widely used *digital musical instrument model* [199], as described in Figure 6.1. The user achieves gestures on a controller. These stimuli are then mapped to the synthesis parameters

through a cloud of conversion equations, that can be based on physical, perceptual or physiological assumptions [11, 179]. In this representation, the performer receives two feedbacks, one haptic F_h (touching the object), and one acoustic F_a (hearing the sound).

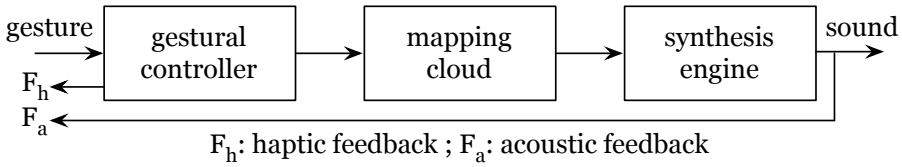

Figure 6.1: Description of the digital musical instrument model: gestures are achieved on a gestural controller, these stimuli are mapped to sound synthesis parameters. The user receives two feedbacks: one haptic F_h and another acoustic F_a.

This model is particularly suitable for the representation of data flows. However it implicitly leads to consider that gestural control, mapping and synthesis correspond to three different areas that should be optimized separately, by addressing different challenges. The model shifts from a representative purpose to a methodological purpose, which leads to split the scientific design of digital musical instruments into different research topics: signal processing, human/computer interaction (HCI), music analysis, etc. This view ignores the complexity of the instrument maker/musician relation.

Our work aims at preserving these specific properties, through a closer understanding of luthery, and it results in several assumptions that are presented in this Chapter:

- The dislocation of the instrument making activity – seen from the point of view of the artisanal practice – into separated topic-related tasks does not necessarily decrease the complexity of the design, as it is expected when a complex problem is split into simpler issues. On the contrary, complexity and abstraction increase.
- The relegation of the musical practice as a testing/benchmarking step in the process does not take the whole benefit of the instrument maker/musician relation. This typical roadmap forgets that practicing an instrument is often much more a source of innovation by itself than strictly a validation process.

In Section 6.2 we give more details on sound synthesis (more precisely on voice synthesis) and on human/computer interaction assessment. We also highlight problems encountered in the cross-validation of multimodal systems. Then Section 6.3 presents the theoretical basis that leads our approach: intimacy and embodiment, and it results

in an integrated methodology, called the *Luthery Model* (LM). Finally Section 6.5 describes the *Analysis-by-Interaction* (AbI), which aims at using the LM in analysis for sound production understanding. This property is further examined in Chapter 8, with the help of the instrument described in Chapter 7: the HandSketch.

6.2 Prototyping digital musical instruments

A digital musical instrument is a multimodal device, aiming at converting gestural control inputs into sounding outputs. The digital musical instrument model, illustrated in Figure 6.1, gives a clear representation of the data flow. This three-block diagram usually leads to a separated optimization of the different modules [199].

The assumption that such a problem has to be split along pre-defined research areas is not really discussed. In traditional instrument making, we see that the splitting strategy preserves some interdisciplinarity at each step. Indeed the setting of a string, or the shaping of a body, equally questions acoustics, haptics, mechanics or aesthetics [176].

In this Section, we give more details about optimization strategies that are used in existing topics. More precisely we particularize this description of optimization strategies to the validation of voice synthesis engines, and human/computer interaction devices. Then we argue on the fact that the combination of both leads to complexity, abstraction, and finally to arbitrary decisions.

6.2.1 Validation of voice synthesis engines

The recommendations provided to competitors in the well-know Blizzard Challenge [21] mention a quite limited set of criteria to evaluate the quality of a voice synthesis engine. For example, properties such as intelligibility and naturalness are recurrently highlighted. The measurement of naturalness and intelligibility of a single voice is often evaluated by a *Mean Opinion Score* (MOS) [184].

When we address voice transformation issues or expressive speech synthesis, it is common to compare synthetic samples with a given target. Once stimuli have to be compared, ABX or AB-BA [71] tests can be implemented. Not to mention the wide use of *Analysis of Variance* (ANOVA) in the context of emotional speech assessment [30].

However all these techniques are based on the same idea. Starting from the analysis of data, a model is proposed. Then a voice synthesis engine is designed. This engine is launched many times in order to produce a given amount of stimuli. Then these stimuli are organized into listening sessions, and participants are asked to evaluate various aspects of the resulting speech: intelligibility, naturalness, likelihood, etc [22].

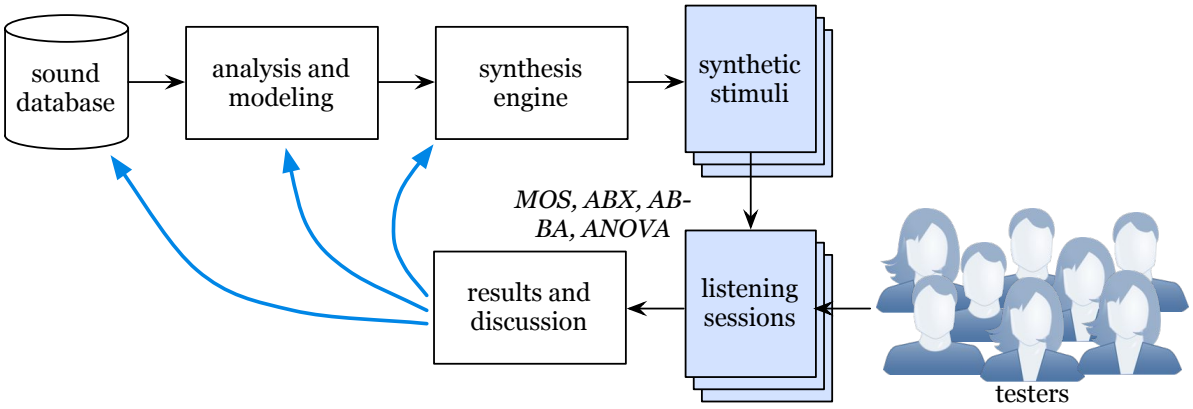

Figure 6.2: Illustration of the validation of a voice synthesis engine: resulting from data analysis, and modeling, the synthesis engine is launched for generating stimuli. Then these stimuli are rated by participants and results are discussed within some interpretation techniques. The process is repeated with next assumptions.

Figure 6.2 illustrates this general view on voice synthesis assessment. The main aspect that we want to highlight is the iterative multi-subjective property of this process. Indeed the synthesizer is progressively improved by following the recommendations of successive testing populations. The large amount of testers guaranties the average coherency.

6.2.2 Validation of HCI devices

Human/computer interaction (HCI) studies how to adequately establish a connection between a humain being and the computer. It is essentially a reflection around interaction devices, their evaluation and their potential. As described in [198], it includes the definition of representative tasks in order to compare devices [31], the proposition of models of aimed movements [91, 90] and the classification of input devices [36].

HCI also suggests a methodological context for managing innovation in interactive applications. Iterating seems to be a native concept in the design of HCI prototypes. We find these ideas in theories such as the HCI spiral [138] or the user-centered iterative design [144]. Usually a sequence of four steps is followed:

- defining requirements: studying the situation and formulating the needs of users;
- choosing the design: choosing the architecture in order to best fit requirements;
- implementing the prototype: integrating hardware and software modules;
- evaluating the prototype: proposing the device to a testing population and assess its behavior along several axes: reliability, ergonomics, usability [1, 38].

These steps can be represented as a spiral, because each revolution along the four steps leads to a new situation. It creates as an iterative movement away from the starting point (cf. Figure 6.3).

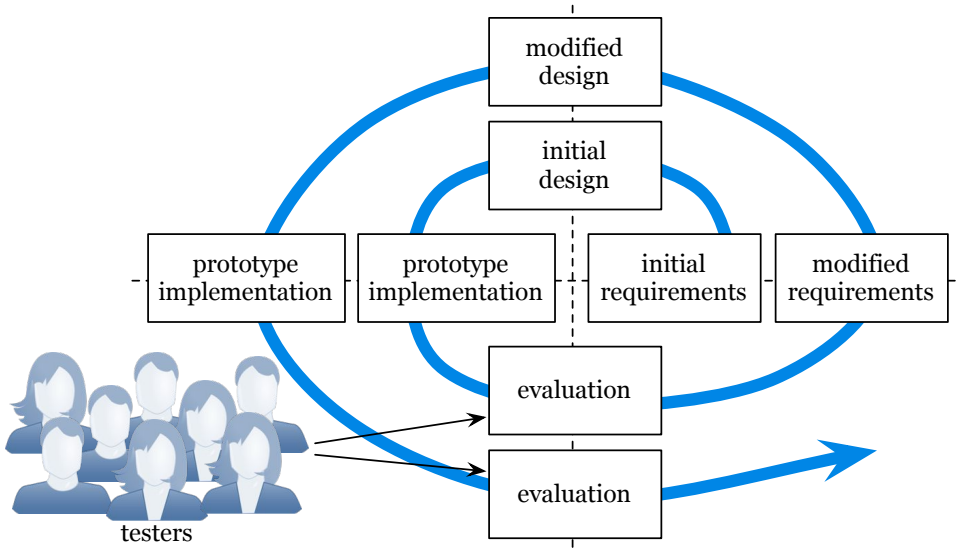

Figure 6.3: Iteration in the validation of human/computer interaction devices: setting requirements, defining a design, implementing a prototype and evaluating the prototype with the help of a testing population.

6.2.3 DMI: the multimodal case study

Recent discussions about the integration of multimodal applications often present the sound synthesis engine as one module imbricated in the HCI optimization process [146]. This assumption is particularly compatible with the digital musical instrument model. Indeed it suggests that the imbrication is achieved by choosing an appropriate mapping.

This workflow is particularly appropriate in the study of typical functional interaction problems: querying information, manipulating data, medical assistance [107], etc.

Unfortunately, addressing expressive interaction issues is essentially seen today as an extension of existing practices, both in sound synthesis and HCI.

Considering that expressive interaction is a generalization of usual functional interaction can be advantageous for research. Indeed extending tools always appears to be more efficient than reinventing them. However it is relevant only within the functional interaction context. For instance, developing the emotional awareness of a talking machine is useful, but it remains a typical functional problem of querying information.

What makes digital instrument making an interesting case study is that, in the context of musical performance, the interaction is significantly centered on the emotional behavior, and somehow less functional. Consequently, the strategy aiming at extending the existing functional prototyping with emotion-related knowledge does not really work.

The state of the art in the understanding of emotional interaction does not benefits from any clear consensus yet [15]. Both models and methodologies are actively discussed. Within a context that is probably totally different from the functional interaction, answers delivered by existing tools are complex, abstract and sometimes inappropriate.

For instance, minimizing the Meyer's law [160] – a relationship describing the structure of a gesture by the number of its sub-movements – is not necessary "good for musical performance". It seems clear that the violin would probably never succeed such a HCI assessment. And it is interesting to highlight that most of the exciting instruments produced during the last decade come from idiosyncratic approaches, i.e. from approaches that are peculiar to an individual, without any generalized procedure.

Building an expressive object relies on underlying emotional mechanisms that are not clear enough today to deduce a systematic approach. Facing this problem requires to restart an initiative from the ground, closer to traditional instrument making principles.

6.3 Intimacy and embodiment

There are many ways of defining the skillfulness of a human in manipulating an object. As it has been described in Section 6.2 the HCI framework proposes an evaluation of this ability by measuring the duration and the amount of movements required to perform representative tasks, and iterates for minimizing these cost functions.

Our work is inspired by another approach. Moore proposes that "the best musical instruments are the ones that exhibit an important quality called *intimacy*" [142]. Intimacy is the degree of fitting between the behavior of someone's body and the object. The more the performer and the instrument develop an intimate relation, the more the performer intuitively transforms desirable sounds into psychophysiological controls.

This property can also be perceived as *embodiment*: an intimate human/object interaction reveals the way the object has been integrated into the behavior of the human's body, like an extension. The degree of intimacy/embodiment depends on many factors, as described by Fels [79], but mainly depends on the type of interaction.

6.3.1 The four types of interaction

The degree of embodiment depends on the context, and mainly on the type of interaction that is involved in a given interactive application. Fels proposes a classification in four different types of interaction [79], and discusses their influence on expressivity:

- The person communicates with the object in a dialogue. This is the functional type of interaction. The person controls the device, and the device communicates back, in a sequence of clear causes and consequences.
- The person embodies the object. The person integrates the object and its behavior into his/her own sense of self, as a part the body. This is clearly the case which best corresponds to musical performance.
- The object communicates with the person. This is the passive and contemplative interaction mode, like when looking at a painting. The object delivers some information and the person receives it. There is no interaction.
- The object embodies the person. This relates to recent cases of immersive media installations where the image and/or sound of the visitor is used as input material. The person can interact with the object within this feeding strategy.

6.3.2 Expression and embodiment in musical performance

The situation of musical performance generally corresponds to the second interaction mode: the person embodies the object. Fels argues that a lack of intimacy leads to poor

communication abilities. In contrast, a high level of intimacy allows to communicate complex ideas and emotions through the psychophysiological control.

Coming back on the definition of expressivity that has been proposed in the Introduction of this work, the possibility of expressing oneself in altering a formal language requires a high level of understanding and practice of this language. Thus intimacy and embodiment are the results of a long learning process and appropriate conditions.

6.4 The Luthery Model: optimization based on intimate assessment

In this work, we aim at proposing a workflow that emphasizes the embodiment, and thus we try to fulfill the two main conditions of its achievement: long term learning and appropriate conditions for continuous learning. In a word: we try to get closer to traditional instrument making. This is why we call this strategy the Luthery Model (LM) of digital instrument making.

The intimate relation that we expect to create requires a long-time involvement. However inventing a new instrument inherently leads to a paradox:

- What can we do during ten years if we have to wait for skilled practice ?
- What can we practice during ten years if we do not have any instrument ?

The research in digital instrument making partially answered this question in trying to borrow existing musical practice. It concerns the whole field of augmented instruments or "instrumentoïds" (violin-like, guitar-like, woodwind-like, etc) [125].

When developing new instrumental paradigms, we cannot benefit from any existing skilled practice. In order to unlock the above-mentioned paradox, we have to merge practicing and prototyping into the same activity.

Several conditions have to be verified, in order to establish such a parallel workflow:

- the initial idea for the musical instrument has to lead directly to the possibility of practicing it; thus the first prototype should be simple, focused and efficient;

- we consider that the HCI spiral remains relevant in this context, but each step exhibits much more inertia: requirements, design and implementation evolve slowly, in order to preserve the increasing playing skills of the performer;
- in order to compensate for this slow progression between consecutive steps, the amount of iteration increases with time; the intensification of this communication builds the particular instrument maker/musician relation;

Following these guidelines leads to the Luthery Model, a reorganization of the common prototyping scheme into a methodology focused on embodiment. After a few years of iterating in this framework, the shape of the instrument reaches a stable state, and the degree of skillfulness associated with this instrument are acknowledged to be high level.

One other significant result of the LM is that the prototyping steps – inspired by the HCI spiral – start to merge. Indeed the evolution of the instrument becomes a one-block strategy where requirements, design, implementation and testing all happen simultaneously. This is why Figure 6.4 represents the LM strategy as a converging spiral.

This aspect probably also explains the great success of idiosyncratic approaches.

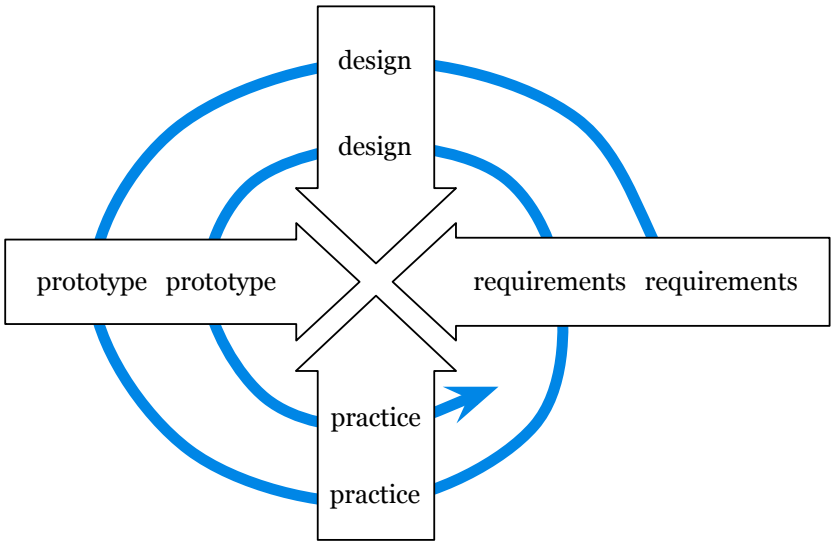

Figure 6.4: Iterative prototyping spiral for an HCI device, reinterpreted within the Luthery Model. Each activity (requirements, design, prototype and practice) has its own internal evolution, and the whole process converges into one integrated strategy.

6.5 Analysis-by-Interaction

One main advantage of the LM is that the instrument stability and the high level practice are reached simultaneously after some years of practice and design. The performer owns this intuitive "know how" of the instrument and can use it for many purposes.

Better still, such a performer/instrument embodiment can be used for analyzing the performance, for research purposes. Indeed it can be used to propose practice-based models for the production of different signals, such as those related to emotional interaction.

Once the studied signal can convincingly be imitated by a performer playing the dedicated instrument, the understanding of this phenomenon can be approached from the gestural point of view. Indeed the digital controller is totally accessible and each particularity of mimicking gestures can be measured and analyzed precisely.

This property of the LM leads to a new framework, which can be applied in a much wider context than in musical performance alone. We call this method *Analysis-by-Interaction* (AbI), in which the analysis of a given signal is extended by the use of an appropriate interactive application and a skilled performer.

6.6 Conclusions

In this Chapter, we have highlighted that the methodology used in this thesis was not following typical analysis/synthesis/control pipeline that is suggested by the digital musical instrument model. Our work has been structured differently, probably closer to the traditional musical instrument making process.

We have described the typical validation processes used separately in prototyping of both sound synthesis engines and human/computer interaction devices. Then we examined the main drawbacks of this dislocated approach and proposed some recommendations, which led us to define the Luthery Model (LM).

Using the encouraging results of the LM, we introduced a new approach in signal analysis, called Analysis-by-Interaction. This new methodology provides an alternate way of analyzing some unaccessible signals, by imitating them with an appropriate digital musical instrument. This idea relies on the long term practice of this instrument in

order to reach a particularly convincing imitation. Finally the imitated signal is studied through the analysis of imitative gestures on the instrument.

In this thesis, we use AbI in the context of expressive voice understanding, and particularly for high quality singing synthesis. Our use of the LM leads us to develop the HandSketch, a tablet-based digital instrument explained in Chapter 7. Then this instrument, and more precisely the analysis of performing gestures, are used in order to propose a new model for the vibrato in singing, in Chapter 8.

Chapter 7

HandSketch: Bi-Manual Control of Voice Quality Dimensions

" I was really interested in touching the sounds. "
— Michel Waisvisz (1949–2008)

7.1 Introduction

Following the recommendations that have been presented in Chapter 6, we describe, in this Chapter, the development of a new musical instrument, called the HandSketch. This new instrument is developed in respect with the Luthery Model (cf. Section 6.4). It means that this instrument has to be practicable right from the beginning of the prototyping, in order to allow the progressive embodiment of the object. This specificity leads us to focus our control paradigm on a particularly embodied skill: the writing.

The HandSketch is a digital instrument made for the bi-manual control of voice quality dimensions: pitch, instensity, glottal flow parameters [54]. It is made of purshasable devices: a pen tablet and force sensing resistors (FSRs). More precisely it is built around a Wacom^{TM} graphic tablet [196], played vertically along the upper part of the body. The HandSketch uses a particular polar transformation of the control space in order to fit the requirements of ther prefered hand. A sensing strategy inspired by woodwind and string instruments is adapted to FSRs for the use of the non-prefered hand. It is

important to highlight that the instrument evolved in nine consecutive versions – being now called HS1.8 – and thus reached a more stable shape and behavior. The most recent playing situation (controller and attitude) is illustrated in Figure 7.1.

Figure 7.1: Typical playing position when performing the HandSketch in 2009: sitting down, arms and hands surounding the controller. This setup also have the particularity of using a headset microphone, as a way of inputing realtime voice.

In this chapter we first propose a discussion on the pen-based control of music (Section 7.2). In the same Section we continue by addressing a serie of issues related to the improvement of pen-based gestures. In Section 7.3 we describe choices that have been made concerning the non-prefered hand. Finally a significant part of this chapter is devoted, in Section 7.4, to discussing the long-term practice of this instrument and its influence on expressivity.

We also want to notice that the HandSketch project does not attempt to "prove" any superiority or relevance, compared to the wide instrument making community. As it has been shown in Chapter 6, the assessment of a musical instrument remains an open proble. We can argue that a systematic approach is used in order to define our control strategies, but we can not totally pretend that this instrument does not rely on any idiosyncratic idea. The relevance of this instrument is rather justified by its ability to achieve some AbI protocols, as it is described in Chapter 8.

7.2 Pen-based musical control

Graphic tablets, which are initially developed and sold to meet the needs of image professionals (designers, architects, editors, etc.), can today be considered as a common device in computer music. They have actually been used since the 70's, for example in the Xenakis' UPIC system [137]. More recently the compositional and scientific work of Wright [200], Momeni [140] or Kessous [10] are considered as significant.

Figure 7.2: Two video archives. On the left, I. Xenakis playing on the control surface of the UPIC system (1987). On the right, M. Wright doing timeline-scrubbing with a realtime sinusoidal model (2006) on a Wacom^{TM} tablet.

Today we can observe an unanimous use of Wacom^{TM} products. Indeed most of the models provide a large number of parameters, with high precision and low latency, structured around our intuitive writing abilities. For instance a professional model sends values for the x axis, in a range of 0 to 65535 (16bits), with a samplerate of about 100Hz. These properties make tablets really good candidates to fit the Hunt and Kirk's *real-time multi-parametric control system* criteria[1] [106, 199]. The availability of many softwares which bridge the Wacom^{TM} parameters, through OSC or with a direct plugin, such as Max/MSP, also contributes to the wide dissemination of the controller.

In this Section, we present our work in the mapping of pen-based gestures with attributes of the vocal expressivity: pitch, loudness and voice quality. First we describe the early tablet-based prototype, called RealtimeCALM (in 7.2.1). Then we give some motivations in the use of pen-based gestures for the precise control of pitch (cf. 7.2.2). Finally we propose some improvements in the ergonomy of the tablet playing (cf. 7.2.3).

[1] It is also intersting to highlight that these performances are far beyond what MIDI can propose.

7.2.1 First prototyping with RealtimeCALM

In the early years of this thesis, there has been quite a lot of emulation in the design of a controller which aimed at manipulating the voice quality dimensions of the RealtimeCALM synthesizer [52]. In this early work we proposed and demonstrated two instruments, one of which already used the tablet – an A6 Wacom™ Graphire™ – as the main device for achieving expressive vocal sounds. That insight happened after an extensive use of the glove as a speech synthesis controller – following what Fels did with GloveTalk [80] and GRASSP [154] – as way of moving to the production of singing.

In our first prototype, the horizontal axis of the tablet x is mapped to the fundamental frequency. Concrete performative situations – typically improvisation – show that 2 or 3 octaves can be managed on a A5/A6 tablet, after some musical training. Vertical axis of the tablet y controls the voice quality, with the use of the "presfort" dimension that has been described in Section 4.4.1. Finally the pressure on the tablet p controls the loudness of the sound, through the modification of E, the amplitude of the GFD negative peak. This mapping is illustrated in Figure 7.3.

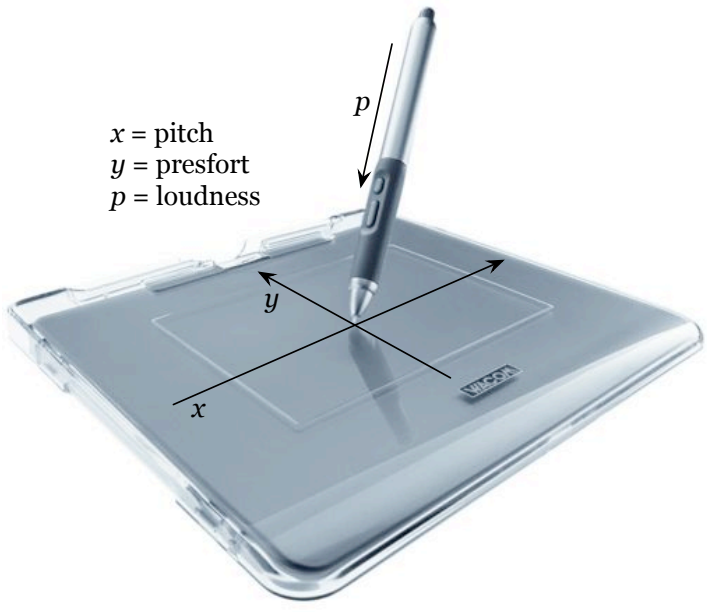

Figure 7.3: Mapping used in the RealtimeCALM system: x controls the fundamental frequency, y is mapped to the "presfort" axis, and p controls the loudness.

7.2.2 Pen-based gestures and fundamental frequency

Prosody and melody play a strategic role in expressive speech and singing production, respectively. We therefore significantly focus the design of our new controller on the accurate and realtime control of pitched sounds. Surprisingly, there is not much literature on pen-based continuous *pitch* and *intensity* gestures, as opposed of course to that of continuous pitch acoustical instruments, like the violin [203], but to that of some electrical devices, like the theremin [174].

The HandSketch can be seen as a new digital case of fretless playing, known to be difficult but powerful. One of the most advanced formalization concerns the helicoidal representation of notes in the Voicer [116], involving the well known Shepard circularity in perception of fundamental frequency [175]. Let also mention the Kyma [124] initiative, which developed a great framework for Wacom™ control of sound synthesis, but without formally considering (*pitch*, *intensity*) issues. In this work, we aim at formalizing the pen-based interaction, essentially by the solving of ergonomic problems.

7.2.3 Solving ergonomic issues

In this Section, we introduce a particular framework for expressive pen-based (*pitch*, *intensity*) musical gestures. This structure is much more based on ergonomic issues and on their impact on sound synthesis, than on psychoacoustic representations. Our approach considers that natural pen movements are mainly forearm- and wrist-centered soft curves (cf. Figure 7.4), easier to perform than lines [52] or complete circles [117]. Then come finger movements which have a refinement purpose.

Therefore we define a strategy in which pitch information results from a transformation of $(x,\ y)$ cartesian coordinates into polar coordinates, but where the the center of the circle position $(x_C,\ y_C)$ is tweakable, in order to fit forearm and wrist circular movements. Typically this center will clearly be out of the drawing surface, close to tablet border, where the forearm is supported. This concept is part of the playing diagram that is visible on Figures 7.5 and 7.6. The conversion is presented in equations 7.1 and 7.2.

$$R = \sqrt{(x - x_C)^2 + (y - y_C)^2} \tag{7.1}$$

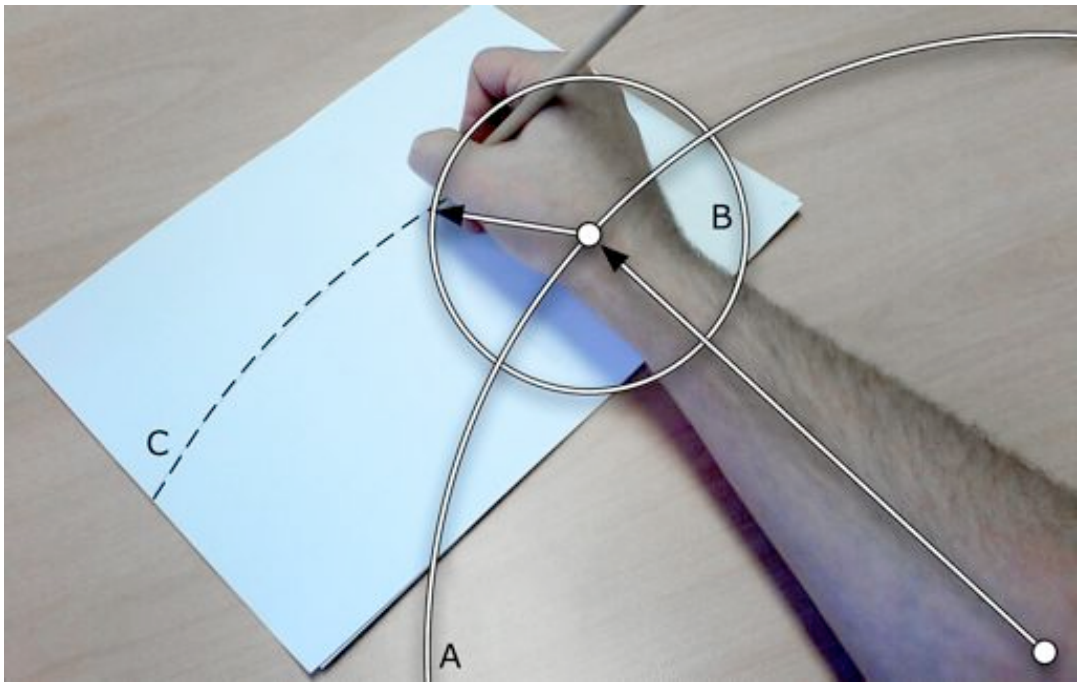

Figure 7.4: Pen drawing soft natural curve (C) on a surface. It can be seen as a mechanical combination of forearm- (A) and wrist-centered (B) movements.

$$\theta = arctan(\frac{y - y_C}{x - x_C}) \tag{7.2}$$

with R and θ respectively the radius and the angle of the (x, y) point, measured in polar coordinates, with the center localized in (x_C, y_C), instead of (0, 0). In Figure 7.4, we show the decomposition of the circular movement, from wrist and arm submovements. The resulting curve is C, here supposed also circular. (x_C, y_C) is considered as the center of this circle C, achieved for a particular value of R.

Mapping of the angle

As pitch control is now related to θ, angular information will be normalized and modified in order to lay out a range of notes in which every semitone (in tempered scale) corresponds to the same angle. Then an arbitrary parameter to set is the number of octaves that are mapped on the whole angle variation (typically between 2 and 4). The conversion is obtained with equations 7.3 and 7.4.

$$f_0 = f_{0_R} \times 2^{\frac{i}{12}} \tag{7.3}$$

$$i = N \times 12 \times \frac{\theta - \theta_B}{\theta_E - \theta_B} \tag{7.4}$$

where N is the number of octaves we want on the playing surface, θ_B is the leftmost angle visible on the playing surface, θ_E is the rightmost angle visible on the playing surface and f_{0_R} is the reference frequency corresponding to the θ_B position. A typical pitch modification on this diagram is illustrated in Figure 7.5.

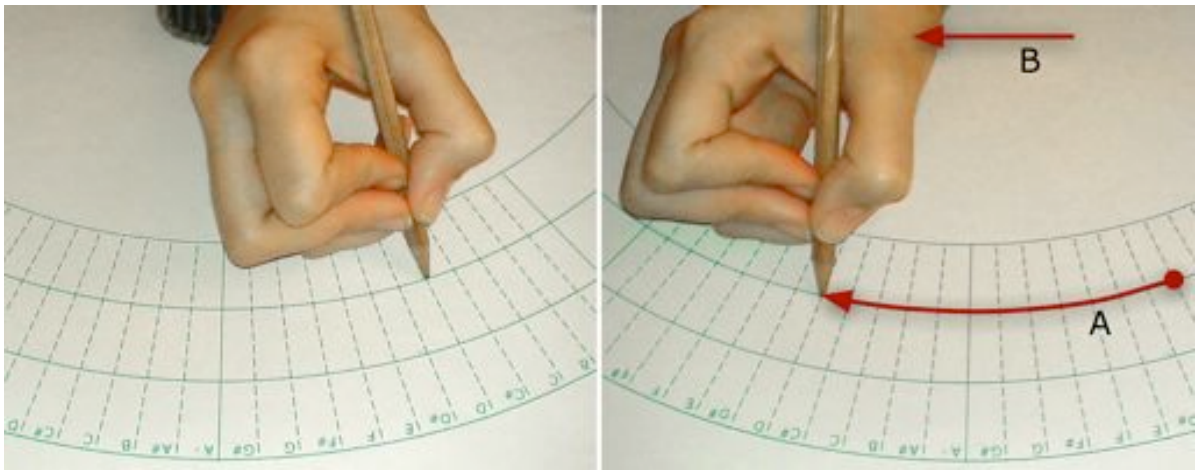

Figure 7.5: Two snapshots (left: before, right: after) in the demonstration of a forearm/wrist movement (B) achieving a simple pitch modification (A).

Mapping of the pressure

Concerning intensity mapping, we decided to keep the same approach as in the RealtimeCALM control model [52], in which sound intensity and stylus pressure were linked. It appears to be relevant, because based on the drawing metaphor, "making sounds" is related to "using pen", and pen is indeed used when pressed on the playing surface. A logarithmic distorsion function can also be added, depending on the sensivity that we want to simulate, while touching the tablet. This add-on is directly inspired by non-linear mappings typically available for MIDI keyboard velocity.

Mapping of the radius: interest in finger-based gestures

Some timbre features have to be controlled coherently with (*pitch*, *intensity*) gestures. A typical situation is singing synthesis control. Indeed voice quality inflections often appear synchronously with pitch and intensity modifications, and combined control of these parameters effectively contributes to the expressivity of the resulting sound [53].

Linking radius R with voice quality dimensions leads to curves which are more complex than in Figure 7.5, where R dynamically changes. Nevertheless underlying forearm and wrist movements remain the same as in Figure 7.5, and refined training just consists in integration of finger flexions. A typical mixed modification on the playing diagram is illustrated in Figure 7.6. We can see the wrist movement B, combined with the finger flexions C, resulting in the mixed gesture A.

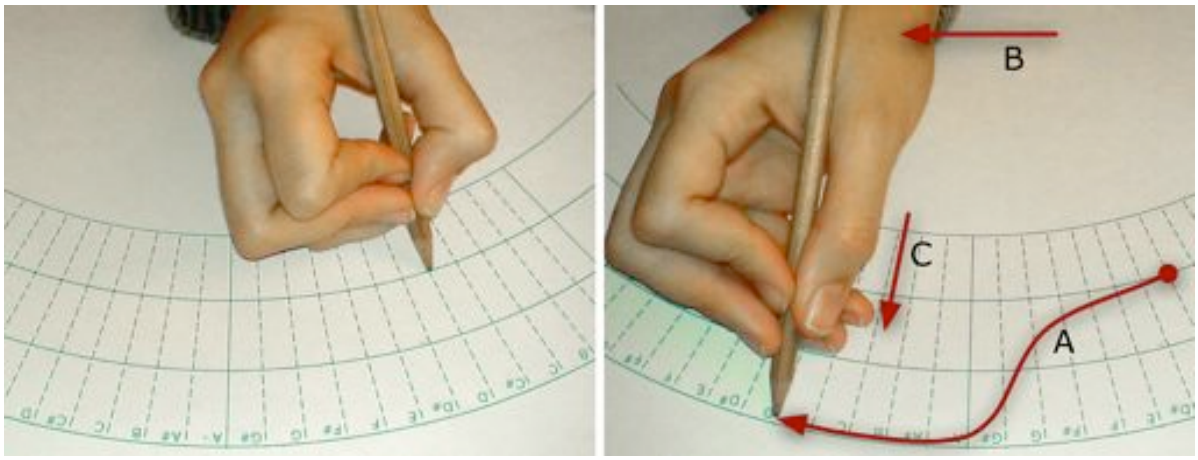

Figure 7.6: Two snapshots (left: before, right: after) in the demonstration of mixed θ and R modification (A) involving both forearm/wrist (B) and fingers (C).

Another interesting aspect of our layout concerns vibrato synthesis. Indeed we know that oscillations do not concern only pitch, but also energy and several spectral parameters [192]. In addition it appears that pen-based vibrato can easily be achieved by little circular movements around a fixed point. In such a gesture, f_0, R and p are all involved in the achievement of the vibrato, which offers good opportunities to develop flexible multi-dimensional mappings around vibrato effects. These issues are extensively discussed in Chapter 8, as a application of Analysis-by-Interaction.

7.3 Non-prefered hand issues

The mapping strategies developed in Section 7.2 proposed some ergonomic improvements, compared to existing tablet-based controls, mainly in pitch and intensity manipulation. Performing on the diagram illustrated in Figures 7.5 and 7.6 makes it possible to learn simple techniques, such as legato or vibrato, in order to reach an interesting level of expressivity for interpretation and improvisation.

However, ore advanced *pitch* and *intensity* structures, like arpeggios, trills, or appoggiaturas are not possible. Moreover even with the large number of parameters transmitted by the stylus, only slow timbre variations can be achieved. We observe that pen-based gestures have a inherent lack in controlling articultions of all kinds.

In this Section, we present a controller for the non-prefered hand, attached to the tablet (cf. Section 7.3.1). Then, in Section 7.3.2, we describe three main challenges that we propose to focus on, resulting in three kinds of gestures that are achieved with this non-prefered hand controller: fretboard, aggregative and articulative controls.

7.3.1 The A+B strategy

Considering the preceding constraints, the use of multiple pressure-sensing surfaces appears to be powerful. In this category, we can find several all-in-one controllers, such as Tactex MTC Express Pad™, Lemur™, or Z-tiles™. We decided to develop an original "on-tablet" shape based on 8 independant FSRs from Infusion Systems™, for technical reasons: portability, unicity, price, latency, and flexibility.

In this configuration, FSRs are separated into 2 groups, A and B. A sensors are aligned to define A thumb positions. In our setup, $A = 3$. B sensors are aligned to achieve four fingers playing techniques. Having one sensor more than the number of available fingers gives particularly creative possibilities, thus we choose a value of $B = 5$. This $5 + 3$ strategy proved to be particulary efficient when playing the instrument. We also want to highlight that this configuration evolved with the instrument, with setups going from $4 + 4$ to $8 + 0$.

A major ergonomic issue of this configuration was to find a comfortable layout. As this problem could not be solved effectively with an horizontal tablet, it has been decided to flip the device vertically, in a position close to accordion playing, as it can be seen in

Figure 7.1. Thus the group of 5 FSRs are placed on the front side, and the group of 3 FSRs on the rear side of the device. It results in the grabbing of the tablet border.

With a longer practice, we can notice that such a movement does not affect the writing abilities required by the prefered hand. Moreover it extends the practice in new directions, as we explain in Section 7.4.1. Figure 7.7 illustrates the front and rear position of the FSR sensors and the way the non-prefered hand interacts with them.

Figure 7.7: Demonstration of front and rear views of a 5+3 playing configuration for the non-prefered hand controller, with a typical hand position.

7.3.2 Non-prefered hand gestures

The $A + B$ strategy is used in order to configure three separate behaviors for the non-prefered hand: fretboard, aggregative and articulative controls. This Section gives an overview of these three mappings. The $5 + 3$ configuration is adapted to the choosing of one of these mappings. Indeed the thumb position (rear panel of the HandSketch) is used in order to select one of those, by pressing on one of the three FSRs.

Fretboard control

This technique is developed in order to allow direct (*pitch*, *intensity*) modifications based on multi-finger playing techniques. It means that a current pitch f_0 is built from the pen-based reference with equations 7.3 and 7.4, then a deviation depending on adopted four fingers position is applied. In the context of singing performance, it can be used to achieve fingering sequences inspired by fretboard playing. A note pointed on the tablet

corresponds to a reference fret on the virtual fretboard. Then pitch can be increased (3 semitones) or lowered (1 semitone), as illustrated in Figure 7.8.

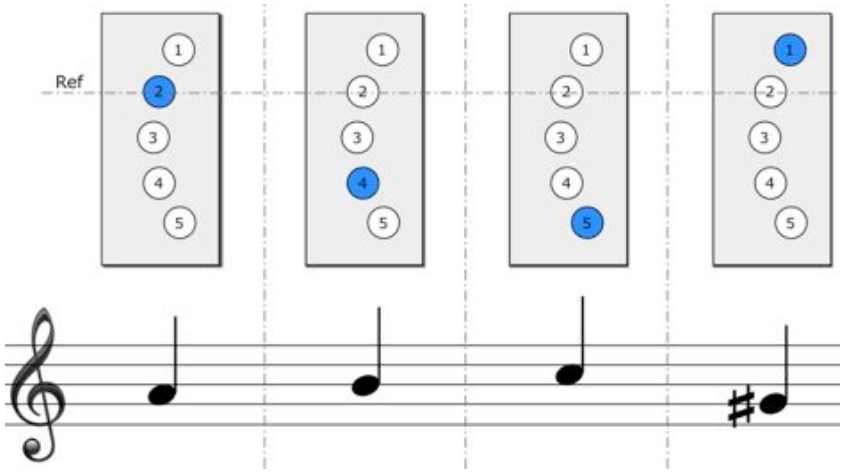

Figure 7.8: Illustration of a non-prefered hand "string-like" playing technique, with captor 2 as the reference fret, corresponding to a $A4$ pointed on the tablet.

Another interesting application is the realtime mimicking of speech intonation. Indeed we know that the f_0 curve in speech can be seen as the combination of a slow ascending/descending slope, plus dynamic inflections synchronized with the syllables [121]. Using the pen for slow slopes and the FSRs for quick inflections is actually really efficient.

Aggregative control

This technique is implemented in order to perform large pen movements, with a structural control on harmonic contents. Thus various finger configurations correspond to pitch and intensity non-linear mappings in a way arpeggios, defined scales or other note articulations can be achieved. Practically the pitch contour is flattened around stable values and the intensity is modulated to sound louder around chosen notes. The amount of this control space distorsion is linked to average FSRs pressure values. This kind of modifications are directly inspired by Steiner's work on the Mapping Library [179].

In Figure 7.9 we observe how the aggregative control modifies pitch and intensity curves. Without any aggregation, pitch and intensity are not modified (green curves). Indeed targeted pitch equals incoming pitch (straight line) and intensity stays at the incoming value A. When aggregation is activated (blue curves), the pitch is flattened around given notes N_i and the intensity decreases between them, in order to attenuate the transition.

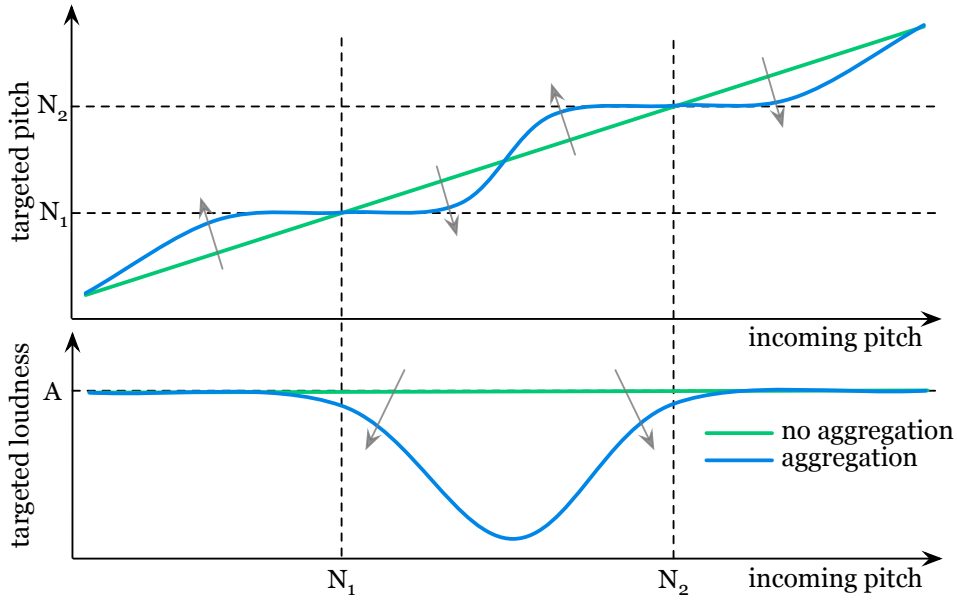

Figure 7.9: Example of aggregative control distorting pitch and intensity curves. Without aggregation pitch and intensity are as incoming from the tablet (green curves). When aggregation is required, pitch is flattened around given notes N_i, and intensity A is reduced between them (blue curves).

Articulative control

Movements on the FSR network reveal to be really dynamic. We have obtained that 10 gestures by second can be reached. Considering that each position on the network can be mapped to a symbolic value, it makes this configuration particularly close to the needed rate for generating phonemes in realtime (i.e. about 10 phoneme/second).

Through GloveTalk and GRASSP, Fels has shown that the achievement of fully hand-controlled speech synthesis (phonemes + prosody) is still an open problem [80], and the adding of voice quality modification even increases the complexity. In this thesis, we highlight that browsing a database from syllable to syllable is really intuitive with the FSR network [55]. But there is probably a really exciting research topic, related to the generalization of the $A + B$ strategy for generating phonetic streams.

In Figure 7.10 we can see how the FSR-based gestures would lead to the generation of a phonetic stream, that would be used as an input for the RAMCESS synthesizer. The mapping for the articulative control could be based on associating some finger positions with a given phoneme, exactly like it is done in the GRASSP framework [80].

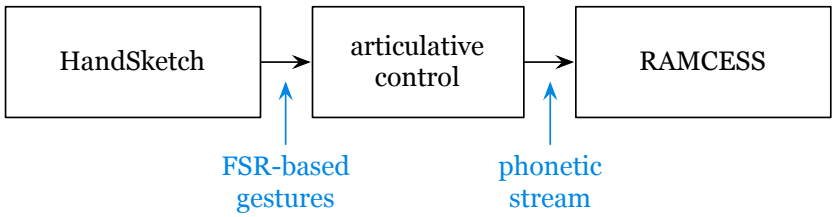

Figure 7.10: FSR-based gestures coming from the HandSketch, mapped to the Ramcess synthesizer in order to produce a phonetic stream in realtime.

7.4 Long-term practice of the instrument

When discussing about the making of new digital instruments, the issues of expertise and feedback, in the practice of these instruments, come recurrently. Poepel [152] or Dobrian [62] already addressed this problem in recent papers. Indeed they focuse on the relevance of virtuosity as a need for expressive interaction with sonic contents. More pragmatically they evoke the need for a larger contribution of skilled performers in the assessement of musical innovations.

It seems that existing frameworks – albeit very useful during conception – only give a part of the answer to this issue. We can highlight e.g. the digital instrument model [11], ergonomic assessments derived from HCI [198], or Cook's recommandations in the development of new controllers [46]. Obviously the amount of new controllers presented each year increases. But if we think about their lifetime and incorporation in the contemporary performer/audience dialogue, we do not have such a clear picture.

In this Section we do not expose improvements achieved on the existing HandSketch, but rather a comment on the fundamental reasons that pushed this instrument to reach its current shape, and a discussion about associated practices. The idea that the HandSketch is a novel instrument is reconsidered and the behavior of the graphic tablet itself as an expressive controller is generalized.

This reconsideration of the HandSketch aims at integrating our approach in the historical picture started by Zbyszynski [204]. He proposed the digital tablet as a major instrument of the contemporary music, and gathered the most significant players, in order to build a set of techniques that can be shared within the community.

In this Section, we discuss the way the vertical playing position evolved along these three years, from a rather static behavior to the position illsutrated in Figure 7.1. This discussion targets the size and the orientation of the tablet (cf. Section 7.4.1). It gives

interesting keys to understand how this position influences the overall attitude of the performer on stage. We also explain how the position justifies and even modifies the behavior of both the preferred and non-preferred hands (cf. Section 7.4.2).

7.4.1 Size and orientation

The first motivation for the current size and orientation of the tablet was the opportunity to develop new kind of gestures based on the writing skills [54]. Indeed most of existing practices associated with the graphic tablet were more or less close to a browsing strategy: taking advantage of the bi-dimensional control surface in order to move into a given sound space. With the use of a large tablet along the natural trajectory of the arm, more dynamic, expressive and "embodied" gestures could be achieved. The fact that the audience could see the control surface has also been highlighted as a interesting performing aspect.

However the decision of flipping and enlarging the sensing area was more or less intuitive. Without having a clear access to underlying reasons, it was difficult to think about transferring the instrumental interest of the HandSketch on other instruments. The use of existing interaction models or assessement strategies provided some answers:

- in the scope of usual HCI assessement, considering ergonomic aspects of the position: precision in moving, speed to reach a given point, etc;
- but also highlighting that using two hands with highly differentiated purposes gives better performances, as it is suggested by Kabbash [113].

In order to get further answers we had to involve time in practicing the instrument and discussing with many people about it[2]. It gives us today the possibility to highlight two mechanisms, considered as really important in order to consolidate the approach of playing vertical tablet:

- the fact that the gravity field and centers of gravity of the body play an important role in the way the performer and the instrument are connected;
- the way an object (i.e. its position, shape and size) influences the attitude of a performer and thus the expressive contents of his/her playing.

[2] The HandSketch participated to more than 30 events (concerts, demonstrations, workshops, etc) and has probably been tried (with different levels of involvment) by about 100 people.

Gravity-related performing issues

Research in applied physiology shows that the shape and the position of the human body is strongly related to the alignment of forces applied on different segments, such as shoulders or knees [202]. This can be seen as an intrinsic strategy for positioning ourselves in the gravity field. It defines how balance and tension are underlying our overall attitudes. Explicit use of the gravity can be found in advanced practices of several instruments, e.g. in the idea of moving *passively* the fingers during the bowing gesture[3] [119].

But it is interesting to notice how this topic is missing in the digital instrument making literature. Few contributors discuss the influence of gravity (and its impact on body/object interaction) in their practice of the instrument [166]. However this is probably one of the most important aspects of the vertical tablet playing in its way of highlighting the body expression, and the tilting of the tablet from horizontal to vertical playing highlights the importance of gravity.

In the *normal* use of the tablet (i.e. in horizontal position, in front of the performer), the pressure is achieved in the same direction as gravity, with the arm rather far from the center of gravity of the body. Consequently the body is static and comfortable, as the performer achieves browsing movements. Risk and effort, two aspects crucially involved in the interest of a live performance, appear not to be accessible for the audience.

At the beginning of the HandSketch, the tablet was placed fully vertically on the knees while sitting down on a chair. Then the overall position progressively moved from the formal sitting on a chair to a different attitude: sitting on the ground (cf. Figure 7.1). There are three significant differences between the former and the current positions.

1. Verticality is broken. As the device is supported by the lateral part of the knees on one side, and by the upper part of the chest on the other side, it makes an angle V of 30-40 degrees with the vertical direction, as illustrated in Figure 7.11.

2. The angle of the tablet V is correlated with the movement of the spine. Therefore this angle can vary, as illustrated in Figure 7.11. This aspect is really important because the behavior of the spine is correlated with the emotional state [32].

[3] The body/object relation and the effect of gravity are also an important issue in other activities such as martial arts, and more precisely with the *Bokken* [130].

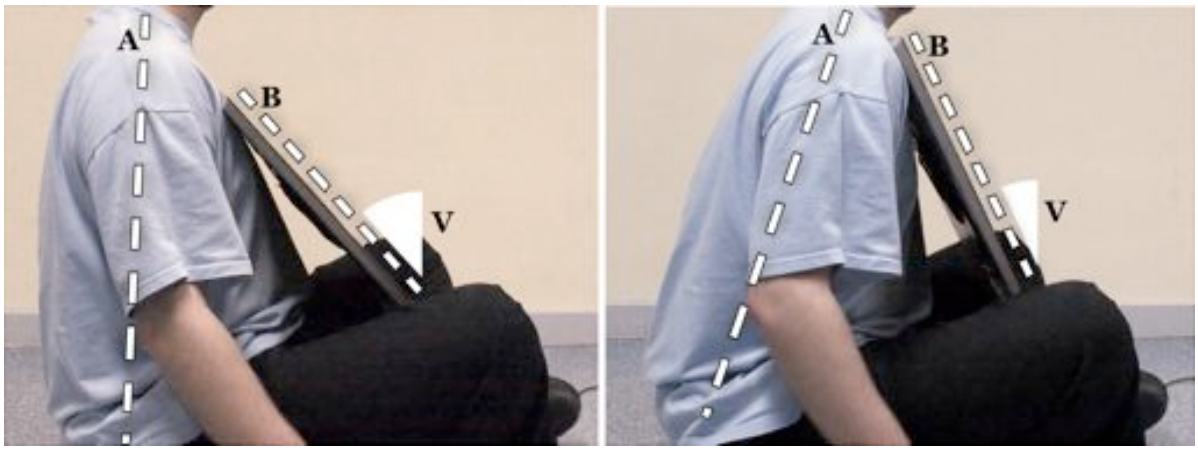

Figure 7.11: Tilt (B) of angle V due to spine movements (A).

3. A given position on the sensing area becomes a *suspended* situation – pressing on a tilted surface is unstable – and requires concentration (cf. Figure 7.12). Playing that way for a long period reveals that this unstable connection between the body (through the behavior of the spine) and the located pressure on the surface helps the audience for understanding the risk and the difficulty of the performance.

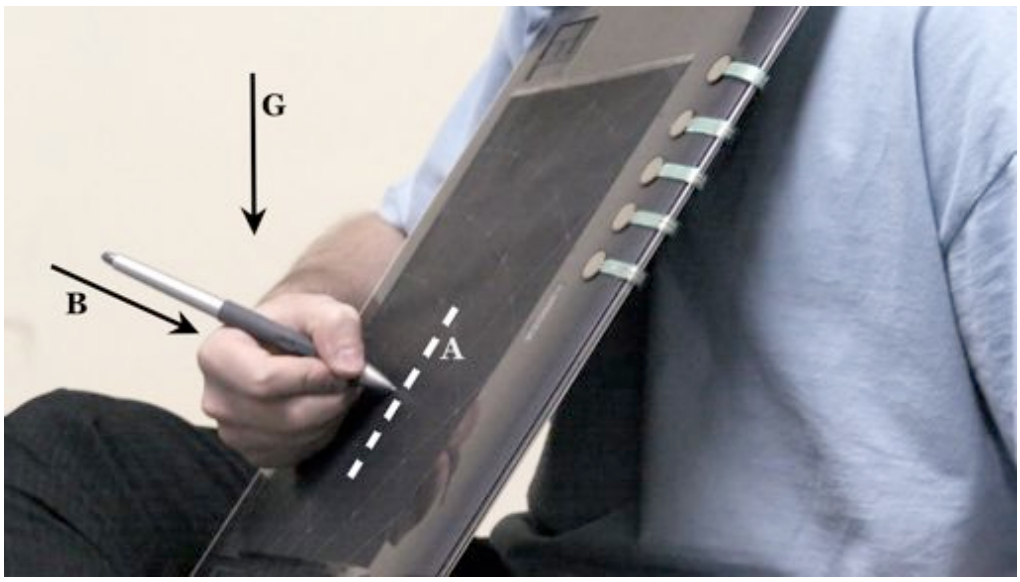

Figure 7.12: Gravity (G) and pressure (B) on the tilted area (A).

Keeping the concept, changing the size

Evoking the concept of category in the context of digital instrument making is difficult. All electronic or digital instruments are often classified in one big cluster: that of "not

acoustic" instruments. Decomposing a given instrumental concept (e.g. bowed strings) and developing different practices mainly due to the size (e.g. violin, cello, double-bass) is not obvious in the digital world. Except for the digital keyboard, and its number of keys leading to various sizes, sizing rarely happens for digital instruments [135].

Playing the graphic tablet gives the opportunity of accessing various sizes (from A6 to A3) with the same resolution and features. The wacom object for Max/MSP can send (x, y) coordinates as a relative position between 0 and 1 for all the supported tablets. Therefore the size of the controller can be changed easily without disturbing sound synthesis. It creates a comfortable context to test the influence of size.

Figure 7.13: Position when the A6 tablet is played.

As it could be expected the relative size of the tablet – compared to the size of the body – plays an important role in the attitude proposed by the performer. Smaller tablets (A5 and A6) can be played on one knee (cf. Figure 7.13). The performer is more comfortable and invites the audience to focus on what is happening around that location. With bigger devices the way of playing is much more imposed by the shape of the controller, and the connection with the spine is stronger. The performer and the tablet become like a unique system, and expressing results from the body behavior.

7.4.2 Generalizing the aim of each hand

In Sections 7.2 and 7.3, the hands are described from their functional point of view. Considering Cadoz's typology [33] it means that the preferred hand makes a modulation gesture through the pen scrubbing the tablet, and the non-preferred hand performs selection gestures on the FSR network. However we explain in Section 7.4.1 that the preferred hand plays a much more important role in the achievement of expressivity. Indeed the contact point between the pen and tablet is a complex combination between the tilt of the surface and how the position of the arm is influenced by gravity. Therefore the preferred hand can be seen as the tensing factor of the performing behavior. If the performer relaxes this hand, the contact point slips out of the sensing area and the relation stops. In our mapping, the sound would stops as well.

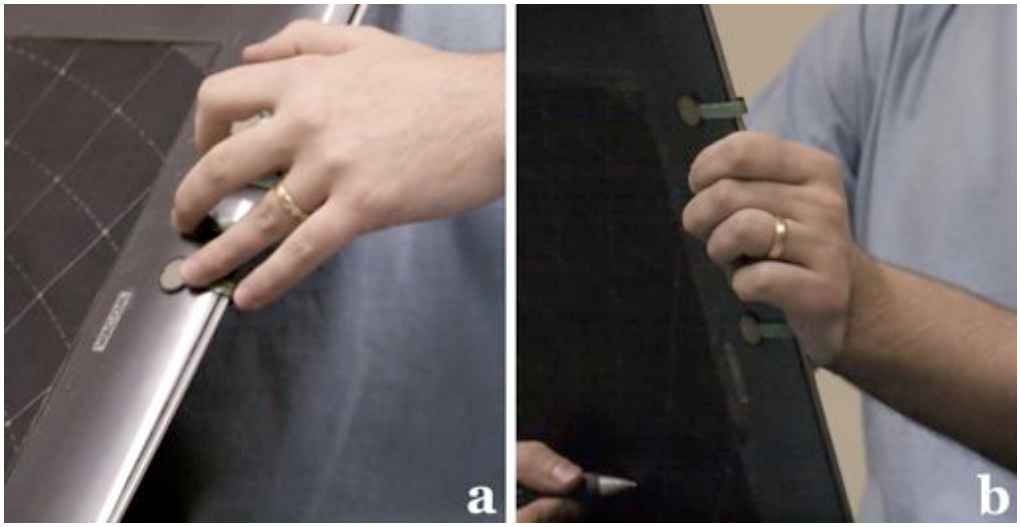

Figure 7.14: Different attitudes with the non-preferred hand.

The practice of the tablet reveals that the non-preferred hand also has a kind of hidden purpose. Indeed we explain in Section 7.4.1 that the tablet is linked with spine movements, creating a strong correlation with the behavior of the upper part of the body. In this context the non-preferred hand is intuitively used in order to develop the body movements in other directions e.g. tilting the tablet further than vertical.

Finally the exact configuration of sensors for the non-preferred hand is not so crucial as soon as there is a continuum in the grabbing attitude, from the full acceptance to the total rejection, respectively illustrated in Figures 7.14a and 7.14b.

7.5 Conclusions

In this Chapter, we described the development of a new digital instrument, based on a graphic tablet and attached FSR sensors: the HandSketch. The prototyping of this instrument results from the Analysis-by-Interaction methodology that has been explained in Chapter 6. Here we present several important aspects of this work:

Innovative mapping for pen-based gestures

The main aspect of the HandSketch is the use of pen-based gestures for the combined control of various aspects of voice production. In this Chapter, we have first introduced how the tablet, a 3-axes controller, could be mapped to some voice production parameters: pitch, loudness and voice quality dimensions. Solutions to some ergonomic problems have also been proposed, leading to an adapted circular representation of pitch, and the use of radial finger-based movements for voice quality modifications.

Embedded FSR network and vertical playing

The role of the non-prefered hand has also been discussed and a controller, embedded on the tablet, has been proposed. This non-prefered hand controller is based on a FSR network. The unusual configuration of FSR sensors (five sensors on the front panel and three sensors on the rear panel) has modified the playing position from horizontal to vertical. Three mappings have been proposed for this FSR network, based on various purposes for the non-prefered hand: fretboard, aggregative and articulative controls.

Long term practice of the instrument

The development of the HandSketch takes the benefit of three years of playing, and eight successive prototypes. This continuous combination of prototyping and practice gave the opportunity to discuss in details the underlying aspects of this tablet-based musical performance. In this discussion, new properties have been highlighted in the playing, such as the impact of size and orientation of the tablet on the overall performing behavior. These new properties are important in order to plan further development of the HandSketch and extend the interest of tablet playing to new instruments.

Chapter 8

Performing Vocal Behaviors

“ L’oreille humaine n’aime pas le son nu. ”
— Jean-Pierre Blivet

8.1 Introduction

In Chapter 6, we discuss the development of a new methodology, called Analysis-by-Interaction (AbI). This methodology uses the long-term practice of a digital instrument, as a way of exploring sound production understanding differently. In Chapter 7 we describe the whole development of our prototype: the HandSketch.

Straightforwardly this last Chapter presents some results obtained by applying AbI with the HandSketch, in the context of expressive voice production. More precisely, the high level practice with the instrument is used to study some glottal properties of the singing vibrato. This study leads to a new model for synthesizing expressive vibrato.

In Section 8.2 we demonstrate with a simple experiment that the HandSketch playing is effectively highly embodied. Then we study the synthesis of vibrato synthesis in Section 8.3. After some background explanation concerning the vibrato effect, we show various properties of HandSketch gestures. Finally we formulate a production model for the synthesis of the vibrato, based on glottal flow parameters: F_0, O_q, α_M and T_L.

8.2 Validation of embodiment in HandSketch practice

As explained in Chapter 6, our approach is based on the skilled practice of the instrument, characterized by a high embodiment. In this Section, we discuss a small experiment that aim at demonstrating this high embodiment with the HandSketch.

The same melody is asked to be performed once a day, during five days. The session starts with the hearing of a simple MIDI reference. Then the performance with the instrument is recorded. Gestures are sampled every 10 ms. Moreover the two last performances of the melody are achieved without audio feedback.

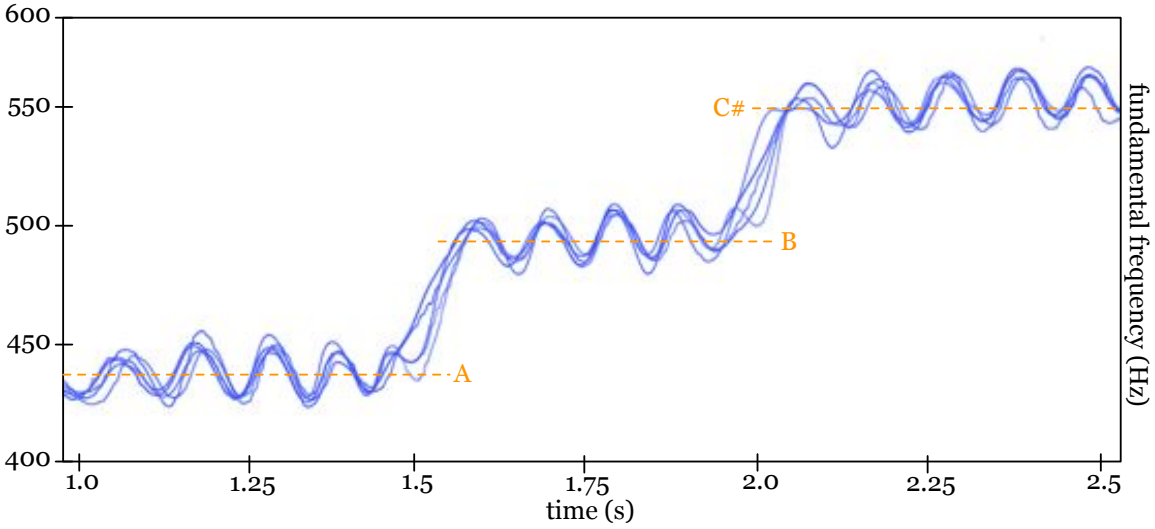

Figure 8.1: Superimposition of five performances of a reference melody with the HandSketch. The five pitch curves (f_0) overlaps, despite the long period between each take, and despite no audio feedback for two of them.

Figure 8.1 illustrates the superimposition of the five performances, for a given part of the melody (A - B - $C\#$). Despite the long period between each recording session (about 24 hours) and despite the canceling of the audio feedback for two of them, the pitch curves (blue lines) perfectly overlap. Such a precision and a repeatability in the achievement of a gestural trajectory illustrates the high embodiment of the instrument.

8.3 Case study: vibrato in singing synthesis

Vibrato is a musical effect that has been introduced, in the XVIIth century, in the Western music, as a way of emphasizing a note. It was used in various kinds of instruments:

bowed string, flute or singing voice. In the XIXth century, vibrato has been extended as a more expressive technique. Most of musical instruments includes a vibrato effect [170].

For the singing voice, the vibrato is achieved by a complex laryngeal modulation. This modulation has an influence on the whole air flow. It influences the fundamental frequency, the intensity, but also the spectral envelope of the sound [185].

It is particularly interesting to use vibrato as a case study, because vibrato is a good example of an intimate gesture. Indeed the expressivity and naturalness of the vibrating singing voice are intimately related to the control of the voice production [23].

Three years of practicing and refining the HandSketch led us to be able to produce expressive and natural vibrato sounding [55]. In the context of the AbI methodology, we study corresponding gestures, and compare them with the state of the art in synthesis of vibrato in singing voice. This study leads us to propose a new approach.

8.3.1 Background in vibrato for the singing voice

Vibrato is perceived as an overall vibrating quality of the sound timbre. In [192], Verfaille *et al.* propose a deep review of vibrato properties and describe the generalized vibrato effect, for various kinds of instruments, along three axes: frequency modulation (FM), amplitude modulation (AM) and spectral envelope modulation (SEM):

- Frequency Modulation (FM): vibrato can alter the melodic curve i.e. the time domain evolution of the fundamental frequency. When it happens, pulsations are superimposed on the trajectory of F_0. This deviation is not perceived as melodic modifications, but as an overall vibrating quality of the timbre.
- Amplitude Modulation (AM): vibrato can alter the intensity curve i.e. the time domain envelope of the sound. When it happens, pulsations are superimposed on the main energy of the sound. This deviation is not perceived as a variation of the volume, but as an overall vibrating quality of the timbre.
- Spectral Envelope Modulation (SEM): vibrato also has a rich effect on the evolution of the spectral envelope of the sound. Indeed the FM and/or AM vibration can be synchronous with a cycle in the shape of the spectral envelope.

For singing voice, the main effect is that of frequency modulation. However Verfaille mentiones that spectral envelope modulation is also encountered and is involved in the

naturalness and expressivity of the vibrato [192]. The vibrato results from a complex laryngeal modulation [23], so that impact on the spectral envelope can not be neglected.

In this Section, we describe the influence of the vibrato effect on the F_0 contour (FM). Then we discuss how the generalized vibrato model introduces AM and SEM, and what is the impact of this combination on trajectories of harmonics of the magnitude spectrum.

F_0 contour in the singing voice

In singing, the main structure of the fundamental frequency curve results from the notes that are produced. In the equally tempered scale, each note corresponds to a unique F_0 value. The basic melodic line can thus be seen as a step function.

Vibrato is superimposed on the trajectory of the fundamental frequency. The short term effect of vibrato is a detuning of the expected F_0, alternatively below and above this reference. Figure 8.2 show features of this deviation: frequency, amplitude and shape.

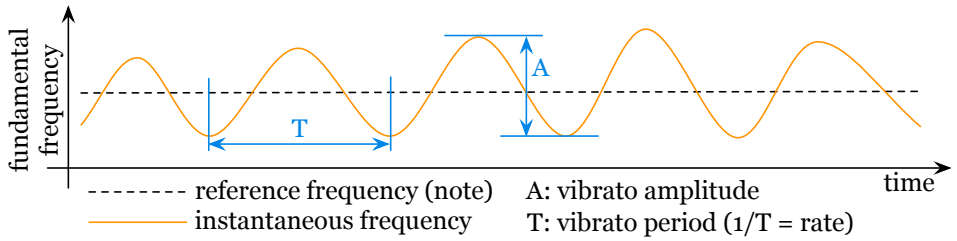

Figure 8.2: Frequency modulation of the vibrato: an detuning of the reference note, alternatively below and above the reference frequency (dashed). This detuning is characterized by its frequency $f = 1/T$, amplitude A, and shape, here mainly sinusoidal.

The vibrato frequency $f = 1/T$ is generally around 6 Hz with a variation of about $\pm 8\%$ [153]. Depending on the singing style and singer, this frequency can go from 4 to 12 Hz [60]. Prame also mentions that the frequency increases at note endings [153]. The amplitude of vibration A goes from 0.6 to 2 semitones, depending on singing style and singer [187].

The perceptual impact of the shape of the oscillation is not a widely addressed topic. It appears that different shapes can be performed, depending on the musical context: sinusoidal, triangular, trapezoidal and even less identifiable [104, 187]. However the sinusoidal shape is the simplest to implement [165], as illustrated in equation (8.1).

$$f_0(t) = i(t) + e(t) \times cos(\phi(t)) \tag{8.1}$$

where $i(t)$ is the step function corresponding to the reference note, $e(t)$ is the envelope function modulating the amplitude of the oscillation, and $\phi(t) = \omega(t) \times t + \phi$ is the phase of the oscillation: $\omega(t)$ is the time-varying angular speed, and ϕ the original phase.

The effect of vibrato on note transition has also been studied. Indeed there is a strong interaction between the phase $\phi(t)$ and the expectation of a note transition [48, 59]. Results show that singers anticipate transitions and intuitively adjust $\phi(t)$ – and more precisely $\omega(t)$ – in order to synchronize an ascending (descending) note transition with an ascending (descending) segment of the vibrato oscillation.

This time domain alignment between the note transition (step) and the vibrato phase $\phi(t)$ is reinforced by another phenomenon. When a singer goes from one note to another – ascending or descending – without pausing the phonation, the F_0 curve of the note transition is slightly larger than the note interval, resulting in two effects:

- preparation: the F_0 curve slightly decreases (increases) in the opposite direction, right before an ascending (descending) note transition;
- overshoot: the F_0 curve slightly overruns above (below) the frequency of the note right after the ascending (descending) note transition.

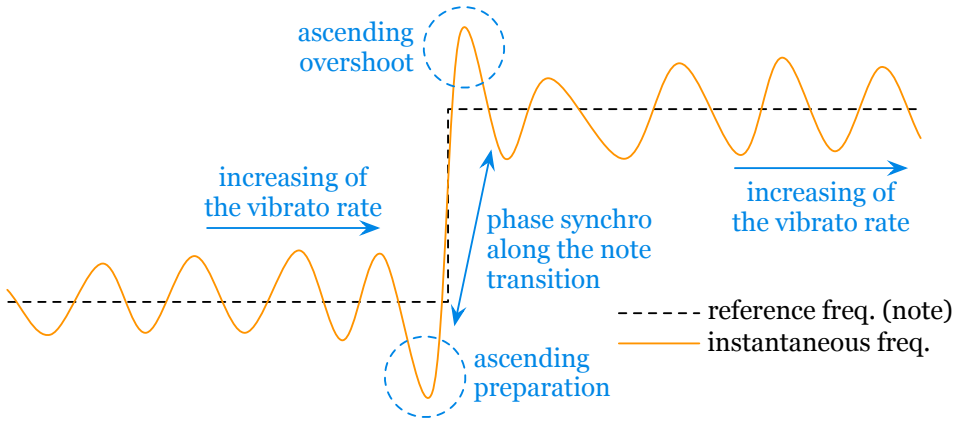

Figure 8.3: Frequency modulation of the vibrato on a note transition. Several phenomena are highlighted: the increasing of the vibrato frequency at note endings, the synchronization of $\phi(t)$ within the note transition, preparation and overshoot.

Figure 8.3 illustrates the combination of these phenomena for a note transition. The reference step (dashed) underlies the more complex F_0 curve (orange). We can see the oscillation, its increasing at note endings, the synchronization of the phase $\phi(t)$ with the note transition, and preparation/overshoot around the note transition.

Spectral effect of combined AM, FM and SEM

Sundberg describes the vibrato as a pseudo-periodic modulation of the air flow, achieved by the glottal source. This modulation also influences vocal tract resonances, due to some coupling effects [185]. As mentioned above, this modulation causes the oscillation of the fundamental frequency (FM), but it also produces a vibrating behavior on the intensity (AM) and the spectral envelope (SEM), somehow synchronized with the pitch curve.

In the generalized vibrato model [192], the movement of each harmonic of the magnitude spectrum[1] is deeply studied, as a function of the time t. We define $a_h(t)$ and $f_h(t)$, respectively the evolution of amplitude and frequency of the harmonic h.

Vibrato is defined as periodic trajectories of harmonics in the (frequency, amplitude, time) space. Let us call it (f, A, t). Figure 8.4 gives an example of an arbitrary evolution of the harmonic h. We can see that $a_h(t)$ and $f_h(t)$ are projections of the trajectory in the (f, A, t) space, respectively on (A, t) and (f, t) subspaces.

Obviously $a_h(t)$ and $f_h(t)$ are not independent functions. Indeed $a_h(t)$ always results from the scanning of the spectral envelope at a given time t, and for a given frequency, depending on $f_h(t)$. We define $\xi(f, t)$, the spectral envelope of the sound at time t. $\xi(f, t)$ is a function in the subspace (f, A), and can be estimated by various techniques: LPC [134], discrete cepstrum [85] or interpolation between harmonics [173].

FM only Most of singing synthesis algorithms and speech-to-singing conversion tools consider that vibrato only consists in FM [185]. This commonly accepted assumption means that the spectral envelope remains unchanged during vibrato, thus $\xi(f, t) = \xi(f)$. The spectral behavior of the sound is completely defined by deviations applied on $f_h(t)$.

Another common assumption is to consider that the deviation, which is applied on the frequency of each harmonic h, has the sinusoidal form, as in equation (8.2).

[1] This representation implicitly means that we work with sinusoidal modeling of the singing voice (SMS), which has been mentioned to be relevant in many studies [173, 24].

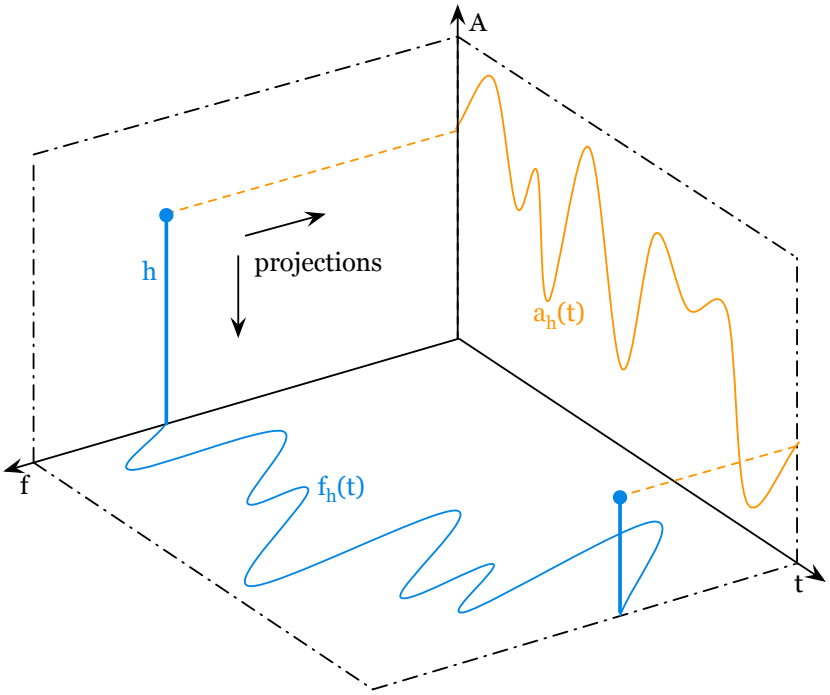

Figure 8.4: Trajectory of a harmonic – from the sinusoidal model of the singing voice – in the (f, A, t) space. This trajectory can be projected on (A, t) and (f, t) subspaces, in order to obtain respectively the $a_h(t)$ and $f_h(t)$ functions.

$$f_h^{FM}(t) = f_h^{ref}(t) + A_h^{FM}(t) \times cos(\omega_h^{FM}(t) \times t + \phi_h^{FM}) \tag{8.2}$$

where $f_h^{ref}(t)$ is the reference frequency of the harmonic h (belonging to the underlying flat magnitude spectrum), and $A_h^{FM}(t)$, $\omega_h^{FM}(t)$, ϕ_h^{FM} respectively the amplitude, the angular speed and the initial phase of the harmonic h oscillation.

If we consider that the signal is still harmonic during the application of the vibrato effect, we can simplify this equation, by expressing that the phase is the same for all harmonics, and that the amplitude preserves the relation $f_h = h \times f_0$, as in equation (8.3). Then the amplitude of each harmonic can be computed with equation (8.4).

$$f_h^{FM}(t) = f_h^{ref}(t) + A^{FM}(t) \times h \times cos(\omega^{FM}(t) \times t + \phi^{FM}) \tag{8.3}$$

$$a_h^{FM}(t) = \xi(f_h^{FM}(t)) \tag{8.4}$$

FM + AM Some singing synthesis algorithms propose the introduction of AM [103, 161]. In this case, the spectral envelope corresponds to a constant shape $\xi(f)$, multiplied by an oscillating factor $\gamma^{AM}(t)$, resulting in the time-varying spectral envelope in equation (8.5). Sinusoidal AM is commonly accepted, as illustrated in equation (8.6).

$$\xi^{AM}(f,t) = \gamma^{AM}(t) \times \xi(f) \tag{8.5}$$

$$\gamma^{AM}(t) = 1 + A^{AM}(t) \times cos(\omega^{AM}(t) \times t + \phi^{AM}) \tag{8.6}$$

where $A^{AM}(t)$, $\omega^{AM}(t)$, ϕ^{AM} are respectively the amplitude, the angular speed and the initial phase of the sinusoidal AM. As the whole vibrato is periodic, we consider that $\omega^{FM}(t) = \omega^{AM}(t)$. In [192] the same assumption is done for the initial phase: $\phi^{FM} = \phi^{AM}$.

Finally, the amplitude of each harmonic can be computed with a modified version of equation (8.4), with the time-varying $\xi^{AM}(f,t)$, here represented in equation (8.7).

$$a_h^{FM+AM}(t) = \xi^{AM}(f_h^{FM}(t), t) \tag{8.7}$$

FM + AM + SEM We find some singing synthesis techniques that integrate SEM [133, 136]. One technique proposes that the spectral envelope $\xi(f,t)$ is a linear interpolation between two static spectral envelopes: $\xi_-(f)$ and $\xi_+(f)$ [133], as described in equation (8.8). The pulsation that interpolates between $\xi_-(f)$ and $\xi_+(f)$ is also assumed to be sinusoidal (between 0 and 1), as described in equation (8.9).

$$\xi^{SEM}(f,t) = (1 - \beta^{SEM}(t)) \times \xi_+(f) + \beta^{SEM}(t) \times \xi_-(f) \tag{8.8}$$

$$\beta^{SEM}(t) = \frac{1 + cos(\omega^{SEM}(t) \times t + \phi^{SEM})}{2} \tag{8.9}$$

where $\omega^{SEM}(t)$ and $\phi^{SEM}(t)$ are respectively the angular speed and the initial phase of the interpolating pulsation. If we consider that the whole vibrato effect is periodic, we can assume that $\omega^{SEM}(t) = \omega^{AM}(t) = \omega^{FM}(t)$. In [192], the same assumption is done for the initial phase: $\phi^{SEM} = \phi^{AM} = \phi^{FM}$. Figure 8.5 illustrates the SEM process.

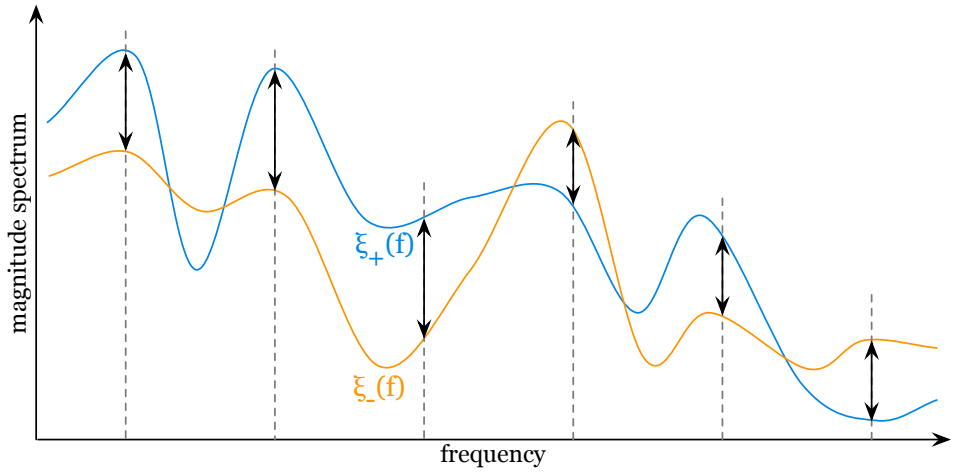

Figure 8.5: Two spectral envelopes are taken as extrema: $\xi_-(f)$ (orange) and $\xi_+(f)$ (blue). $\beta^{SEM}(t)$ linearly interpolates between these two situations, with a sinusoidal shape: going from $\xi_-(f)$ to $\xi_+(f)$, and symmetrically coming back to $\xi_-(f)$.

Combining AM and SEM effects is achieved with equation (8.10). Harmonic amplitudes when FM, AM and SEM are combined result from the scanning of this oscillating spectral envelope with the oscillating frequencies, as represented in equation (8.11).

$$\xi^{AM+SEM}(f,t) = \gamma^{AM}(t) \times \xi^{SEM}(f,t) \tag{8.10}$$

$$a_h^{FM+AM+SEM}(t) = \xi^{AM+SEM}(f_h^{FM}(t),t) \tag{8.11}$$

Perceptive tests show that SEM significantly improves the naturalness of the generated vibrato [192]. This improvement of the quality is attributed to the more complex pattern achieved by harmonics in the (f, A) subspace. Indeed we can understand that FM alone ($f_h(t)$ scanning) emphasizes the unchanging behavior of the spectral envelope.

8.3.2 Drawbacks of the generalized vibrato model

The generalization of the vibrato effect in a combination of FM, AM and SEM appears to be an interesting approach, in order to generate natural vibrating sounds. However we would like to highlight several drawbacks in the current implementation, considering that we want the vibrato effect to have expressive and interactive properties.

Parameterization of SEM

As illustrated by equations (8.8) and (8.9), one advanced way of implementing SEM is the linear interpolation between two extreme spectral envelopes $\xi_-(f)$ and $\xi_+(f)$. These two spectral envelopes are measured on real waveforms, at extreme instants of the vibrato cycle. This measurement explains the assumption $\phi^{SEM} = \phi^{AM} = \phi^{FM}$.

For the singing voice, this strategy means that spectral envelopes $\xi_-(f)$ and $\xi_+(f)$ have to be evaluated on real singing waveforms with vibrato, and for each phonetic context. Consequently, expect for the f_0 modification and the corresponding scanning of $f_h(t)$, the spectral behavior of the vibrato effect is totally determined by the database.

However we know that the vibrato in singing results from a complex laryngeal behavior, with a high level of intuitive control from the singer [23]. The generalized vibrato model proposes the control of a overall phase, and amounts of AM and FM applied on the signal. We expect a vibrato model to be more flexible at the production level.

Hysteresis in harmonic pattern

When the trajectory of harmonics is evaluated on real signals, and plotted in the (f, A) subspace, we can observe that the curve is not exactly equal in forward and backward movements of the vibrato. Harmonics achieve *hysteresis* along their trajectories with one cycle of vibrato. It means that vibrato is not a totally symmetric process.

In [192], the issue of hysteresis is addressed for the generalized vibrato model. Considering that modulating parameters are fixed, as represented in equations (8.12) to (8.13), a condition for no hysteresis is established: the flatness of $\xi_-(f)$ and $\xi_+(f)$ around each harmonic h. In any other case, the generalized vibrato model produces hysteresis.

$$\omega^{FM}(t) = \omega^{AM}(t) = \omega^{SEM}(t) = \omega \tag{8.12}$$

$$A^{FM}(t) = A^{FM} \; ; \; A^{AM}(t) = A^{AM} \; ; \; f_h^{ref}(t) = f_h^{ref} \tag{8.13}$$

However we have implemented the generalized vibrato model in our interactive synthesis context, and we did not find any situation, respecting equations (8.12) to (8.13), that produce hysteresis in harmonic patterns, with any kinds of spectral envelopes.

We propose an experimental study of the generalized vibrato model and show that conditions for the systematic production of hysteresis are different. Our procedure first sets two random spectral envelopes $\xi_-(f)$ and $\xi_+(f)$, thus not respecting the flatness suggested in [192]. Then we apply FM, AM and SEM with the phase alignment:

$$\phi^{FM} = \phi^{AM} = \phi^{SEM} \tag{8.14}$$

Figure 8.6 illustrates that for any kind of spectral envelope $\xi_-(f)$ and $\xi_+(f)$, forward (blue) and backward (green) trajectories[2] are completely overlapped, for a given harmonic h. In this example, f_h^{ref} is set to 500 Hz and $A_{FM} = 40$ Hz.

Regarding these results, we postulate that hysteresis are due to other kind of phenomena within the vibrato cycle. These phenomena lead to asymmetrical forward/backward movements. Thus we consider that the generalized vibrato model has to be used with other assumptions, or that another model has to be proposed:

- some components of the vibrato effect use a different initial phase;
- some components of the vibrato effect are not sinusoidal;
- the transition between $\xi_-(f)$ and $\xi_+(f)$ is not a linear interpolation.

[2] Forward (backward) trajectory means the movement from the minimum (maximum) to the maximum (minimum) position of the sinusoidal cycle. Forward is $\phi = 0 \rightarrow \pi$, and backward is $\phi = \pi \rightarrow 2\pi$.

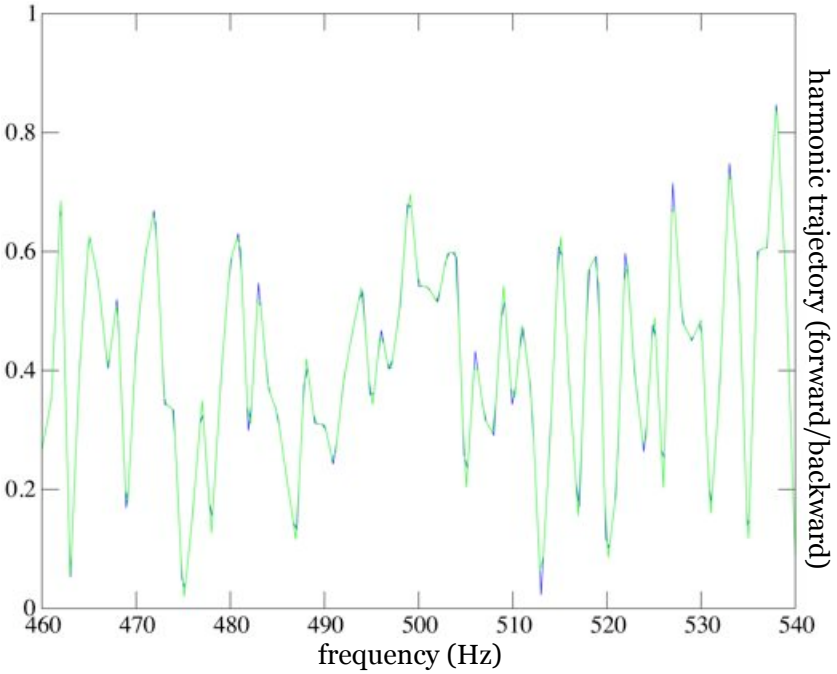

Figure 8.6: For any kind of spectral envelope $\xi_-(f)$ and $\xi_+(f)$, forward (blue) and backward (green) trajectories are completely overlapped, for a given harmonic h.

8.3.3 AbI with HandSketch-based gestures

We reach a situation where the use of the Analysis-by-Interaction methodology (cf. Chapter 6) becomes particularly interesting. Indeed the understanding of the real laryngeal activity – in the sense of production parameters – during vibrato would lead to more realistic and controllable models of the singing voice. But this laryngeal activity is not precisely measurable, as it has been described in Section 3.3.1. On the other hand, vibrato coming out of the HandSketch has been awarded as being really natural.

Consequently we postulate that the analysis of HandSketch gestures provide new information, potentially interesting for proposing a new model of vibrato for the singing voice. In that sense, we trust the skilled performer in his ability to *recreate* the underlying glottal behavior. This experimental work has been achieved by Ooge [149].

In this experimentation, ongoing values of f_0, O_q, α_M and T_L are recorded every 10 ms during the performance of a vibrato, with the HandSketch plugged in the Ramcess synthesizer. Then a few seconds are plotted, with their amplitudes fitted in the same range, in order to better observe the phase of each parameter.

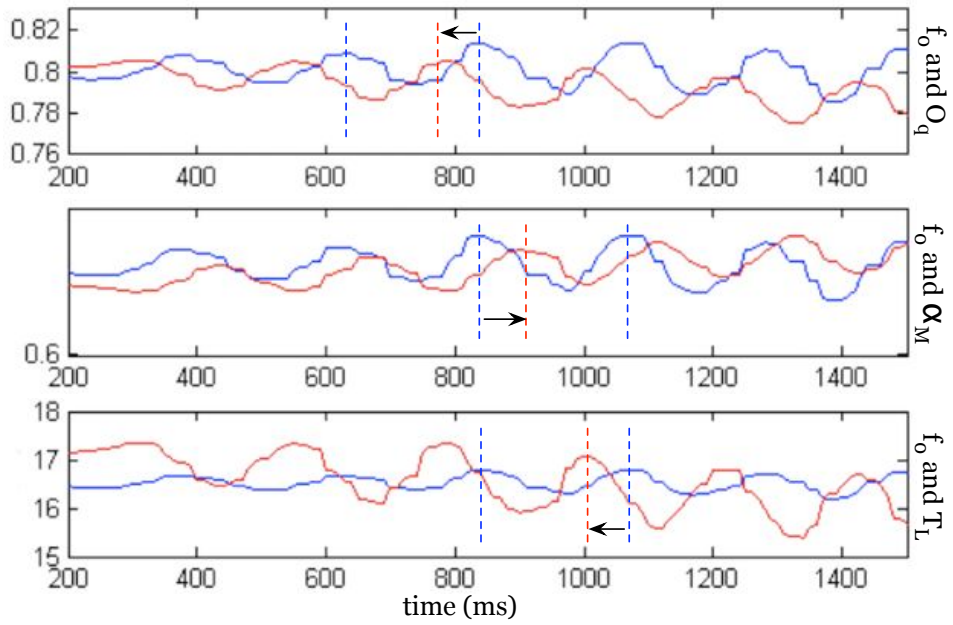

Figure 8.7: Evolution of O_q (top), α_M (middle) and T_L (bottom) (red), superimposed to f_0 (blue), for a few period of vibrato with the HandSketch.

Figure 8.7 represents the evolution of O_q (top), α_M (middle) and T_L (bottom), superimposed to f_0, for a few period of vibrato. We observe a dephasing between the effect of vibrato on the fundamental frequency (FM) and on production parameters of the glottal source (SEM). Average dephasing (f_0 as reference) is reported in Table 8.1.

$\Delta\phi$	O_q	α_M	T_L
f_0	$-\pi/2$	$+\pi/2$	$-\pi/2$

Table 8.1: Average dephasing $\Delta\phi$ between the effect of vibrato on f_0 and on glottal source parameters O_q, α_M and T_L, as estimated on HandSketch gestures.

The way of producing the vibrato effect with the HandSketch is significantly different from the assumptions used in the generalized vibrato model. Indeed the dephasing of $\frac{\pi}{2}$ that is observed reveals that forward and backward movements of the voice spectrum are not symmetrical. Moreover the idea of hysteresis is now related to physiological features, leading to more flexibility and control, compared to spectral envelope interpolation.

8.3.4 Vibrato model for the control of SELF

The good results obtained with the HandSketch vibrato, and the fact that this quality is due to an underlying hysteresis in the glottal source production mechanisms, lead us to propose a vibrato model for the control of SELF (cf. Section 4.3).

In this vibrato model, all oscillations are sinusoidal, and relying on the same pulsation ω, just as it as been suggested in equation (8.12). The input of the production model consists in a set of reference parameters: $\{f_0^{ref}, E^{ref}, O_q^{ref}, \alpha_M^{ref}, T_L^{ref}\}$. Then the amplitude of the vibrato is adjusted on each parameter with a set of deviation factors $\{\delta f_0, \delta E, \delta O_q, \delta \alpha_M, \delta TL\}$. Each oscillation is achieved as in equation (8.1).

The particularity in this model consists in introducing a hysteresis factor, through the variable initial phase $\phi(t)$. This phase shift is introduced in the oscillation of O_q, α_M and T_L in the same direction as it has been measured on HandSketch gestures. We see the mathematical development of this vibrato model in equations (8.15). Figure 8.8 shows how the vibrating parameters are plugged in the Ramcess system.

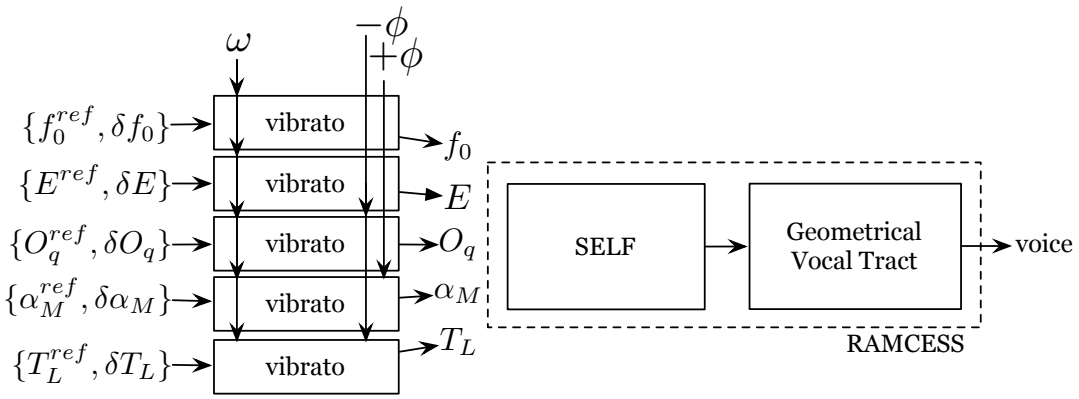

Figure 8.8: Vibrato models applied on glottal source parameters, and then being plugged in the Ramcess synthesizer (more precisely the SELF model). A positive or negative phase shift is introduced in vibrations of O_q, α_M and T_L.

$$\begin{cases} f_0(t) &= f_0^{ref}(t) + \delta f_0(t) \times cos(\omega(t) \times t) \\ E(t) &= E^{ref}(t) + \delta E(t) \times cos(\omega(t) \times t) \\ O_q(t) &= O_q^{ref}(t) + \delta O_q(t) \times cos(\omega(t) \times t - \phi(t)) \\ \alpha_M(t) &= \alpha_M^{ref}(t) + \delta \alpha_M(t) \times cos(\omega(t) \times t + \phi(t)) \\ T_L(t) &= T_L^{ref}(t) + \delta T_L(t) \times cos(\omega(t) \times t - \phi(t)) \end{cases} \quad (8.15)$$

Equations (8.15) correspond to the general vibrating behavior of the glottal source. All the reference parameters $X^{ref}(t)$ correspond to the underlying production without vibrato. Fundamental frequency and intensity – through $E(t)$ – are vibrating in phase, as in [192]. $\phi(t) = 0$ leads to the same kind of interpolation between two spectral envelopes. But increasing $\phi(t)$ introduces some asymmetry in the whole vibration process.

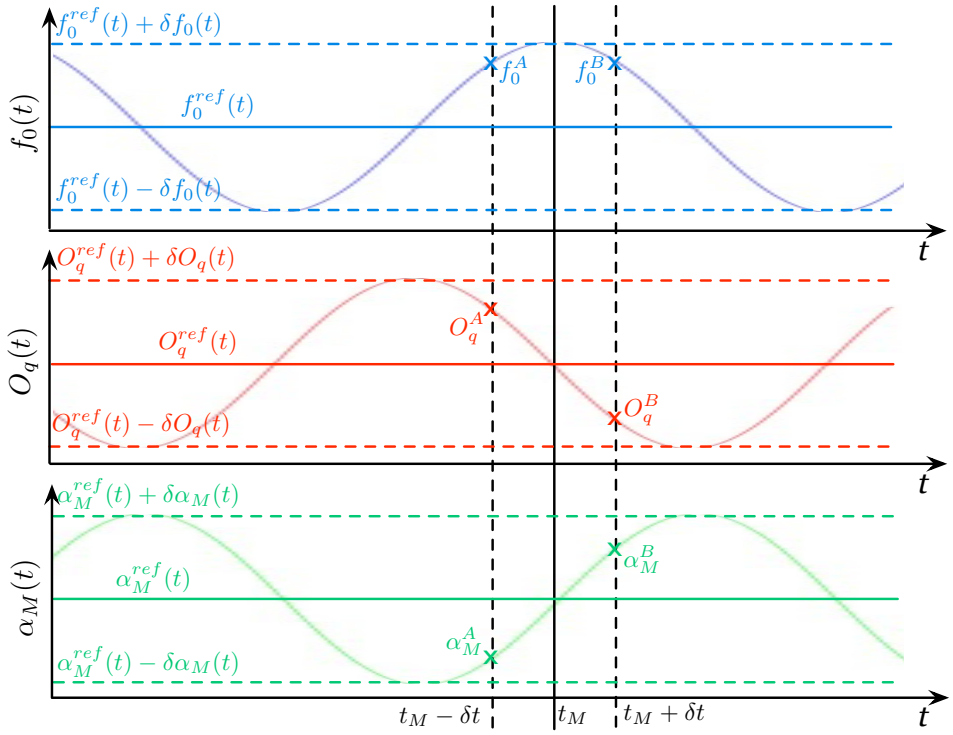

Figure 8.9: Evolution of O_q (top), α_M (middle) and T_L (bottom) (red), superimposed to f_0 (blue), for a few period of vibrato with the HandSketch.

In Figure 8.9 we show the oscillations of $f_0(t)$ (blue), $O_q(t)$ (red) and $\alpha_M(t)$ (green), with $\phi(t) = \pi/2$. We see their respective oscillations within the boundaries imposed by equations (8.15). If we define the instant t_M where the oscillation of $f_0(t)$ is maximum, we can examine the values of $f_0(t)$, $O_q(t)$ and $\alpha_M(t)$ at $t_M - \delta t$ and $t_M + \delta t$, defined symmetrically around t_M. Obviously $f_0(t)$ has the same value before and after t_M, thus $f_0^A = f_0^B$. But this is not the case for $O_q(t)$ and $\alpha_M(t)$, $O_q^A \neq O_q^B$ and $\alpha_M^A \neq \alpha_M^B$.

Thus a dephasing $\phi(t) > 0$ leads to different trajectories for glottal source parameters in ascending and descending parts of the vibrato cycle. This hysteresis in glottal source parameters leads in different ascending/descending movements in the spectral envelope along the vibrato cycle, and thus leads to hysteresis in the harmonics.

8.4 Conclusions

In this Chapter, we have shown how the HandSketch, involved in the AbI methodology, leads to reach interesting results in the field of voice modeling, through the case study of the synthesis of vibrato in singing. Here we present the main axes of this study:

Three years of HandSketch practice lead to highly embodied skills

We have shown that the three years of practicing the HandSketch lead to reach high precision and great embodiment in the achievement of performing gestures. Indeed this property has been illustrated through a playing experiment, where the same melody leads to significantly similar pitch contour, even without audio feedback.

New model for the synthesis of vibrato in singing, based on AbI and HandSketch

Playing the HandSketch revealed the possibility of producing expressive and spectrally rich vibrato with a rather simple singing synthesizer. Based on AbI assumptions, the corresponding gestures have been analyzed and an hysteresis has been highlighted, in the oscillating movements of glottal source parameters.

Comparing the generalized vibrato model – using frequency, amplitude and spectral envelope modulations – with our own technique, we could determine that the HandSketch-based vibrato led to more complex harmonic trajectories, resulting in a more natural sound. Consequently a new vibrato model has been proposed, parametrizing the hysteresis that had been observed in HandSketch gestures.

Chapter 9

Conclusions

" I don't feel like I scratched the surface yet. "
— Jody Fisher

This Chapter concludes this thesis work. Its structure follows the five main axes explained in the Introduction, from [A1] to [A5]. Conclusions are presented as following: definition of realtime (cf. Section 9.1), analysis of vocal expressivity (cf. Section 9.2), resynthesis of expressive voice contents (cf. Section 9.3), description of voice quality dimensions (cf. Section 9.4) and the Analysis-by-Interation methodology (cf. Section 9.5).

Let us remember that these five axes come from the strong assumption that have been made in this thesis, concerning expressivity (cf. Section 1.3). We postulated that expressive qualities of a voice synthesis system mainly lead on its interactive capacities.

9.1 Definition of realtime [A1]

The interactivity of the RAMCESS system has been a transversal consideration throughout this thesis. The whole analysis/resynthesis process relies on GCI-centered frames with the length $2 \times T_0$. During the analysis, voice production model parameters are estimated locally and independently for each GCI-centered frame, systematically avoiding the use of non-causal or delay-prone algorithms (e.g. no parameter smoothing, no Viterbi optimization, etc). During the synthesis, the voice database is browsed frame by frame, and for each frame, these voice production model parameters can be modified

in realtime. Achieving various expressions with the RAMCESS system results from this highly interactive behavior.

9.2 Analysis of vocal expressivity [A2]

The main purpose of the RAMCESS analysis is the parameterization of the glottal flow signal over a large connected speech database. The state of the art in glottal flow analysis has been presented in Chapter 2. This thesis does not propose any new paradigm for the extraction of the glottal waveform from prerecorded voice signals. However, we used a pragmatic approach, by combining two promising existing algorithms, ZZT and ARX-LF, in order to reinforce the whole analysis process (cf. Chapter 3).

ZZT-based decomposition uses zeros of the Z-Transform of a given signal, in order to separate anticausal and causal components. Bozkurt has shown that, in specific windowing conditions, the anticausal component had great relevance in the estimation of the glottal formant. However ZZT-based decomposition is not robust to noise and very sensitive to GCI estimation. In this thesis, we examine this robustness issue and propose an optimization algorithm, based on the measurability of the glottal formant frequency.

Based on the glottal formant frequency, open quotient O_q and asymmetry coefficients α_M – parameters of the glottal flow anticausal component in the time domain – are estimated by fitting the LF model in the spectral domain.

These estimations of O_q and α_M, for each frame, are used as a way of reducing the complexity of the ARX-LF optimization. Indeed, ARX-LF is a modified version of the LP algorithm, using the LF model as the source component. The minimization of the prediction error is achieved over a large codebook of possible glottal flow derivative (GFD) waveforms. With the previous computation of O_q and α_M, the amount of waveforms to be tested is drastically decreased, as the only varying parameter is the return phase T_a.

Finally the RAMCESS analysis pipeline is evaluated. Considering the mean modeling error, RAMCESS remains less efficient than ARX-LF alone. Indeed ARX-LF also includes some refinements for the estimation of high frequencies, that have not been included in RAMCESS yet. However we propose new indicators, related to the stability (short and long term fluctuations) of extracted glottal source parameters. These indicators show that RAMCESS led to more expected means (considering usual values encountered with

a normal male voice), narrower variances, and smoother transitions of O_q and α_M values over the whole database, compared to ARX-LF alone.

9.3 Resynthesis of expressive voice contents [A3]

The RAMCESS synthesizer is based on the realtime convolution of the LF-based glottal source component and LP coefficients of the vocal tract, estimated in the analysis process and prerecorded in a database. Convolution is achieved frame by frame, and each frame can be queried (by choosing one GCI_k in the datbase) and modified in realtime.

Realtime interaction with all the voiced frames of the database is achieved thanks to two signal processing modules: the realtime generator of the GF/GFD signal and the interpolable tube-based vocal tract filter (cf. Chapter 4).

We propose a new model for the synthesis of the GF/GFD in realtime, called SELF (Spectrally Enhanced LF). The LF model appeared to rather be limited to several phonation types, typically the normal male voice. SELF computes the anticausal component of the glottal signal in the time domain, based on simplified LF equations. Then the return phase is processed in the spectral domain, and the range of variation of this return phase has been adapted, in order to propose a larger range of phonation, such as e.g. continuous transitions between quasi-sinusoidal voice and creaky voice.

The vocal tract filter is designed as a tube-based model, implemented as an all-pole lattice filter. LP coefficients, estimated by the RAMCESS analysis module, are converted into reflection coefficients, and then into relative area coefficients. These coefficients exhibit interesting interpolation properties, that are used to continuously interpolate vocal tract impulse responses, between consecutive frames queried by the user.

9.4 Description of voice quality dimensions [A4]

New mappings between voice quality dimensions and glottal flow parameters are proposed. Following the idea of realtime interaction, the continuous control space has been preferred to the usual classification of voice production into typical expressions: soft, tensed, creaky, etc. This assumption leads us to define several voice quality control spaces. In this thesis, we have proposed two different configurations (cf. Chapter 4).

The first mapping, called the "presfort" approach, aims at being appropriate for controlling voice quality with a limited amount of dimensions. This mapping gathers all the timbre transformations due to voice quality variation into one single axis, linearly interpolating the voice production parameters between "soft and lax" and "loud and tensed" phonations. This mapping is particularly appropriate for controlling voice quality with 3-axis controllers, such as 3D accelerometers, gloves, joysticks or graphic tablets.

The other proposed strategy aims at integrating a much larger amount of mechanisms encountered in voice quality variation. Several perceptual dimensions, such as vocal effort, tenseness and registers, are combined into one continuous control space. In particular, a realtime representation of the phonetogram (non-linear dependency between fundamental frequency, vocal effort and voice registers) is proposed.

9.5 Analysis-by-Interaction methodology [A5]

A quite unusual approach is used in the realization of this thesis work. Indeed, we do not follow the typical analysis/synthesis/control pipeline, as encountered in the prototyping of multimodal user interfaces. Our methodology is rather inspired by the traditional instrument making process (cf. Chapter 6).

We describe the validation protocols used separately in the design of both sound synthesizers and human/computer interaction devices. Examining the main drawbacks of this dislocated methodology, we formulate some recommendations, which leads us to propose a new approach for building digital instruments: the Luthery Model (LM).

LM aims at promoting the regular practice of the digital instrument, right from its prototyping. This approach relies on the theory of embodiment, which argues that only a strong embodiment of the object within the performer's body leads to high expressive

skills. LM develops these high expressive skills during the prototyping of the instrument, in order to use them for validation.

Based on the LM, the Analysis-by-Interaction (AbI) approach is proposed. This new methodology provides an alternate way of analyzing signals, by imitating them with an appropriate digital musical instrument. High expressive skills developed by the performer are used in order to reach a particularly convincing imitation of the signal. The imitated signal can then be studied by the analysis of imitative gestures.

In this thesis, we use AbI in the context of expressive voice modeling, and particularly for high quality singing synthesis. The use of LM leads us to develop the HandSketch, a tablet-based digital instrument (cf. Chapter 7). Then this instrument, or more precisely the analysis of performing gestures, are used to propose a new SELF-based vibrato model in singing (cf. Chapter 8).

Bibliography

[1] J. Accot and S. Zhai, "Performance evaluation of input devices in trajectory-based tasks: An application of the steering law," in *Proc. ACM Conference on Human Factors in Computing Systems*, 1999, pp. 466–472.

[2] M. Airas and P. Alku, "Emotions in Vowel Segments of Continuous Speech: Analysis of the Glottal Flow Using the Normalised Amplitude Quotient," *International Journal of Phonetic Science*, vol. 63, no. 1, pp. 26–46, 2006.

[3] O. Akanden and P. J. Murphy, "Improved Speech Analysis for Glottal Excited Linear Predictive Speech Coding," in *Proc. Irish Signals and Systems Conference*, 2004, pp. 101–106.

[4] P. Alku, "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulse Form," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 1992, pp. 29–32.

[5] ——, "Glottal Wave Analysis with Pitch Synchronous Iterative Adaptative Inverse Filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–117, 1992.

[6] P. Alku, H. Strik, and E. Vilkman, "Parabolic Spectral Parameter: a New Method for Quantification of the Glottal Flow," *Speech Communication*, vol. 22, pp. 67–79, 1997.

[7] P. Alku, J. Svec, E. Vilkman, and F. Sram, "Analysis of Voice Production in Breathy, Normal and Pressed Phonation by Comparing Inverse Filtering and Videokymography," in *Proc. International Conference on Spoken Language Processing*, 1999, pp. 885–888.

[8] P. Alku and E. Vilkman, "Estimation of the Glottal Pulseform Based on Discrete All-Pole Modeling," in *Proc. International Conference on Spoken Language Processing*, 1994, pp. 1619–1622.

[9] ——, "A Comparison of Glottal Voice Source Quantification Parameters in Breathy, Normal and Pressed Phonation of Female and Male Speakers," *Folia Phoniatrica et Logopaedica*, vol. 48, pp. 240–254, 1996.

[10] D. Arfib, J. M. Couturier, and L. Kessous, "Expressiveness and Digital Musical Instrument Design," *Journal of New Music Research, Special Issue on Expressive Gesture in Performing Arts and New Media*, vol. 34, no. 1, pp. 125–136, 2005.

[11] D. Arfib, J. M. Couturier, L. Kessous, and V. Verfaille, "Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces," *Organized Sound*, vol. 7, no. 2, pp. 127–144, 2002.

[12] I. Arroabarren and A. Carlosena, "Glottal Spectrum Based Inverse Filtering," in *Proc. European Conference on Speech Communication and Technology*, 2003, pp. 57–60.

[13] T. Backstrom, P. Alku, and E. Vilkman, "Time-Domain Parametrization of the Closing Phase of Glottal Airflow Waveform from Voices Over a Large Intensity Range," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 186–192, 2002.

[14] J. Barnes, P. Davis, J. Oates, and J. Chapman, "The Relationship Between Professional Operatic Soprano Voice and High Range Spectral Energy," *Journal of Acoustical Society of America*, vol. 116, no. 1, pp. 530–538, 2004.

[15] L. F. Barrett, "Solving the Emotion Paradox: Categorization and the Experience of Emotion," *Personality and Social Psychology Review*, vol. 10, no. 1, pp. 20–46, 2006.

[16] J. W. Beauchamp, "Analysis of Simultaneous Mouthpiece and Output Waveforms," *Journal of the AES*, no. 1626, pp. 1–11, 1980.

[17] R. Beaufort and A. Ruelle, "eLite : Systme de Synthse de la Parole Orientation Linguistique," in *Proc. Journées d'Études de la Parole*, 2006, pp. 509–512.

[18] M. Bellemare and C. Traube, "Verbal Description of Piano Timbre : Exploring Performer-Dependent Dimensions,," in *Proc. Conference on Interdisciplinary Musicology*, 2005.

[19] G. Berndtsson and J. Sundberg, "The MUSSE DIG Singing Synthesis," in *Proc. Stockholm Music Acoustic Conference*, no. 79, 1994, pp. 279–281.

[20] P. Birkholz, "Articulatory Synthesis of Singing," in *Proc. Interspeech*, 2007, pp. TuC.SS–1.

[21] A. W. Black and K. Tokuda, "The Blizzard Challenge: Evaluating Corpus-Based Speech synthesis on Common Datasets," in *Proc. Eurospeech*, 2005, pp. 77–80.

[22] L. Blin, O. Boeffard, and V. Barreaud, "Web-Based Listening Test System for Speech Synthesis and Speech Conversion Evaluation," in *International Conference on Language Resources and Evaluation*, 2008, pp. 2270–2274.

[23] J.-P. Blivet, *Les Voies du Chant.* Fayard, 1999.

[24] J. Bonanda and X. Serra, "Synthesis of the Singing Voice by Performance Sampling and Spectral Models," *Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[25] B. Bozkurt, "New Spectral Methods for Analysis of Source/Filter Characteristics of Speech Signals," Ph.D. dissertation, University of Mons, Mons, 2004.

[26] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp Group Delay Analysis of Speech Signals," *Speech Communication*, vol. 49, no. 3, pp. 159–176, 2007.

[27] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of the Z-Transform Representation with Application to Source-Filter Separation in Speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005.

[28] B. Bozkurt, F. Severin, and T. Dutoit, "An Algorithm to Estimate Anticausal Glottal Flow Component from Speech Signals," *Lecture Notes in Computer Science*, pp. 338–343, 2005.

[29] J. Bretos and J. Sundberg, "Measurements of Vibrato Parameters in Long Sustained Crescendo Notes as Sung by Ten Sopranos," *Journal of Voice*, vol. 17, no. 3, pp. 343–352, 2003.

[30] M. Bulut, S. Narayanan, and A. Syrdal, "Expressive Speech Synthesis Using a Concatenative Synthesizer," in *Proc. International Conference on Spoken Language Processing*, 2002.

[31] W. A. S. Buxton, "The Haptic Channel," *Human-Computer Interaction: A Multidisciplinary Approach*, pp. 357–365, 1987.

[32] J. T. Cacioppo, D. J. Klein, G. G. Berntson, and E. Hatfield, *The Psycho-Physiology of Emotion.* New York Guilford Press, 1993.

[33] C. Cadoz, "Instrumental Gesture and Musical Composition," in *Proc. International Computer Music Conference*, 1988, pp. 1–12.

[34] N. Campbell, "High-Definition Speech Synthesis," *Journal of Acoustical Society of America*, vol. 100, no. 4, p. 2850, 1996.

[35] ——, "Databases of Expressive Speech," *Journal of Chinese Language and Computing*, vol. 14, no. 4, 2004.

[36] S. K. Card, J. D. Mackinlay, and G. G. Robertson, "A Morphological Analysis of the Design Space of Input Devices," *ACM Transactions on Information Systems*, vol. 9, no. 2, pp. 99–122, 1991.

[37] M. Castellengo, B. Roubeau, and C. Valette, "Study of the Acoustical Phenomena Characteristic of the Transition Between Chest Voice and Falsetto," in *Proc. Stockholm Music Acoustic Conference*, vol. 1, 2002, pp. 113–123.

[38] C. Castillo, H. R. Hartson, and D. Hix, "Remote Usability Evaluation: Can Users Report their own Critical Incidents ?" in *Proc. ACM Conference on Human Factors in Computing Systems*, 1998, pp. 253–254.

[39] N. Chafai, C. Pelachaud, and D. Pelé, "A Case Study of Gesture Expressivity Breaks," *Language Resources and Evaluation*, vol. 41, no. 3, pp. 341–365, 2007.

[40] F. Charpentier and M. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 3, 1986, pp. 2015–2018.

[41] D. Childers and C. Ahn, "Modeling the Glottal Volume Velocity for Three Voice Types," *Journal of Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.

[42] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. Wiley and Sons, Inc., 1999.

[43] D. G. Childers and C. K. Lee, "Vocal Quality Factors: Analysis, Synthesis and Perception," *Journal of Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[44] V. Colotte and R. Beaufort, "Synthèse Vocale par Sélection Linguistiquement Orientée d'Unités Non-Uniformes : LiONS," in *Proc. Journées d'Études de la Parole*, 2004.

[45] P. Cook, "SPASM: A Real-Time Vocal Tract Physical Model Editor/Controller and Singer," *Computer Music Journal*, vol. 17, no. 1, pp. 30–44, 1992.

[46] ——, "Principles for Designing Computer Music Controllers," in *Proc. New Interfaces for Musical Expression*, 2001.

[47] D. C. Coplay, "A Stroboscopic Study of Lip Vibrations in a Trombone," *Journal of Acoustical Society of America*, vol. 99, pp. 1219–1226, 1996.

[48] C. d'Alessandro and M. Castellengo, "The Pitch in Short-Duration Vibrato Tones," *Journal of Acoustical Society of America*, vol. 95, no. 3, pp. 1617–1630, 1994.

[49] C. d'Alessandro, B. Doval, and K. Scherer, "Voice Quality: Functions, Analysis and Synthesis," in *Proc. ISCA ITRW VOQUAL*, 2003.

[50] N. d'Alessandro, O. Babacan, B. Bozkurt, T. Dubuisson, A. Holzapfel, L. Kessous, A. Moinet, and M. Vlieghe, "RAMCESS 2.x Framework - Expressive Voice Analysis for Realtime and Accurate Synthesis of Singing," *Journal of Multimodal User Interfaces*, vol. 2, no. 2, pp. 133–144, 2008.

[51] N. d'Alessandro, B. Bozkurt, R. Sebbe, and T. Dutoit, "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis," in *Proc. European Signal Processing Conference*, 2005.

[52] N. d'Alessandro, C. d'Alessandro, S. L. Beux, and B. Doval, "RealtimeCALM Synthesizer, New Approaches in Hands-Controlled Voice Synthesis," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 266–271.

[53] N. d'Alessandro, B. Doval, S. L. Beux, P. Woodruff, Y. Fabre, C. d'Alessandro, and T. Dutoit, "Realtime and Accurate Musical Control of Expression in Singing Synthesis," *Journal of Multimodal User Interfaces*, vol. 1, no. 1, pp. 31–39, 2007.

[54] N. d'Alessandro and T. Dutoit, "HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet," in *Proc. New Interfaces for Musical Expression*, 2007, pp. 78–81.

[55] ——, "RAMCESS/HandSketch: A Multi-Representation Framework for Realtime and Expressive Singing Synthesis," in *Proc. Interspeech*, 2007, pp. TuC.SS–5.

[56] J.-P. Dalmont, J. Gilbert, and S. Ollivier, "Nonlinear Characteristics of Single-Reed Instruments: Quasistatic Volume Flow and Reed Opening Measurements,"

Journal of Acoustical Society of America, vol. 114, pp. 2253–2262, 2003.

[57] I. U. E. Database, http://theremin.music.uiowa.edu/mis.html.

[58] A. de Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Journal of Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[59] P. Desain and H. Honing, "Modeling Continuous Aspects of Music Performance: Vibrato and Portamento," in *Proc. International Conference on Music Perception and Cognition*, 1996.

[60] P. Desain, H. Honing, R. Aarts, and R. Timmers, *Rhythm Perception and Production*. Lisse: Swets and Zeitlinger, 1999, ch. Rhythmic Aspects of Vibrato, pp. 203–216.

[61] K. Ding and H. Kasuya, "A Novel Approach to the Estimation of Voice Source and Vocal Tract Parameters from Speech Signals," in *Proc. International Conference on Spoken Language Processing*, 1996, pp. 1257–1260.

[62] C. Dobrian and D. Koppelman, "The E in NIME: Musical Expression with New Computer Interfaces," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 277–282.

[63] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: An analytic study," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 1997, pp. 446–452.

[64] ——, "The Spectrum of Glottal Flow Models," *Acta Acustica*, vol. 92, pp. 1026–1046, 2006.

[65] B. Doval, C. d'Alessandro, and N. Henrich, "The Voice Source as a Causal/Anticausal Linear Model," in *Proc. ISCA ITRW VOQUAL*, 2003, pp. 15–19.

[66] T. Drugman, B. Bozkurt, and T. Dutoit, "Chirp Decomposition of Speech Signals for Glottal Source Estimation," in *Proc. Non-Linear Speech Processing Conference*, 2009.

[67] T. Drugman, T. Dubuisson, A. Moinet, N. d'Alessandro, and T. Dutoit, "Glottal Source Estimation Robustness," in *Proc. IEEE International Conference on Signal Processing and Multimedia Applications*, 2008.

[68] T. Dubuisson and T. Dutoit, "Improvement of Source-Tract Decomposition of Speech Using Analogy with LF Model for Glottal Source and Tube Model for Vocal Tract," in *Proc. International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2007, pp. 119–122.

[69] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Springer (1st Edition), 2001.

[70] T. Dutoit and H. Leich, "MBR-PSOLA: Text to Speech Synthesis Based on a MBE Resynthesis of the Segments Database," *Speech Communication*, vol. 13, 1993.

[71] H. Duxans, A. Bonafonte, A. Kain, and J. V. Santen, "Including Dynamic and Phonetic Information in Voice Conversion Systems," in *Proc. International Conference on Spoken Language Processing*, 2004.

[72] C. dAlessandro, "Voice Quality in Vocal Communication: Tutorial," in *Proc. Interspeech*, 2007.

[73] N. dAlessandro, A. Moinet, T. Dubuisson, and T. Dutoit, "Causal/Anticausal Decomposition for Mixed-Phase Description of Brass and Bowed String Sounds," in *Proc. International Computer Music Conference*, vol. 2, 2007, pp. 465–468.

[74] A. Edelman and H. Murakami, "Polynomial Roots from Companion Matrix Eigenvalues," *Mathematics of Computation*, vol. 64, no. 210, pp. 763–776, 1995.

[75] M. Edgington and A. Lowry, "Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis," in *Proc. International Conference on Spoken Language Processing*, vol. 3, 1996, pp. 1425–1428.

[76] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[77] G. Fant, *Acoustic Theory of Speech Production*. Mouton and Co. Netherlands, 1960.

[78] ——, "The LF-Model Revisited, Transformations and Frequency Domain Analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 2004.

[79] S. Fels, "Intimacy and Embodiment: Implications for Art and Technology," in *Proc. ACM Workshops on Multimedia*, 2000, pp. 13–16.

[80] S. S. Fels, *Radial Basis Function Networks 2, New Advances in Design*. Physica-Verlag, 2001, ch. Using Radial Basis Functions to Map Hand Gestures to Speech, pp. 59–101.

[81] S. S. Fels, J. E. Lloyd, I. Stavness, F. Vogt, A. Hannam, and E. Vatikiotis-Bateson, "ArtiSynth: A 3D Biomechanical Simulation Toolkit for Modeling Anatomical Structures," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 14, no. 3, pp. 964–971, 2006.

[82] J. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer-Verlag (2nd Expanded Edition), 1972.

[83] Q. Fu and P. Murphy, "Adaptative Inverse Filtering for High Accuracy Estimation of the Glottal Source," in *Proc. Non-Linear Speech Processing Conference*, 2003, p. 13.

[84] ——, "Robust Glottal Source Estimation Based on Joint Source-Filter Model Optimization," *IEEE Transations on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[85] T. Galas and X. Rodet, "An Improved Cepstral Method for Deconvolution of Source-Filter Systems with Discrete Spectra: Application to Musical Sounds," in *Proc. International Computer Music Conference*, 1990, pp. 82–88.

[86] W. R. Garner, "Modeling and Quantization Techniques for Speech Compression Systems," Ph.D. dissertation, University of California, San Diego, 1994.

[87] D. B. Gerhard, "Computationally Measurable Differences Between Speech and Song," Ph.D. dissertation, Simon Fraser University, Burnaby, 2003.

[88] C. Gobl, "The Voice Source in Speech Communication," Ph.D. dissertation, KTH Speech, Music and Hearing, Stockholm, 2003.

[89] A. Group, http://www.acapela-group.com.

[90] Y. Guiard, "Disentangling Relative from Absolute Amplitude in Fitts Law Experiments," in *Proc. ACM Conference on Human Factors in Computing Systems*, 2001.

[91] Y. Guiard, M. Beaudouin-Lafon, and D. Mottet, "Navigation as a Multiscale Pointing, Extending Fitts Model to Very High Precision Tasks," in *Proc. ACM Conference on Human Factors in Computing Systems*, 1999, pp. 450–457.

[92] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM Expressive Speech Synthesis System," in *Proc. International Conference on Spoken Language Processing*, 2004.

[93] H. M. Hanson, "Glottal Characteristics of Female Speakers: Acoustic Correlates," *Journal of Acoustical Society of America*, vol. 101, pp. 466–481, 1997.

[94] H. M. Hanson and E. S. Chuang, "Individual variations in glottal characteristics of female speakers," *Journal of Acoustical Society of America*, vol. 106, no. 2, pp. 1064–1077, 1999.

[95] F. Hemke, *The Early History of the Saxophone*. University of Wisconsin-Madison, 1975.

[96] N. Henrich, "Étude de la source glottique en voix parlée et chantée," Ph.D. dissertation, Université Paris VI, France, 2001.

[97] ——, "Mirroring the Voice from Garcia to the Present Day: Some Insights into Singing Voice Registers," *Logopedics Phoniatrics Vocology*, vol. 31, pp. 3–14, 2006.

[98] N. Henrich, B. Doval, and C. d'Alessandro, "Glottal Open Quotient Estimation Using Linear Prediction," in *Proc. International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 1999.

[99] N. Henrich, B. Doval, C. d'Alessandro, and M. Castellengo, "Open Quotient Measurements on EGG, Speech and Singing Signals," in *Proc. International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, 2000.

[100] N. Henrich, C. dAlessandro, M. Castellengo, and B. Doval, "Glottal Open Quotient in Singing: Measurements and Correlation with Laryngeal Mechanisms, Vocal Intensity, and Fundamental Frequency," *Journal of Acoustical Society of America*, vol. 117, pp. 1417–1430, 2005.

[101] N. Henrich, C. dAlessandro, B. Doval, and M. Castellengo, "On the Use of the Derivative of Electroglottographic Signals for Characterization of Non-Pathological Phonation," *Journal of Acoustical Society of America*, vol. 115, pp. 1321–1332, 2004.

[102] N. Henrich, G. Sundin, and D. Ambroise, "Just Noticeable Differences of Open Quotient and Asymmetry Coefficient in Singing Voice," *Journal of Voice*, vol. 17, pp. 481–494, 2003.

[103] P. Herrera and J. Bonada, "Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework," in *Proc. Digital Audio Effects Conference*, 1998.

[104] Y. Horii, "Frequency Modulation Characteristics of Sustained /a/ Sung in Vocal Vibrato," *Journal of Speech and Hearing Research*, vol. 32, pp. 829–836, 1989.

[105] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 1996, pp. 373–376.

[106] A. Hunt and R. Kirk, "Mapping Strategies for Musical Performance - Trends in Gestural Control of Music," *Trends in Gestural Control of Music*, pp. 231–258, 2000.

[107] A. D. Hunt, M. Paradis, and M. Wanderley, "The Importance of Parameter Mapping in Electronic Instrument Design," *Journal of New Music Research, Special Issue on New Interfaces for Musical Performance and Interaction*, vol. 32, no. 4, pp. 429–440, 2003.

[108] M. Iseli and A. Alwan, "Inter- and Intra- Speaker Variability of Glottal Flow Derivative Using the LF Model," in *Proc. International Conference on Spoken Language Processing*, 2000, pp. 477–480.

[109] ——, "An Improved Correction Formula for the Estimation of Harmonic Magnitudes and its Application to Open Quotient Estimation," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. 669–672.

[110] L. B. Jackson, "Non-Causal ARMA Modeling of Voiced Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1606–1608, 1989.

[111] Y. Jiang and P. J. Murphy, "Production Based Pitch Modification of Voiced Speech," in *Proc. International Conference on Spoken Language Processing*, 2002, pp. 2073–2076.

[112] W. Johnson and T. J. Bouchard, "The Structure of Human Intelligence: it is Verbal, Perceptual and Image Rotation (VPR), Not Fluid and Crystallized," *Intelligence*, vol. 33, no. 4, pp. 431–444, 2004.

[113] P. Kabbash, W. Buxton, and A. Sellen, "Two-Handed Input in a Compound Task," in *Proceedings of the SIGCHI Conference*, 1995, pp. 417–423.

[114] B. F. G. Katz, F. Prezat, and C. d'Alessandro, "Human Voice Phoneme Directivity Pattern Measurements," *Journal of Acoustical Society of America*, vol. 120, no. 5, pp. 3359–3359, 2006.

[115] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based f0 Extraction: Possible Role of a Repetitive Structure in Sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[116] L. Kessous and D. Arfib, "Bi-Manuality in Alternate Musical Instruments," in *Proc. New Interfaces for Musical Expression*, 2003, pp. 140–145.

[117] L. Kessous, "Bi-Manual Mapping Experimentation, with Angular Fundamental Frequency Control and Sound Color Navigation," in *Proc. New Interfaces for Musical Expression*, 2002.

[118] Y. E. Kim, "Singing voice analysis, synthesis and modeling," *Handbook of Signal Processing in Acoustics*, pp. 359–374, 2008.

[119] J. Kjelland, *Orchestral Bowing: Style and Function.* Alfred Publishing Company, 2004.

[120] D. Klatt and L. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers," *Journal of Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[121] K. J. Kohler, "Macro and Micro F0 in the Synthesis of Intonation," *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pp. 115–138, 1990.

[122] R. Kortekaas and A. Kohlrausch, "Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech Waveform Manipulation Technique Using Single-Formant Stimuli," *Journal of Acoustical Society of America*, vol. 101, no. 4, pp. 2202–2213, 1997.

[123] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research," *Journal of Speech and Hearing Research*, vol. 36, pp. 21–40, 1993.

[124] Kyma, http://www.symbolicsound.com.

[125] O. Lähdeoja, "An approach to instrument augmentation: the electric guitar," in *Proc. New Interfaces for Musical Expression*, 2008.

[126] B. Larsson, "Music and Singing Synthesis Equipment (MUSSE)," *STL-QPRS*, vol. 18, no. 1, pp. 38–40, 1977.

[127] P. D. Lehrman, *MIDI for the Professional*. Music Sales America (1st Edition), 1993.

[128] J. M. López, R. Gil, R. García, I. Cearreta, and N. Garay, "Towards an Ontology for Describing Emotions," *Emerging Technologies and Information Systems for the Knowledge Society*, vol. 5288, pp. 96–104, 2008.

[129] Loquendo, http://www.loquendo.com.

[130] D. Lowry, *Bokken: Art of the Japanese Sword*. Black Belt Communications, 1986.

[131] H. L. Lu, "Toward a High-Quality Singing Synthesizer with Vocal Texture Control," Ph.D. dissertation, Stanford University, California, 2002.

[132] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George, "Concatenation-Based MIDI-to-Singing Voice Synthesis," in *Audio Engineering Society International Conference*, vol. 103, 1997.

[133] R. C. Maher and J. Beauchamp, "An Investigation of Vocal Vibrato for Synthesis," *Applied Acoustics*, vol. 30, pp. 219–245, 1990.

[134] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of IEEE*, vol. 63, pp. 561–580, 1975.

[135] J. Malloch and M. M. Wanderley, "The T-Stick: from Musical Interface to Musical Instrument," in *Proc. New Interfaces for Musical Expression*, 2007, pp. 66–69.

[136] S. Marchand and M. Raspaud, "Enhanced Time-Stretching using Order-2 Sinusoidal Modeling," in *Proc. Digital Audio Effects Conference*, 2004, pp. 76–82.

[137] G. Marino, M. H. Serra, and J. M. Raczinski, "The UPIC System: Origins and Innovations," *Perspectives of New Music*, vol. 31, no. 1, pp. 258–269, 1993.

[138] J. Martin, E. McKay, and L. Hawkins, "The Human-Computer Interaction Spiral," in *InSITE 2006*, 2006, pp. 183–196.

[139] Y. Meron, "High Quality Singing Synthesis using the Selection-Based Synthesis Scheme," Ph.D. dissertation, University of Tokyo, 1999.

[140] A. Momeni, "Composing Instruments: Inventing and Performing with Generative Computer-Based Instruments," Ph.D. dissertation, University of California, Berkeley, 2005.

[141] Mondofacto, http://www.mondofacto.com.

[142] F. R. Moore, "The Dysfunctions of MIDI," *Computer Music Journal*, vol. 12, no. 1, pp. 19–28, 1988.

[143] M. Mori, "The uncanny valley," *K. F. MacDorman and T. Minato, Trans.*, vol. 7, no. 4, pp. 33–35, 1970.

[144] P. Mulhem and L. Nigay, "Interactive Information Retrieval Systems: From User Centred Interface Design to Software Design," in *Proc. of SIGIR*, 1996, pp. 326–334.

[145] J. Mullen, D. M. Howard, and D. T. Murphy, "Waveguide Physical Modeling of Vocal Tract Acoustics: Flexible Formant Bandwidth Control from Increased Model Dimensionality," *Computer Music Journal*, vol. 17, no. 1, pp. 30–44, 1992.

[146] L. Nigay, J. Bouchet, D. Juras, B. Mansoux, M. Ortega, M. Serrano, and L. Lawson, *Multimodal User Interfaces: from Signals to Interaction*. Springer, 2008, ch. Software Engineering for Multimodal Interactive Systems, pp. 201–218.

[147] J. Ohala, "Ethological Theory and the Voice Expression of Emotion in the Voice," in *Proc. International Conference on Spoken Language Processing*, 1996.

[148] C. Oliveira, "Estimation of Source Parameters by Frequency Analysis," in *Proc. Eurospeech*, 1993, pp. 99–102.

[149] C. Ooge, "Keyboard-Based Singing Synthesis," 2008.

[150] K. K. Paliwal and L. Alsteris, "Usefulness of Phase Spectrum in Human Speech Perception," in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[151] M. D. Plumpe and T. F. Quatieri, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–585, 1999.

[152] C. Poepel, "On Interface Expressivity: a Player-Based Study," in *Proc. New In-*

terfaces for Musical Expression, 2005, pp. 228–231.

[153] E. Prame, "Vibrato Extent and Intonation in Professional Western Lyric Singing," *Journal of Acoustical Society of America*, vol. 102, no. 1, pp. 616–621, 1997.

[154] B. Pritchard and S. S. Fels, "GRASSP: Gesturally-Realized Audio, Speech and Song Performance," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 272–276.

[155] T. F. Quatieri and R. J. McAulay, "Shape-Invariant Time-Scale and Pitch Modification of Speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, 1992.

[156] M. Rahim, C. Goodyear, B. Kleijn, J. Schroeter, and M. Sondhi, "On the Use of Neural Networks in Articulatory Speech Synthesis," *Journal of Acoustical Society of America*, vol. 93, no. 2, pp. 1109–1121, 1993.

[157] E. L. Riegelsberger and A. K. Krishnamurthy, "Glottal Source Estimation: Methods of Applying the LF Model to Inverse Filtering," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 1993, pp. 542–545.

[158] X. Rodet, "Time-Domain Formant Wave Function Synthesis," *Computer Music Journal*, vol. 8, no. 3, pp. 9–14, 1984.

[159] X. Rodet, Y. Potard, and J. Barriere, "CHANT: de la Synthèse de la Voix Chantée la Synthèse en Général," *Rapports de recherche IRCAM*, no. 35, 1985.

[160] D. A. Rosenbaum, *Human Motor Control*. Academic Press, 1991.

[161] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette, "Vibrato: Detection, Estimation, Extraction, Modification," in *Proc. Digital Audio Effects Conference*, 1999.

[162] M. Russ, *Sound Synthesis and Sampling*. Focal Press, 1997.

[163] J. M. Rye and J. N. Holmes, "A Versatile Software Parallel-Formant Speech Synthesizer," *Joint Speech Research Unit Report*, no. 1016, 1982.

[164] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "HMM-Based Singing Voice Synthesis System," in *Proc. Interspeech*, 2000, pp. 1141–1144.

[165] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique

to Singing Voices," in *Proc. of IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, no. 10, 2007, pp. 215–218.

[166] S. Schiesser and C. Traube, "On Making and Playing an Electronically-Augmented Saxophone," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 308–313.

[167] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction: High-Quality Speech at Very Low Bit Rates," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 10, 1985, pp. 937–940.

[168] R. Schulman, "Articulatory Dynamics of Loud and Normal Speech," *Journal of Acoustical Society of America*, vol. 85, no. 1, pp. 295–312, 1989.

[169] D. Schwarz, "Data-Driven Concatenative Sound Synthesis," Ph.D. dissertation, Ircam - Centre Pompidou, Paris, 2004.

[170] C. E. Seashore, *The Vibrato*. University of Iowa studies, New Series, 1932, vol. 225.

[171] S. Serafin, F. Avanzini, and D. Rocchesso, "Bowed String Simulation Using an Elasto-Plastic Friction Model," in *Proc. Stockholm Music Acoustic Conference*, 2003.

[172] X. Serra and J. Bonada, "Sound Transformations Based on the SMS High Level Attributes," in *Proc. Digital Audio Effects Conference*, 1998.

[173] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: a Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[174] R. B. Sexton and D. Haussner, *Method for the Theremin Bk. I*. Tactus Press, 1996.

[175] R. N. Shepard, "Circularity in Judgements of Relative Pitch," *Journal of Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[176] R. H. Siminoff, *The Luthier's Handbook: A Guide to Building Great Tone in Acoustic Stringed Instruments*. Hal Leonard, 2002.

[177] J. O. Smith, "Physical Modeling using Digital Waveguides," *Computer Music Journal*, vol. 16, no. 4, pp. 74–91, 1992.

[178] M. Södersten and P. A. Lindestad, "Glottal Closure and Perceived Breathiness

During Phonation in Normal Speaking Subjects," *Journal of Speech and Hearing Research*, vol. 33, pp. 601–611, 1990.

[179] H. C. Steiner, "Towards a Catalog and Software Library of Mapping Methods," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 106–109.

[180] H. Strik, "Automatic Parametrization of Differentiated Glottal Flow: Comparing Methods by Means of Synthetic Flow Pulses," *Journal of Acoustical Society of America*, vol. 103, no. 5, pp. 2659–2669, 1998.

[181] H. Strik, B. Cranen, and L. Boves, "Fitting a LF-Model to Inverse Filter Signals," in *Proc. Eurospeech*, vol. 1, 1993, pp. 103–106.

[182] Y. Stylianou, "Concatenative Speech Synthesis Using a Harmonic Plus Noise Model," in *Proc. ESCA/COCOSDA Workshop on Speech Synthesis*, 1998, pp. 261–266.

[183] ——, "Voice Quality Compensation System for Speech Synthesis Based on Unit Selection Speech Database," March 1999, uS Patent 6266638.

[184] D. Suendermann, A. Bonafonte, H. Duxans, and H. Hoege, "Tc-Star: Evaluation Plan for Voice Conversion Technology," in *German Annual Conference on Acoustics*, 2005.

[185] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[186] J. Sundberg, I. R. Titze, and R. C. Scherer, "Phonatory Control in Male Singing: a Study of the Effects of Subglottal Pressure, Fundamental Frequency, and Mode of Phonation on the Voice Source," *Journal of Voice*, vol. 7, pp. 15–29, 1993.

[187] R. Timmers and P. Desain, "Vibrato: Questions and Answers from Musicians and Science," in *Proc. International Conference on Music Perception and Cognition*, 2000.

[188] I. Titze, T. Riede, and P. Popolo, "Nonlinear Source-Filter Coupling in Phonation: Vocal Exercises," *Journal of Acoustical Society of America*, vol. 123, no. 4, pp. 1902–1915, 2008.

[189] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2000,

pp. 1315–1318.

[190] C. Traube and P. Depalle, "Timbral Analogies Between Vowels and Plucked String Tones," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 4, 2004, pp. 293–296.

[191] H. Traunmuller and A. Eriksson, "Acoustic Effects of Variation in Vocal Effort by Men, Women, and Children," *Journal of Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[192] V. Verfaille, C. Guastavino, and P. Depalle, "Perceptual Evaluation of Vibrato Models," in *Proc. Conference on Interdisciplinary Musicology*, 2005, pp. 1–19.

[193] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF Glottal Source Parameters Based on ARX Model," in *Proc. International Conference on Spoken Language Processing*, 2005, pp. 333–336.

[194] ——, "A New Method for Speech Synthesis and Transformation Based on an ARX-LF Source-Filter Decomposition and HNM Modeling," in *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2007, pp. 525–528.

[195] Vocaloïd, http://www.vocaloid.com.

[196] Wacom, http://www.wacom.com.

[197] J. Walker and P. Murphy, "A Review of Glottal Waveform Analysis," *Progress in Nonlinear Speech Processing*, vol. 4391, pp. 1–21, 2007.

[198] M. M. Wanderley, N. Orio, and N. Schnell, "Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI," *Computer Music Journal*, vol. 26, no. 3, pp. 62–76, 2002.

[199] M. Wanderly and P. Depalle, "Gestural Control of Sound Synthesis," *Proceedings of the IEEE : Special Issue on Engineering and Music - Supervisory Control and Auditory Communication*, vol. 92, no. 4, pp. 632–644, 2004.

[200] D. Wessel, M. Wright, and S. A. Khan, "Preparation for Improvised Performance in Collaboration with a Khyal Singer," in *Proc. International Computer Music Conference*, 1998, pp. 497–503.

[201] D. Y. Wong, J. D. Markel, J. Augustine, and H. Gray, "Least Square Glottal Inverse Filtering from the Acoustic Waveform," *IEEE Transactions on Acoustics,*

Speech and Signal Processing, vol. 27, pp. 350–355, 1979.

[202] A. M. Woodhull, K. Maltrud, and B. L. Mello, "Alignment of the Human Body in Standing," *European Journal of Applied Physiology*, vol. 54, no. 1, pp. 109–115, 1985.

[203] R. Woof, *Technique and Interpretation in Violin-Playing*. Read Country Books, 2006.

[204] M. Zbyszynski, M. Wright, A. Momeni, and D. Cullen, "Ten Years of Tablet Musical Interfaces at CNMAT," in *Proc. New Interfaces for Musical Expression*, 2006, pp. 100–105.

List of Figures

List of Tables