HAL open science

Glottal source and vocal-tract separation

Gilles Degottex

To cite this version:

Gilles Degottex. Glottal source and vocal-tract separation. Signal and Image processing. Université Pierre et Marie Curie - Paris VI, 2010. English. NNT : . tel-00554763v2

HAL Id: tel-00554763

https://theses.hal.science/tel-00554763v2

Submitted on 17 Nov 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Glottal source and vocal-tract separation

Estimation of glottal parameters, voice transformation and synthesis using a glottal model

Gilles Degottex

Presented on November the 16th 2010, to obtain the grade of Docteur de l'Université Paris VI - Pierre et Marie Curie (UPMC) with specialization in signal processing from the Ecole Doctorale Informatique, Télécommunications et Electronique (EDITE), to the following jury:

Chairman	Thierry Dutoit	Professor - Faculté Polytechnique de Mons
Reviewers	Yannis Stylianou	Professor - University of Crete
	Christophe d'Alessandro	CNRS Research Director - LIMSI
Examiners	Nathalie Henrich	CNRS Researcher - Grenoble INP
	Olivier Rosec	Research Engineer - Orange Labs
	Jean-Luc Zarader	Professor - UPMC
	Olivier Boëffard	Professor - University of Rennes
Thesis director	Xavier Rodet	Emeritus Researcher - Ircam

This work has been supervised by Axel Röbel and Xavier Rodet, realized at the Institut de Recherche et Coordination Acoustique/Musique (Ircam) and partially funded by a grant of the Centre National de la Recherche Scientifique (CNRS-UMR9912-STMS).

Ircam - CNRS-UMR9912-STMS
Sound Analysis/Synthesis Team
1, place Igor Stravinsky
75004 Paris, FRANCE

December 7, 2010

Final version

The content of the companion CD is currently available at:
`http://recherche.ircam.fr/anasyn/degottex/index.php?n=Main.ExDegottex2010`

Abstract

This study addresses the problem of inverting a voice production model to retrieve, for a given recording, a representation of the sound source which is generated at the glottis level, the glottal source, and a representation of the resonances and anti-resonances of the vocal-tract. This separation gives the possibility to manipulate independently the elements composing the voice. There are many applications of this subject like the ones presented in this study, namely voice transformation and speech synthesis, as well as many others such as identity conversion, expressivity synthesis, voice restoration which can be used in entertainment technologies, artistic sound installations, movies and music industry, toys and video games, telecommunication, etc.

In this study, we assume that the perceived elements of the voice can be manipulated using the well known source-filter model. In the spectral domain, voice production is thus described as a multiplication of the spectra of its elements, the glottal source, the vocal-tract filter and the radiation. The second assumption used in this study concerns the deterministic component of the glottal source. Indeed, we assume that a glottal model can fit one period of the glottal source. Using such an analytical description, the amplitude and phase spectra of the deterministic source are linked through the shape parameter of the glottal model. Regarding the state of the art of voice transformation and speech synthesis methods, the naturalness and the control of the transformed and synthesized voices should be improved. Accordingly, we try to answer the three following questions: 1) How to estimate the parameter of a glottal model? 2) How to estimate the vocal-tract filter according to this glottal model? 3) How to transform and synthesize a voiced signal using this glottal model?

Special attention is given to the first question. We first assume that the glottal source and the impulse response of the vocal-tract filter are mixed-phase and minimum-phase signals respectively. Then, based on these properties, various methods are proposed which minimize the mean squared phase of the convolutive residual of an observed spectrum and its model. A last method is described where a unique shape parameter is in a quasi closed-form expression of the observed spectrum. Additionally, this study discusses the conditions a glottal model and its parametrization have to satisfy in order to ensure that the parameters estimation is reliable using the proposed methods. These methods are also evaluated and compared to state of the art methods using synthetic and electroglottographic signals. Using one of the proposed methods, the estimation of the shape parameter is independent of the position and the amplitude of the glottal model. Moreover, it is shown that this same method outperforms all the compared methods.

To answer the second and third questions addressed in this study, we propose an analysis/synthesis procedure which estimates the vocal-tract filter according to an observed spectrum and its estimated source. Preference tests have been carried out and their results are presented in this study to compare the proposed procedure to existing ones. In terms of pitch transposition, it is shown that the overall quality of the voiced segments of a recording can be improved for important transposition factors. It is also shown that the breathiness of a voice can be controlled.

Keywords: voice separation, glottal model, shape parameter estimation, mean squared phase, voice transformation, speech synthesis.

Resumé

Cette étude s'intéresse au problème de l'inversion d'un modèle de production de la voix pour obtenir, à partir d'un enregistrement audio de parole, une représentation de le source sonore qui est générée au niveau de la glotte, la source glottique, ainsi qu'un représentation des résonances et anti-résonances créées par le conduit vocal. Cette séparation permet de manipuler les éléments composant la voix de façon indépendente. On trouve de nombreuses applications de ce sujet comme celles présentées dans cette étude (transformation de la voix et synthèse de la parole) et bien d'autres comme la conversion d'identité, la synthèse d'expressivité, la restauration de la voix qui peuvent être utilisées dans les technologies de divertissement, des installations sonores, les industries de la musique et du cinéma, les jeux vidéos et autres jouets sonores, la télécommunication, etc.

Dans cette étude, nous supposons que les éléments perçus de la voix peuvent être manipulés en utilisant le modèle source-filtre. Dans le domaine spectral, la production de la voix est donc décrite comme une multiplication des spectres de ses éléments, la source glottique, le filtre du conduit vocal et la radiation. La seconde hypothèse utilisée dans cette étude concerne la composante déterministe de la source glottique. En effet, nous supposons qu'un modèle glottique peut schématiser une période de la source glottique. En utilisant une telle description analytique, les spectres d'amplitude et de phase de la source déterministe sont donc liés par les paramètres de forme du modèle glottique. Vis-à-vis de l'état de l'art des méthodes de transformation de la voix et de sa synthèse, le naturel et le contrôle de ces voix devraient donc être améliorés en utilisant un tel modèle. Par conséquent, nous essayons de répondre au trois questions suivantes dans cette étude : 1) Comment estimer un paramètre de forme d'un modèle glottique. 2) Comment estimer le filtre du conduit vocal en utilisant ce modèle glottique. 3) Comment transformer et synthétiser un signal vocal en utilisant toujours ce même modèle.

Une attention toute particulière à été portée à la première question. Premièrement, nous supposons que la source glottique est un signal à phase mixte et que la réponse impulsionnelle du filtre du conduit vocal est un signal à minimum de phase. Puis, considérant ces propriétés, différentes méthodes sont proposées qui minimisent la phase carrée moyenne du résiduel convolutif d'un spectre de parole observé et de son modèle. Une dernière méthode est décrite où un unique paramètre de forme est solution d'une forme quasi fermée du spectre observé. De plus, cette étude discute les conditions qu'un modèle glottique et sa paramétrisation doivent satisfaire pour assurer que les paramètres sont estimés de façon fiable en utilisant les méthodes proposées. Ces méthodes sont également évaluées et comparées avec des méthodes de l'état de l'art en utilisant des signaux synthétiques et electro-glotto-graphiques. En utilisant une des méthodes proposées, l'estimation du paramètre de forme est indépendante de la position et de l'amplitude du modèle glottique. En plus, il est montré que cette même méthode surpasse toute les méthodes comparées en terme d'efficacité.

Pour répondre à la deuxième et à la troisième question, nous proposons une procédure d'analyse/synthèse qui estime le filtre du conduit vocal en utilisant un spectre observé et sa source estimée. Des tests de préférences ont été menés et leurs résultats sont présentés dans cette étude pour comparer la procédure décrite et d'autres méthodes existantes. En terme de transposition de hauteur perçue, il est montré que la qualité globale des segments voisés d'un enregistrement peut être meilleure pour des facteurs de transposition importants en utilisant la méthode proposée. Il est aussi montré que le souffle perçu d'une voix peut être contrôlé efficacement.

Mots-clés : séparation de la voix, modèle glottique, estimation de paramètres de forme, phase carrée moyenne, transformation de la voix, synthèse de la parole.

Remerciements

Une thèse ne se fait pas seul, ou en tout cas, ce n'était pas le cas de celle-ci. Je ne peux donc que couvrir de remerciements les personnes suivantes qui ont rendu ce travail d'abord possible, puis agréable mais surtout passionnant.

Xavier Rodet et Axel Röbel pour m'avoir accueilli dans leur équipe, m'avoir écouté débuter dans la parole, le traitement du signal et toutes leurs phases, pour toutes ces discussions passionnantes de connaissances, de réflexions et d'autres jeux cérébraux, pour leur temps et leur attention.

Erkki Bianco, Laurier Fagnan, Mette Pederson pour ces expériences inoubliables de spéléologie glottique en ultra-strobo-endo-scopo-electro-glotto-video-graphie ainsi que Emmanuel Fléty pour son aide dans la connexion de tous ces termes, les entreprises *Richard Wolf* et *F-J Electronics* pour le prêt des outils mis en jeux. Une pensée également à tous ces chanteurs ainsi qu'à Mikou qui nous ont prêté quelques vocalises le temps d'immortaliser leurs cordes.

L'Ircam et ses collaborateurs passionnés, l'équipe Anasynth, Thomas Hélie, Rémi Mignot, Chunghsin Yeh, & co. et tout particulièrement le fabuleux groupe Voix, Snorre Farner, Nicolas Obin, Christophe Veaux, Pierre Lanchantin, mon ami de mezzanine et de bouts de ficelles Grégory Beller, ainsi que les nombreux stagiaires qui ont défilés durant ces quatre années, avec une pensée toute particulière pour Pierre ainsi que Baptiste Bohelay pour leurs chaleureuses collaborations dans la synthèse et la transformation de la voix.

Infinite acknowledgments to all the English speakers who so much suffered from my infinitely progressing bounded English and saved me from drowning into the language of sciences, Leigh Smith, Michael Sweeton, Ashleigh Gonzales and Paul Ladyman.

Encore une fois les ircamiens et près d'une centaine de personnes à travers cette belle planète ayant répondu à mes infatigables relances de test de préférences.

Les personnes qui m'ont apporté tous ces conseils, remarques constructives et corrections lors de la rédaction de cette thèse, Βασιλική Ζαχάρη, Snorre, Juan José Burred, Christophe Charbuillet, Charles Françoise, Thomas. Ainsi que Christophe d'Alessandro, Nathalie Henrich et Thierry Dutoit pour leur temps et leur commentaires qui ont évité confusions et boulettes autant que éclaircit la version finale de ce manuscript.

Toutes ces amitiés qui m'ont fait découvrir la fourmillière parisienne, Bruno, Frédérique, Bruno, ces amitiés qui m'ont fait tant apprécier le *selecta* du -2, Pepette, Pipof et Ciccio, encore une fois ce dernier ainsi que Marta Gentilucci pour avoir tant insisté à me faire découvrir la musique electro-acoustico-contemporaine. The incognitos J.J., L.S. & P.L. pour ces heures de musique non-ircamiennes dans un local très ircamien.

Et bien entendu mon ami, collègue et coloc' de bar, et de taf', que j'aurais pu citer dans presque tous les paragraphes ci-dessus, Marcelaoudje Caetano.

Džiloš de Glottex

à Louis,

Contents

Abstract & Resumé **3**

Notations, acronyms and usual expressions **15**

1 Introduction **19**
1.1 Problematics . . . 19
1.1.1 Source-filter model vs. acoustic model . . . 22
1.1.2 The chosen approach: the glottal model . . . 23
1.1.3 Evaluation and validation of the proposed methods . . . 24
1.2 Structure of the document . . . 25

I Voice production and its model **27**

2 The glottal source **29**
2.1 Vocal folds & glottal area . . . 29
2.2 Laryngeal mechanisms & voice quality . . . 32
2.3 Glottal flow vs. Glottal source . . . 33
2.4 Glottal models . . . 35
2.5 Time and spectral characteristics of the glottal pulse . . . 39
2.5.1 Time properties: glottal instants and shape parameters . . . 40
2.5.2 Spectral properties: glottal formant and spectral tilt . . . 41
2.5.3 Mixed-phase property of the glottal source . . . 42
2.5.4 Vocal folds asymmetry, pulse scattering and limited band glottal model . . . 42
2.6 Aspiration noise . . . 44
Conclusions . . . 45

3 Filtering elements and voice production model **47**
3.1 The Vocal-Tract Filter (VTF) . . . 47
3.1.1 Structures of the vocal-tract . . . 47
3.1.2 Minimum-phase hypothesis . . . 49
3.2 Lips and nostrils radiation . . . 50
3.3 The complete voice production model . . . 52
Conclusions . . . 54

II Voice analysis **55**

4 Source-filter separation **57**
4.1 General forms of the vocal-tract filter and the glottal source . . . 57
4.2 Estimation of glottal parameters . . . 58
4.3 Spectral envelope models . . . 60
4.4 The state of the art . . . 62
4.4.1 Analysis-by-Synthesis . . . 62
4.4.2 Pre-emphasis . . . 62
4.4.3 Closed-phase analysis . . . 64
4.4.4 Pole or pole-zero models with exogenous input (ARX/ARMAX) . . . 64
4.4.5 Minimum/Maximum-phase decomposition, complex cepstrum and ZZT . . . 66
4.4.6 Inverse filtering quality assessment . . . 66
Conclusions . . . 67

5 Joint estimation of the pulse shape and its position **69**
5.1 Phase minimization . . . 69
5.1.1 Conditions of convergence . . . 72
5.1.2 Measure of confidence . . . 72
5.1.3 Polarity . . . 73
5.1.4 The iterative algorithm using MSP . . . 74
5.2 Difference operator for phase distortion measure . . . 76
5.2.1 The method using MSPD . . . 77
5.3 Estimation of multiple shape parameters . . . 77
5.3.1 O_q/α_m vs. I_q/A_q . . . 82
Conclusions . . . 83

6 Estimation of the shape parameter without pulse position **85**
6.1 Parameter estimation using the 2nd order phase difference . . . 85
6.1.1 The method based on MSPD^2 . . . 85
6.2 Parameter estimation using function of phase-distortion . . . 87
6.2.1 The method FPD^{-1} based on FPD inversion . . . 87
6.2.2 Conditioning of the FPD inversion . . . 89
Conclusions . . . 94

7 Estimation of the Glottal Closure Instant **95**
7.1 The minimum of the radiated glottal source . . . 97
7.2 The method using a glottal shape estimate . . . 99
7.2.1 Estimation of a GCI in one period . . . 99
7.2.2 Estimation of GCIs in a complete utterance . . . 99
7.3 Evaluation of the error related to the shape parameter . . . 103
Conclusions . . . 104

8 Evaluation of the proposed estimation methods **105**
8.1 Evaluation with synthetic signals . . . 105
8.1.1 Error related to the fundamental frequency . . . 106
8.1.2 Error related to the noise levels . . . 108
8.2 Comparison with electroglottographic signals . . . 111
8.2.1 Evaluation of GCI estimates . . . 112
8.2.2 Evaluation of the shape parameter estimate . . . 113
8.3 Examples of *Rd* estimates on real signals . . . 115
Conclusions . . . 117

III Voice transformation and speech synthesis **121**

9 Analysis/synthesis method **121**
9.1 Methods for voice transformation and speech synthesis . . . 121
9.2 Choice of approach for the proposed method . . . 124
9.3 The analysis step: estimation of the SVLN parameters . . . 124
9.3.1 The parameters of the deterministic source: f_0, Rd, E_e . . . 124
9.3.2 The parameter of the random source: σ_g . . . 125
9.3.3 The estimation of the vocal-tract filter . . . 125
9.4 The synthesis step using SVLN parameters . . . 127
9.4.1 Segment position and duration . . . 127
9.4.2 The noise component: filtering, modulation and windowing . . . 127
9.4.3 Glottal pulse and filtering elements . . . 129
Conclusions . . . 130

10 Evaluation of the SVLN method **131**
10.1 Influence of the estimated parameters on the SVLN results . . . 131
10.2 Preference tests for pitch transposition . . . 132
10.3 Evaluation of breathiness modification . . . 136
10.4 Speech synthesis based on Hidden Markov Models (HMM) . . . 137
Conclusions . . . 139

11 General conclusions **141**
11.1 Future directions . . . 143

A Minimum, zero and maximum phase signals **145**
A.1 The real cepstrum and the zero and minimum phase realizations . . . 146
A.2 Generalized realization . . . 147

B Error minimization and glottal parameters estimation **151**
B.1 Error minimization . . . 151
B.2 Parameter estimation based on glottal source estimate . . . 152
B.2.1 Procedure for the Iterative Adaptive Inverse Filtering (IAIF) . . . 152
B.2.2 Procedure for the Complex Cepstrum (CC) and the Zeros of the Z-Transform (ZZT) 153

C The glottal area estimation method **155**
C.1 The proposed method for glottal area estimation 155
C.2 Spectral aliasing 159
C.3 Equipment 159

D Maeda's voice synthesizer **161**
D.1 Space and time sampling 161
D.2 Minimum-phase property of the vocal-tract 163

Bibliography **167**

Publications during the study **181**

Notations

/x/	:	Phoneme given in IPA (International Phonetic Alphabet).
$\tilde{x}$	:	Approximation of x.
f_s	:	The sampling frequency of a discrete signal.
f_0	:	The fundamental frequency in Hz of a periodic signal.
$T_0 = 1/f_0$	:	The fundamental period.
ω_0	:	The fundamental frequency in radians.
$\overline{X}$	:	The complex conjugate of X
$x(t)$	:	Continuous signal with respect to the time t
$x[n]$	:	Discrete signal with respect to the sample n
$x(t) \circledast y(t)$	:	The convolution of $x(t)$ and $y(t)$
$\delta(t)$	:	The Dirac delta function.
$X[k]$	:	Discrete spectrum with respect to the bin k
$X(\omega)$	:	Continuous spectrum with respect to the circular frequency ω
$\mathcal{F}(x)$	:	Depending on the context (input and output),
		the continuous Fourier Transform (FT),
		the Discrete Time Fourier Transform (DTFT) or
		the Discrete Fourier Transform (DFT)
		(its inverse: $\mathcal{F}^{-1}(X)$)
X_h	:	Discrete spectrum with respect to the harmonic h ($X_h = X(h \cdot \omega_0)$)
$X_-(\omega)$	:	Spectrum of a minimum-phase signal (e.g. the vocal-tract filter).
$X_+(\omega)$	:	Spectrum of a maximum-phase signal (e.g. the Rosenberg glottal model).
$\tilde{x}[n]$	:	Real cepstrum of $x[n]$ ($\tilde{x}[n] = \mathcal{F}^{-1}\big(log\lvert\mathcal{F}(x[n])\rvert\big)$)
$\mathcal{E}(.)$	:	Envelope estimation (e.g. LP, DAP, TE) using a given order o.
$\theta^* + \Delta\theta = \theta$	:	An estimated parameter is equal to the optimal parameter and an additional error.
$e^{j\omega\phi}$	:	Linear-phase term (e.g. the one which controls the position of a glottal pulse).
$G^{\theta}(\omega)$	:	Glottal model parametrized by θ
		(e.g. the Liljencrants-Fant model parametrized by Rd is: $G^{Rd}(\omega)$)
$G^{\Delta\theta}(\omega)$	:	Spectral error due to an error of the glottal parameter θ

Acronyms

AR AutoRegressive

ARMA AutoRegressive and Moving Average

DAP Discrete All-Pole (an AR envelope estimation method)

DFT Discrete Fourier Transform

DTFT The Discrete Time Fourier Transform

EGG ElectroGlottoGraphy

FPD Function of Phase-Distortion

FT Fourier Transform

GCI Glottal Closure Instant (t_e in the LF model)

GOI Glottal Opening Instant (t_s in the LF model)

HMM Hidden Markov Model

HNM Harmonic+Noise Model

HSV High Speed Videoendoscopy

IQR Inter-Quartile Range

LF Liljencrants-Fant glottal model

LFRd Transformed-LF glottal model (LF glottal model parametrized by R_d)

LP Linear Prediction

MSP Mean Squared Phase

PCG Preconditioned Conjugate Gradient (a local search algorithm)

PSOLA Pitch Synchronous Overlap-Add.

RMS$_N$ Root Mean Squared limited to the first N values.

STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum [KMKd99].

SVLN Separation of the Vocal-tract with a Liljencrants-fant model + Noise

VTF Vocal-Tract Filter

VUF Voiced/Unvoiced Frequency (also known as Maximum Voiced Frequency)

WBVPM Wide-Band Voice Pulse Modeling [Bon08]

Usual expressions

Glottal source
Both deterministic and stochastic components of the source of the source-filter model. The glottal source usually covers a multiple number of periods.

Glottal pulse
The shape of the deterministic source in a single period.

Glottal model
A schematized definition of the glottal pulse.

Shape parameter
A parameter which defines the shape of a glottal model only. This parameter is thus independent on the duration and amplitude of the glottal pulse.

Convolutive residual
The result of the deconvolution of a signal by its model.
In spectral domain: $R(\omega) = S(\omega)/M(\omega)$

Additive residual
The difference between a signal and its model: $E(\omega) = S(\omega) - M(\omega)$

Analysis/synthesis method
Given a signal model and its parameters, a method which can entirely encode a voice recording into a sequence of parameters. Given this sequence, the signal can then be resynthesized or transformed, by changing or not the parameters sequence between the analysis and synthesis steps. Using such a method, the observed signal is fully modeled.

Modification method
A method which partly changes a voice recording. Using such a method, the observed signal is partly modeled.

HMM-based synthesis
Parametric speech synthesis using Hidden Markov Models.

Chapter 1

Introduction

Voice production is made of various elements which either generate a sound source or modify this latter by emphasizing or reducing some frequency bands by means of filtering effects like resonances. Once the resulting sound propagates outside of the vocal apparatus, some part is perceived as semantic content whereas the other part is related to the timbre of the voice, its granularity, its color, which is termed *voice quality*. By separating the glottal source and the vocal-tract, the goal is to inverse the voice production from a given recorded waveform, retrieving thus the voice source generated at the glottis level and a representation of the vocal-tract characteristics. Separating these elements is very attractive for voice processing. Indeed, it is a way to analyze, study and understand the properties of voice production. In addition, as addressed in this study, a voice can be transformed and synthesized using independent manipulation of its elements. For example, using a proper separation, the voice quality of a speech recording can be modified while keeping all the other components of the voice untouched. The modification of the voice quality is of large interest to obtain various personalities, identities and expressiveness from a neutral voice in both voice transformation and speech synthesis. Consequently, the main purpose of this work is to study how to obtain such a separation in order to improve the current voice processing techniques.

1.1 Problematics

Languages and singing styles cover a variety of sounds in a manner that voice analysis in its most general sense will not be handled in this study. We will therefore address the voiced sounds only since the voice quality is mainly dependent on this component. Additionally, this subject covers many research areas and can be approached by many point of views. As argued in the next section, a signal processing approach using a linear model of the voice production has been chosen, which is currently widely used in speech technologies.

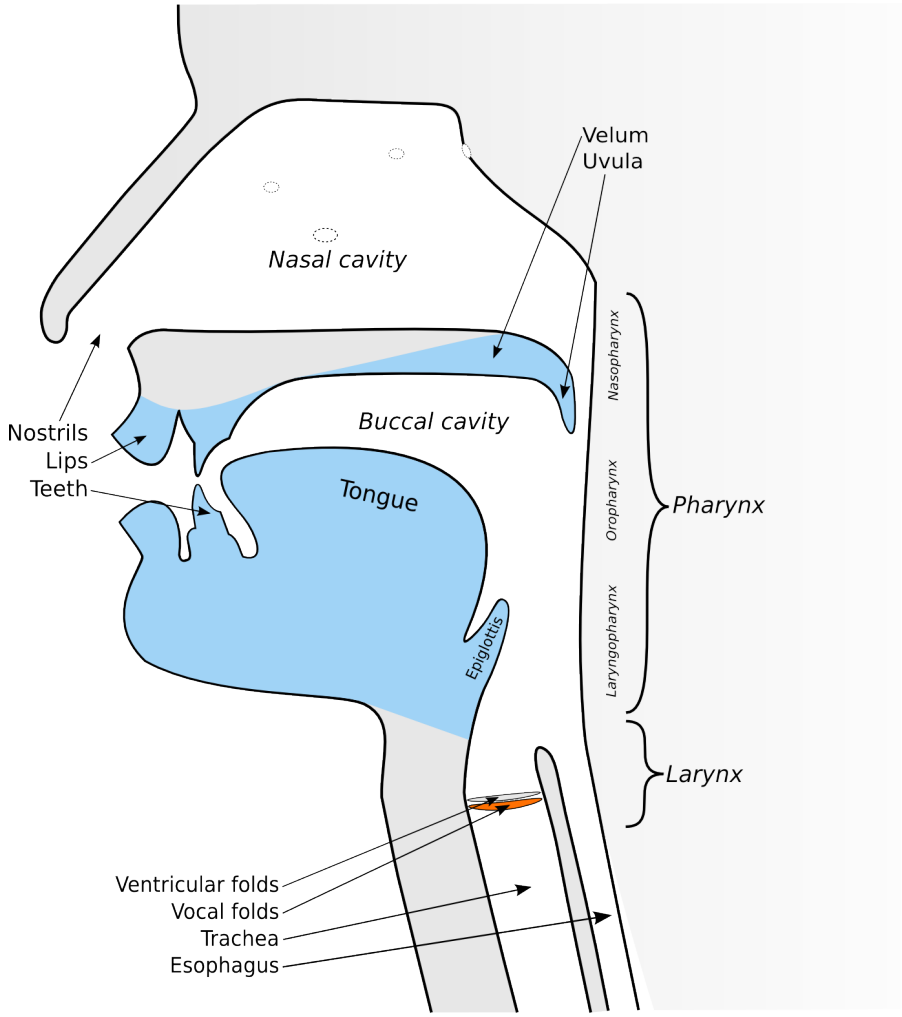


Figure 1.1: Schematic view of a profile cut of the head

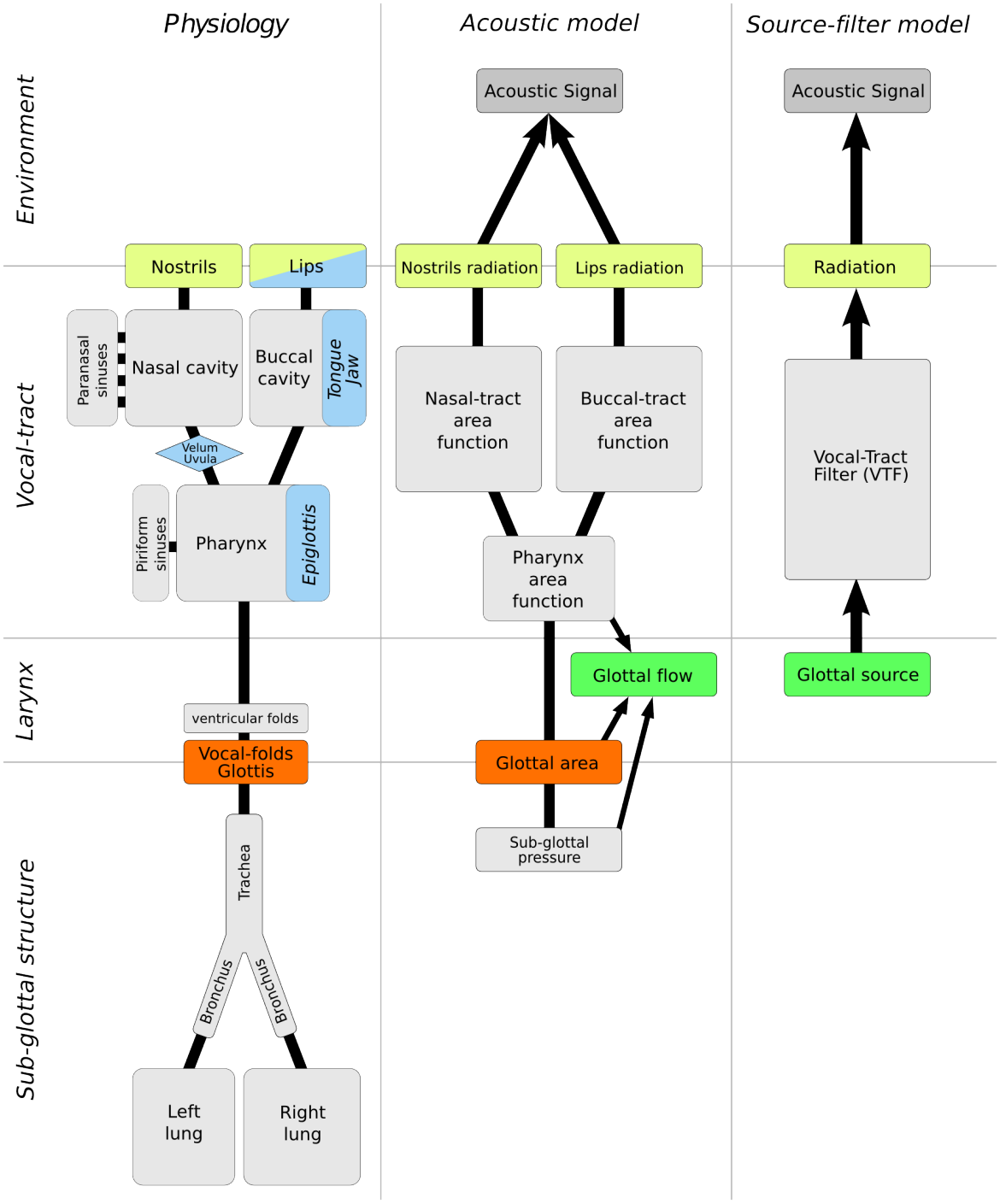


Figure 1.2: Schematic view of voice production models

1.1.1 Source-filter model vs. acoustic model

Figure 1.1 depicts a profile cut of the head to show the various structures of the physiology of voice production whereas the left part of figure 1.2 emphasizes the links between these elements. The articulators are in blue, the passive structures are in grey and the glottis which is acoustically active is in red like the vocal folds. During the realization of a voiced phoneme, pushed by the air coming from the lungs, the vocal folds vibrate like the lips for a brass instrument. Such a modulation of the flow by the glottis creates an acoustic source which is modified according to the resonances and anti-resonances of the vocal-tract due to the shape of this latter. Finally, the acoustic wave radiates outside of the head through the mouth and the nostrils.

In the center of figure 1.2, an acoustic model is depicted [Mae82a, MG76]. In this model, the impedance of the vocal apparatus is represented by area sections and their physical properties all along the structures. The impedance of the larynx is mainly defined by the glottal area. Using such a model, the forward problem of the voice synthesis can be well approximated by numerical integration of the differential equations of the associated problem [Mae82a]. Note that, in such a model, the glottal flow (the air flow going through the glottal area) is an implicit variable of the system. Other models, which are not illustrated here, take into account the influence of the glottal flow on the vocal folds motion [PHA94, ST95]. In this case, the glottal area is an implicit variable which is influenced by the imposed mechanical properties of the vocal folds.

To the right of figure 1.2, the source-filter model is depicted [Mil59]. This model simplifies the acoustic model using two strong assumptions:

I The modulation of the flow by the glottis is independent of the variations of the vocal-tract impedance (i.e. the glottal source is equal to the glottal flow). The possible coupling between the acoustic source at the glottis level and the vocal-tract impedance is thus constant.

II In the time domain, voice production can be represented by means of convolutions of its elements. Therefore in spectral domain, on can write:

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot L(\omega) \tag{1.1}$$

where $G(\omega)$ is the spectrum of the acoustic excitation at the glottis level, resonances and anti-resonances of the vocal-tract are merged into a single filter $C(\omega)$ termed *Vocal-Tract Filter* (VTF) and the radiations at the mouth and nostrils level is merged into a single filter $L(\omega)$ termed *radiation*. Such a system is thus linear and its elements can be commuted.

In addition to these two hypotheses, although the influence of the glottal area on the VTF can be modeled using a time-varying filter, the VTF is usually assumed to be independent of the glottal area variations and stationary during a period of vocal folds vibration.

Regarding the assumptions used in the source-filter model, it is important to compare these two models and their purposes. Indeed, on the one hand, as proposed by the acoustic model, it is interesting to reproduce analytically or numerically the voice production as closely as possible to physical measurements in order to understand and describe it (e.g. understand the mechanical behavior of the vocal folds, study the different levels of coupling between the acoustic source and the vocal-tract impedance). On the other hand, one can be interested in the manipulation of the perceived elements of the voice. Therefore, for this latter purpose, modeling all the physical behaviors of voice production can be unnecessary [Lju86] (e.g. if we assume that the couplings can be neglected for this purpose). This perception based point

of view partly explains why the source-filter model is widely used for voice transformation and speech synthesis. Since this study matches the same objectives, the source-filter model has been used.

Additionally, to transform a given speech recording, it is necessary to estimate the parameters of the model of the voice production model which is used. Regarding this inverse problem, the source-filter has a strong advantage compared to the acoustic model. The source-filter model is more flexible than the interpretation depicted for the voice in figure 1.2. Indeed, the spectrum of any signal can be decomposed in a smooth amplitude envelope and the residual of this envelope. For example, the linear prediction method has not been especially developed for speech [Gol62, Yul27], although this analysis tool has been widely used and its results interpreted in the context of speech [MG76]. Similarly, the minimum/maximum-phase decomposition by means of the complex cepstrum, which can be interpreted as a source-filter separation, has not been developed for voice analysis only but also for radar signals analysis and image processing [OSS68]. Regarding the acoustic model, although it has been shown that the vocal-tract area function can be approximated using linear prediction [Den05, MG76], it has been shown that very different vocal-tract area functions can imply very close vocal-tract impedances [Son79]. The inversion of an acoustic model is thus ill-conditioned and remains, currently, an open problem, although it is of great interest to obtain a model which is closer to voice production than the source-filter model. In this study, due to the inversion issues of the acoustic models and the flexibility of the source-filter model, we propose a separation scheme of this latter.

1.1.2 The chosen approach: the glottal model

For voice transformation and speech synthesis, most of the existing methods tends to model the voiced signal with a minimum number of *a priori* about the voice production. For example, although the phase vocoder [FG66] is designed for the voice, it achieves also efficient pitch transposition and time stretching of musical signals [LD99a]. In other methods, the vocal folds vibration is often assumed to be periodic. Therefore, the fundamental frequency is used in current methods (e.g. WBVPM [Bon08], STRAIGHT [KMKd99], HNM [Sty96], PSOLA [MC90, HMC89]). In this study, using more *a priori* about the voice production is assumed to improve the voice transformation and the speech synthesis quality. Among all the possibilities, this study focuses on an analytical description of the deterministic component of the glottal source, a glottal model.

Regarding the control of voice manipulations, a glottal model should be of great interest. Indeed, regardless of the method, once the voiced signal is modeled, the parameters of the used model have to be modified in order to obtain the target effect (e.g. create a breathy voice from a neutral voice). However, such a link between low-level parameters and high-level voice qualities is not obvious. Therefore, the control of a voice model which uses a glottal model should be easier than using parameters which are close to the signal properties (e.g. frequencies, amplitudes and phases of a sinusoidal model). Indeed, the parameters of a glottal model link its phase and amplitude spectra in a way which respects physical assumptions and constraints. However, the use of a glottal model implies a model of the voice production which is less flexible than the simple residual-filter model. Whereas the latter applies to any signal, a glottal model is limited to the signals it can be fitted to. Accordingly, in this study, we assume that a glottal model can fit the excitation source of the source-filter model. Starting from this assumption, this study mainly intends to answer the three following questions:

1) How to estimate the parameters of a glottal model?

2) How to estimate the Vocal-Tract Filter according to this glottal model?

3) How to transform and synthesize a voiced signal using this glottal model?

Part II of this document tries to answer the first question and try to party answer the second one using a reduced representation of the VTF. An analysis/synthesis procedure is proposed in part 8.3 which fully represents the VTF and tries to handle the third question.

For voice modeling, glottal models have already been used in ARX/ARMAX based methods [Lju86, Hed84]. However, although this approach has been studied since two decades, no implementations of these methods achieved perennial results as other techniques did in voice transformation and synthesis, mainly because of robustness issues [AR09]. Indeed, the following methods have shown capabilities to model the voice production in various applications: WBVPM for pitch transposition in Vocaloid [Bon08]; STRAIGHT for HMM-based synthesis [YTMK01]; PSOLA and phase vocoder for voice transformation [Roe10, FRR09, LD99b]. Consequently, as assumed in other recent studies [AR08, VRC07, Lu02], the use of a glottal model remains a challenging problem for voice transformation and speech synthesis. Regarding to the robustness issues encountered by ARX/ARMAX based methods, in order to reduce the degrees of freedom of the voice production model used in this study, the shape of the glottal model is controlled by only one parameter. Conversely, this drastically reduces the signals space the proposed methods can be applied to. Increasing the flexibility of the proposed solutions should be a next step.

To the best of the author's knowledge, many studies have already been dedicated to time and spectral properties of glottal models [van03, Hen01] whereas the estimation of the parameters of these models held less attention. Accordingly, we consider that developing the knowledge and the understanding of estimation methods of glottal parameters is of great interest. In this study, based on the minimum and maximum phase properties of the speech signal, various methods are proposed. In addition, the conditions a glottal model has to satisfy in order to obtain reliable estimates of its parameters are discussed for the proposed methods.

1.1.3 Evaluation and validation of the proposed methods

Even though the estimation of glottal parameters is currently an active research field [DDM+08, VRC05b, Fer04, Lu02], the validation of the corresponding methods is a tricky problem. Indeed, the ground truth of voice production is far from accessible. A measurement of the glottal flow which is usually associated to the source of the source-filter model could be compared to glottal model estimates. However, a measurement of this flow is an equally difficult problem. Moreover, the acoustic coupling between the glottal flow and the vocal-tract make this comparison difficult to establish.

Before computer resources reached a sufficiently high level to make robust evaluations possible using large databases [PB09, Str98, DdD97], most of the evaluations are made with a few demonstrative pictures. However, it is difficult to know to which extent those *star examples* are representative of a general behavior of the presented methods. Consequently, in this study, a synthetic signals database is used to evaluate the efficiency of the proposed methods. Although this approach allows to evaluate the theoretical limits of the methods, it does not evaluate the methods robustness and the capability of the used voice model to fit an observed signal. Consequently, ElectroGlottoGraphic (EGG) signals are used in this study to create references which are compared to estimated glottal parameters. In current literature, using preference tests, the evaluation of estimation methods is often avoided by evaluating the transformation and synthesis capabilities of procedures using these estimation methods [Lu02];[Yos02, KMKd99];[RSP+08, Alk92]. In this study, this approach is also considered using the analysis/synthesis method presented in chapter 9.

1.2 Structure of the document

Part I: Voice production and its models

The vocal apparatus is described in this part with the necessary details in order to review the usual simplifications and hypothesis made by the source-filter model. However, given the interdisciplinary nature of this part, one could dedicate a whole study for each of its points of view: the physiology of the structures, the acoustic models, the perception of the voice, etc. Therefore, given the purpose of this study, the descriptions given in this part are assumed to be sufficient in order to establish the voice production model which is used in the two following parts. Chapter 2 describes the excitation source which is filtered by the elements presented in chapter 3.

Part II: Analysis

The second part first discusses in chapter 4 the separation problem, its related hypothesis and its state of the art. In the following chapters, various estimation methods of glottal parameters are presented with their related operating conditions listed. Finally, chapter 8 deals with the evaluation of those methods using synthetic and EGG signals.

Part III: Voice transformation and speech synthesis

Glottal source and Vocal-tract separation can be used in many applications: voice transformation, speech synthesis, identity conversion, speaker recognition, clinical diagnosis, affect classification, etc. This part presents an analysis/synthesis method in chapter 9 which uses a glottal model in order to properly separate the glottal source and the vocal-tract filter. Chapter 10 evaluates the efficiency of this method in the context of pitch transposition, breathiness modification and speech synthesis.

All chapters end with a list of conclusive remarks and the last chapter 11 ends this document discussing the contributions of this work and adds remarks for future directions.

Part I

Voice production and its model

Chapter 2

The glottal source

Anything which can vibrate in the vocal apparatus is a possible acoustic source: the vocal folds, the velum and or uvula (snoring), the tongue (alveolar trill /r/), ventricular folds (*Harsh* voice in Metal music), aryepiglottic folds [EMCB08], etc. However, we will consider in this study only the vibration of the vocal folds since it is the source of the voiced phonemes on which voice quality depends.

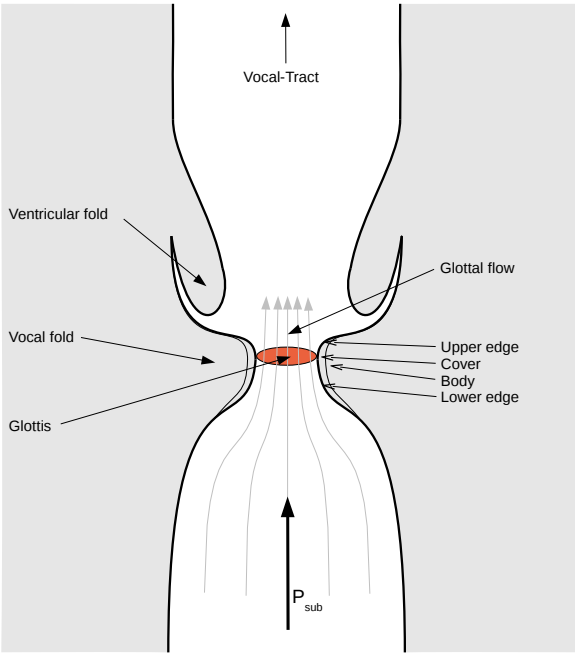
2.1 Vocal folds & glottal area

The larynx contains the cartilages and the muscles which control the mechanical properties, the length, the tension, the thickness and the vibrating mass of the vocal folds. Figure 2.1(a) shows a diagram of a vertical cut of the larynx whereas figure 2.1(b) shows an *in vivo* image of the vocal folds taken downward from the pharynx using High-Speed Videoendoscopy (HSV). Pushed by the air coming from the lungs, the vocal folds vibrate. Most of the time, this vibration is periodic because the vocal folds motion is maintained in a self sustained mechanism. The resulting motion is thus not neurologically controlled. For a standard male voice, the fundamental frequency f_0 of this vibration is around $120Hz$ whereas $200Hz$ corresponds more to female voices and $400Hz$ is easily reached by a child.

Using High-Speed Videoendoscopy (HSV) it is possible to record the vocal folds motion (see fig. (b)). The area of the glottis can be then estimated on each image of such a recording (see Appendix C for the method that we propose). During this study, we used a HSV camera to make 183 recordings with synchronized ElectroGlottoGraphic (EGG) signal and acoustic waveforms [DBR08]. Although we recorded healthy speakers and mainly singers, figure 2.2 shows the diversity of glottal area functions which can exist. Examples (a) and (b) show that the glottal area can be skewed to the left as well as to the right (also shown in [SKA09]). Additionally, example (f) shows that it is possible that ripples appear on the glottal area. Note that, simulations of the glottal flow with the Maeda's model [Mae82a] (see appendix D) can show a glottal flow which is also skewed to these both directions.

Bernoulli's effect

When the air passes through the vocal folds, its speed is larger than when released in the space above the glottis. A Bernoulli's effect is thus created and the vocal folds motion can be influenced by this pressure difference (*Level 2 interaction* in [Tit08]). From the point of view of the source-filter model, we assume that this effect is included in the characteristics of the glottal source (i.e. according to our objectives, it is not necessary to separate this effect from the source).

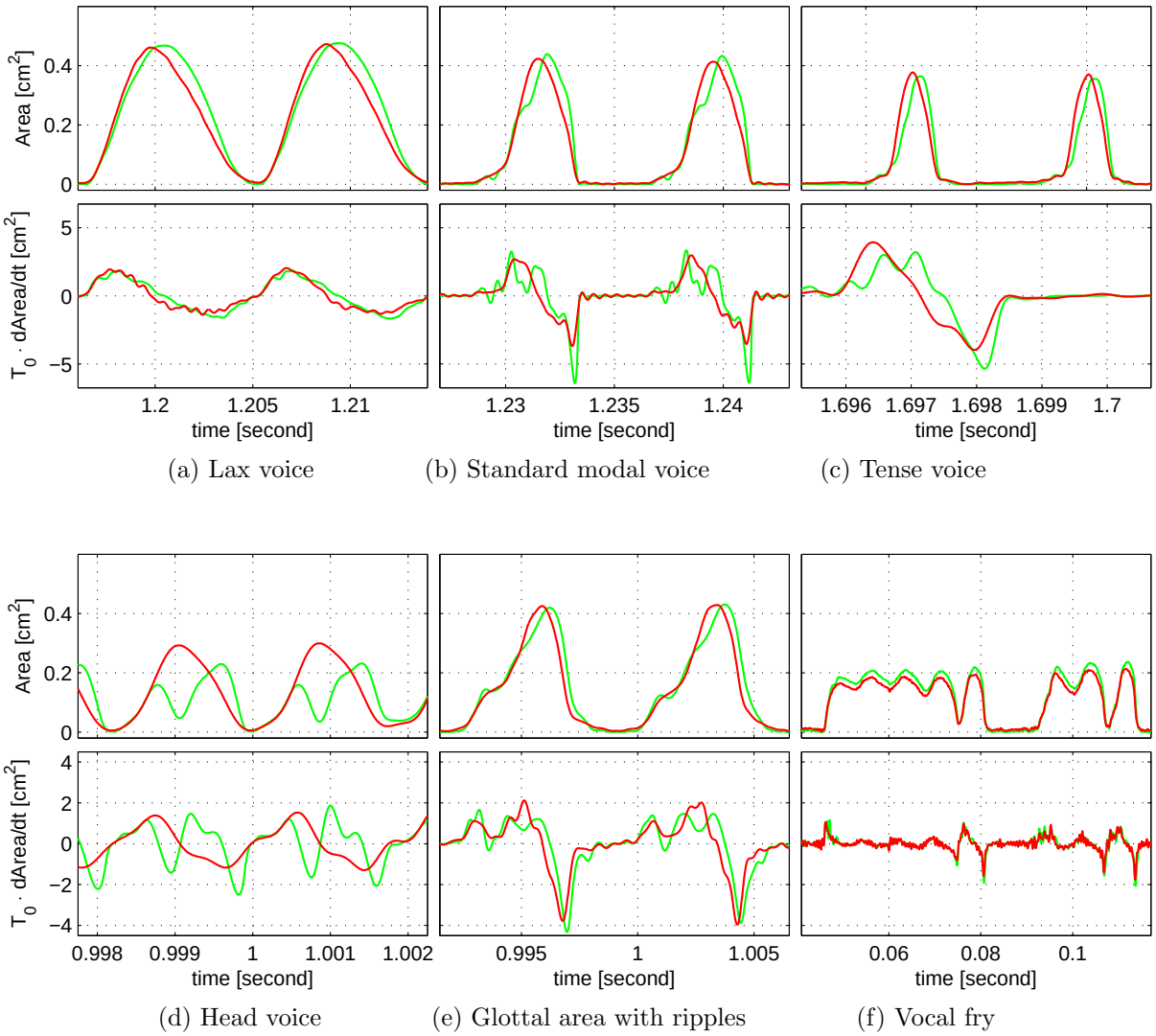


(a) Schematic view of a front vertical cut of the larynx

(b) High-Speed Videoendoscopic (HSV) image

Figure 2.1: (a) Diagram of a vertical cut of the vocal folds; (b) High-Speed Videoendoscopic image of the larynx: This image is taken from the oropharynx (see fig. 1.1) in direction to the larynx. The top of the image corresponds to the back (posterior) of the larynx, the glottis is the dark area in the center which is delimited by the vocal folds (see samples `USC08.ouv_renf.50.*` for the complete video recording at 4000 images/sec)

©2010 All HSV images are properties of Erkki Bianco & Gilles Degottex - Ircam.



(a) Lax voice (b) Standard modal voice (c) Tense voice

(d) Head voice (e) Glottal area with ripples (f) Vocal fry

Figure 2.2: Examples of estimation of glottal area in red line and simulated glottal flow in green line using Maeda's simulator [Mae82a]. (a) is taken from `USC08.60.avi`, (b,c) from `USC08.50.avi`, (d) from `USC08.65.avi`, (e) from `USC08.1.avi` and (f) from `USC07.62.avi`. All examples are taken from men voices except (d). Since the estimated glottal area is given in pixels/second, the maximum area of each recording is normalized to $0.48\,\mathrm{cm}^2$. The simulated glottal flow is computed using an average sub-glottal pressure of 784 Pa. For the comparison, the maximum of the glottal flow is normalized to the maximum of the glottal area.

2.2 Laryngeal mechanisms & voice quality

The vocal folds are made of different layers, mainly the body and the cover (see fig. 2.1(a)). Depending on the muscles tension, the cover can vibrate without a vibration of the main body. Moreover, the folds have a certain thickness. Therefore, the upper edge can vibrate without vibration of the lower edge. These differences of vibration imply laryngeal mechanisms which exclude each other. Through this exclusion, changing from one mechanism to another creates a pitch discontinuity in non trained singers [RHC09, Hen01]. Follows, a brief description of the usual laryngeal mechanisms [RHC09].

- Mechanism 1, known as *Modal voice* or *Chest voice*

  In this mechanism, both body and cover of the vocal folds vibrates. The usual fundamental frequency is about 120 Hz for a male speaker and about 180 Hz for a female speaker. The closed duration of the glottis represent an important part of the fundamental period (e.g. fig. 2.2(b)). This is the most used mechanism in speech.

  Samples: middle of `USC08.50.*`, end of `USC08.16.*`

- Mechanism 2, known as *Falsetto voice*, *Soft voice* or *Head voice*

  The larynx muscles make the vocal folds thinner than in mechanism 1. Consequently, on each vocal-fold, the upper edge vibrates while the lower edge remains fixed. Additionally, whereas the cover move, the body is assumed to be fixed. The fundamental frequency can be easily twice as in mechanism 1. The closed duration can be difficult to define on a HSV recording but the open duration is clearly bigger than in mechanism 1 (e.g. fig. 2.2(d)). This mechanism is mainly used by children and often used by female speakers.

  Samples: `USC08.65.*`, start of `USC08.16.*`

- Whistle voice, (sometimes termed *Mechanism 3*)

  The physiological and mechanical descriptions of this mechanism is not yet described as well as the two previous ones and seems to be fairly similar to the mechanism 2. The fundamental frequency is usually above 1000 Hz. Conversely to the two previous mechanisms, this mechanism is never used is speech. Soprano singers can use this mechanism as well as young children to shout.

- Vocal fry, (sometimes termed *Mechanism 0* or *Creaky voice*)

  In this mechanism, the vocal folds motion is irregular and create a popping of an average frequency between $10 - 50$ Hz (e.g. fig. 2.2(f)). Due to this irregularity, a pitch is not properly perceived. This mechanism is frequent in speech and often appears at the end of sentences of English voice, when the lungs pressure slacken. Note that a lot of variations of this mechanism exist: exhaled, inhaled, alternated.

  Samples: exhaled fry `USC07.62.*`, inhaled fry `USC07.61.*`, end of `snd_arctic_bdl.1.wav`

For all laryngeal mechanisms, various voice qualities exist. For instance as depicted in figure 2.2(a,b,c) for mechanism 1, the larynx can vary between *lax* and *tense* states (see `USC08.50.*` $0.65 < t < 1.80$ s). Additionally, aspiration noise can be generated through a leakage at the posterior side of the glottis which is created by the abduction of the vocal-folds (see sample `USC08.60.*` $t = 1.25$ s and section 2.6 for its description). For the following, we define a voice quality axis termed *breathy-tense axis* which is relevant for speech. This axis is made of a *breathy voice* at one side, a lax voice with significant aspiration noise. Then, by the tightening of the vocal folds, the aspiration noise is reduced and, at high frequencies, the

deterministic source is increased and thus more easily perceived, leading finally to the *tense voice* at the other side of this axis.

2.3 Glottal flow vs. Glottal source

As already seen in the introduction, one of the main difference between the acoustic model and the source-filter model is related to the modeling of the source of voice production. Therefore, in this section, the glottal source of the source-filter model is described and its relations to the glottal flow are discussed. The glottal flow is the air flow coming from the lungs which is modulated by the glottal area. First, compared to the glottal area function, the glottal flow is skewed toward positive time [Tit08, CW94, Fla72a]. Therefore, it is not possible to express the glottal flow with an instantaneous function as $\text{flow}(t) = f(\text{area}(t))$ where $f(x)$ would be independent of t. Secondly, the glottal flow is a non-linear function of the glottal area and the sub-glottal pressure [Fla72a]. Therefore, from these two points, the band limit of the glottal flow is higher than that of the glottal area. Thirdly, the glottal flow is dependent on the vocal-tract shape since this latter defines the acoustic load above the glottis (described as *Level 1 interaction* in [Tit08]). Figure 2.3 shows simulation of glottal flow for various phonemes and various subglottal pressures from a measured glottal area on a HSV recording. A clear skewness effect of the flow is shown with the phoneme /ɛ̃/. Additionally, one can see ripples on the ascending branch of the pulses which vary significantly between the phonemes.

If we assume that the relation between the glottal flow and the flow at the output of the vocal apparatus is linear, the real glottal source of the source-filter model is the glottal flow as described above and illustrated in figure 2.3. Consequently, if we use a glottal source which is defined as independent of the vocal-tract properties, we have to assume the following points: 1) the pulse skewness and the non-linear relation with the glottal area are equal for all vocal-tract shape. 2) The ripples of the glottal flow which depend on the vocal-tract impedance have to be neglected. Accordingly, considering a source-filter model, we define in this study the *glottal source* as a non-interactive signal description of the voice source which is thus independent of the variations of the vocal-tract shape.

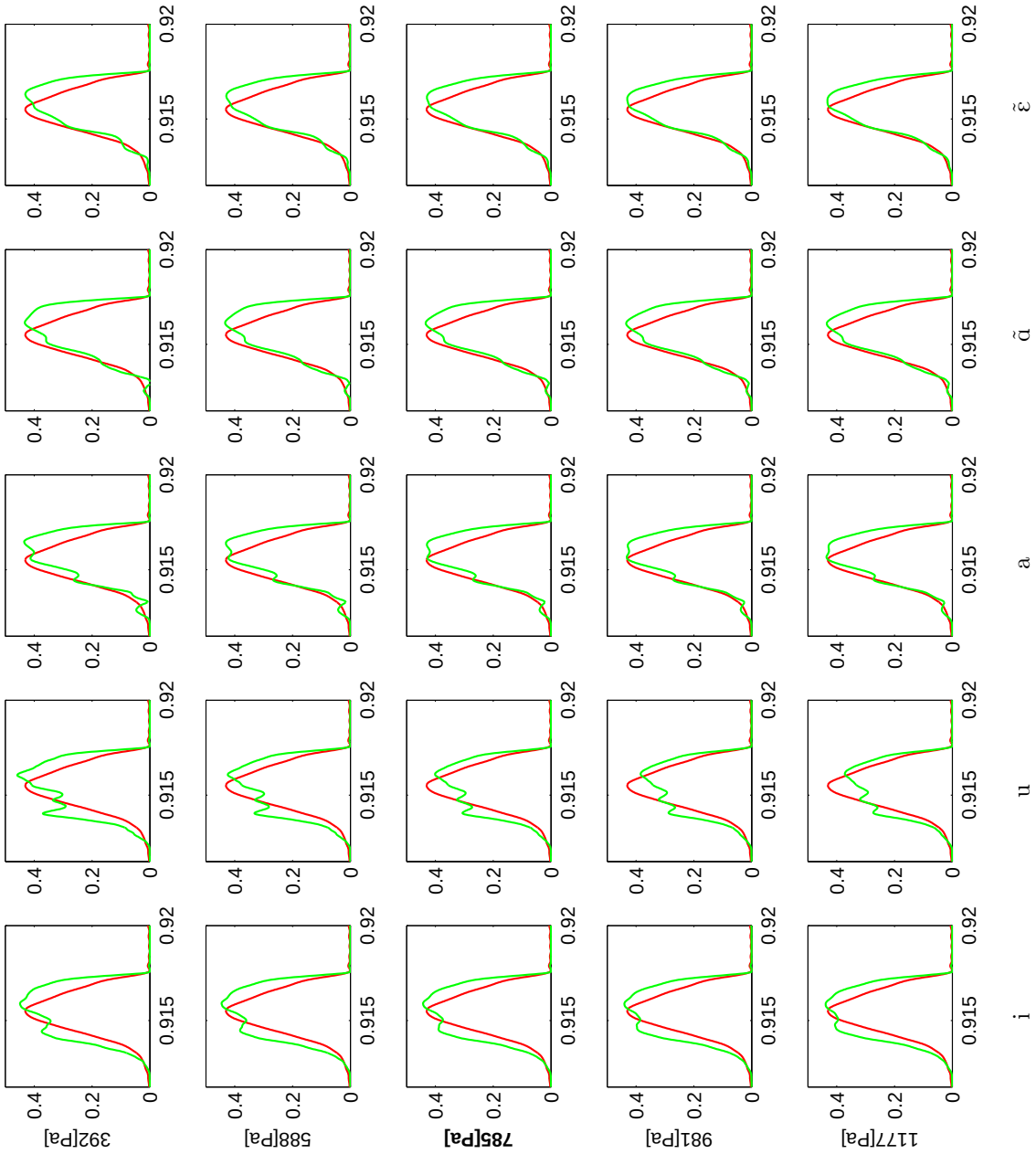


Figure 2.3: Examples of glottal area in red line and simulated glottal flow in green line using Maeda's simulator for different vowels and sub-glottal pressures. 785 Pa is an average sub-glottal pressure for speech.

2.4 Glottal models

Many non-interactive glottal models have been proposed to define analytically one period of the glottal flow [DdH06, CC95]. In the following, these models will be used to describe the deterministic component of the glottal source. As shown in figure 2.4, existing models use mainly a set of time instants (known as T parameters [Vel98]):

t_s Time of the start of the pulse (for the following definitions, we assume $t_s = 0$).[1]

t_i Time of the maximum of the time-derivative.

t_p Time of the maximum of the pulse. This maximum is termed the voicing amplitude A_v.

t_e Time of the minimum of the time-derivative.

t_a The return phase duration.

t_c Time of the end of the pulse.

T_0 Duration of the period $T_0 = 1/f_0$

Then, each glottal model $g(t)$ defines a set of analytical curves passing through these points. Although most of the glottal models are mainly defined in time domain, it has been shown that their amplitude spectrum $|G(\omega)|$ can be stylized as in right plots of figure 2.4 [DdH06]. In the following list, we briefly describe the most known and used glottal models:

- **Rosenberg** [Ros71]

  Initially, Rosenberg proposed 6 different models to fit a glottal pulse estimated by inverse filtering (see sec. 4.4). In [Ros71], a preference test has been used to select the model which sounds as the best source. The best model, the Rosenberg-B model, is thus referred as the Rosenberg model.

  This model is made of 2 polynomial parts:

  $$g(t) = \begin{cases} t^2(t_e - t) & \text{if } 0 < t < t_e = t_c \\ 0 & \text{if } t_c < t < T_0 \end{cases}$$

  This model has only 1 shape parameter: t_e the instant of closure. The instant of maximum flow is proportional to t_e: $t_p = \frac{2}{3}t_e$.

- **KLGLOTT88 (or Klatt)** proposed by Klatt and Klatt [KK90]

  The glottal pulse is synthesized in the same way as the Rosenberg model. A low pass resonator is then added to smooth the shape of the pulse and control the spectral tilt in the high frequencies. This model has 2 shape parameters: the open quotient $OQ = t_e/T_0$ and a parameter controlling the spectral tilt TL in dB down at $3kHz$. The model is mainly used in the KLSYN88 synthesizer [KK90]. An analytical definition of the KLGLOTT88 spectrum can be found in [Dd97].

- **Fujisaki** proposed by Fujisaki and Ljungqvist [FL86]

  This glottal model describes the glottal pulse $g(t)$ in 4 polynomial parts with 6 shape parameters. The analytical description can be found in [FL86]. The Rosenberg and Ananthapadmanabha

[1] t_0 is usually used. To avoid any confusion with $T_0 = 1/f_0$, we chose to use t_s instead of t_0.

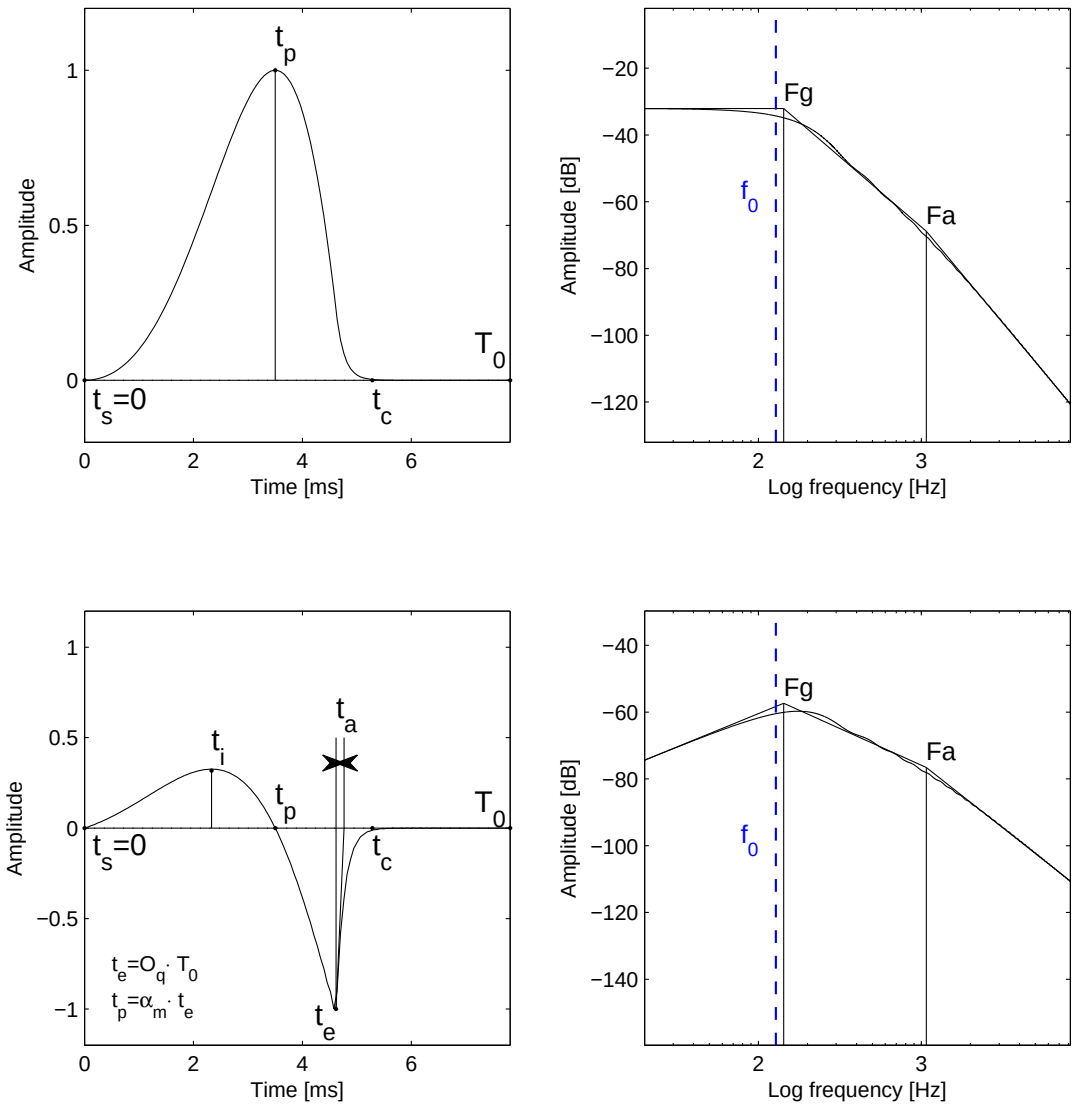


Figure 2.4: To the left in the time domain the main scheme of one glottal pulse used by most of the glottal models. To the right, in the frequency domain the stylization of the glottal amplitude spectrum according to [DdH06].

[Ana84] models can be approximated by this one by setting different parameters to zero. Whereas the Rosenberg and KLGLOTT88 models can model only discontinuities at closure, this model can model discontinuities at both opening and closure. Note that contrarily to all other models, this one does not exclude a negative flow due for instance to a lowering of the vocal folds following the glottal closure.

- **Fant** [Fan79]

  This model is made of 2 sinusoidal parts:

$$g(t) = \begin{cases} \frac{1}{2}(1 - cos(\omega_g t) & \text{if } 0 < t < t_p & \text{the rising branch} \\ K \cdot cos(\omega_g(t - t_p)) - K + 1 & \text{if } t_p < t < t_c = t_p + \frac{arcos\frac{K+1}{K}}{\omega_g} & \text{the descending branch} \\ 0 & \text{if } t_c < t < T_0 & \end{cases}$$

  with $\omega_g = \pi/t_p$. This model has 2 shape parameters: t_p and K which control the slope of the descending branch (if $K = 0.5$ the pulse is symmetric and if $K \geq 1 \Rightarrow t_e = t_c$).

- **LF** proposed by Liljencrants, Fant and Lin [FLL85]

  This model defines the time derivative of the glottal pulse. A first part is made of one exponential part modulated by a sinusoid and a second part is made of one exponential.

$$\dot{g}(t) = \begin{cases} e^{\alpha t} sin(\omega_g t) & \text{if } 0 < t < t_e & \text{the open phase} \\ \frac{-1}{\epsilon t_a}\left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c - t_e)}\right] & \text{if } t_e < t < t_c = T_0 & \text{the return phase} \end{cases}$$

  with $\omega_g = \pi/t_p$ and the synthesis parameters α and ϵ are computed in order to respect $\int_0^{T_0} \dot{g}(t)dt = 0$ given the 3 shape parameters (t_e, t_p, t_a). If $t_a = 0$ the return phase is zero and the LF-model reduces to the L-model (also described in [FLL85]). Besides being the most used model, this one is also the most studied in terms of spectral properties [van03, Hen01, DdH06, FL88]. Note that an analytical definition of the LF spectrum can be found in [Dd97].

- **Transformed-LF (LFRd)** a particular parametrization of the LF model proposed by Fant [Fan95]

  The analytical description is the same as in the LF model but the shape space covered by the LF model with its 3 shape parameters is reduced to a meaningful curve parametrized by only one shape parameter Rd. Fant has shown that this parameter is the most effective parameter to describe voice qualities into a single value [Fan95].

  First, the T parameters can be expressed in a T_0-normalized form (known as R parameters [Vel98]):

  $R_o = t_e/T_0$ is the duration of the open phase.

  $R_g = T_0/(2t_p)$ is the rising speed of the pulse.

  $R_k = (t_e - t_p)/t_p$ is the symmetry of the glottal pulse.

  $R_a = t_a/T_0$ is the T_0-normalized return phase duration (according to [Fan95, Vel98][2]).

  Then, the Rd parameter has been described using relations between the R parameters measured on various speakers from extreme tight adducted phonation to very breathy abducted phonation.

[2] Note that R_a is normalized to the return phase duration in [FLL85]: $R_a = t_a/(t_c - t_e)$

From these measurements, the following statistical regression has been proposed [FKLB94]:

$$Rd = (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a)$$

and the ratio parameters can be predicted from the Rd parameter

$$\begin{aligned} R_{ap} &= (-1 + 4.8Rd)/100 \\ R_{kp} &= (22.4 + 11.8Rd)/100 \\ R_{gp} &= 1/(4((0.11Rd/(1/2 + 1.2R_{kp})) - R_{ap})/R_{kp}) \end{aligned} \tag{2.1}$$

According to [Fan95], the main range of Rd is $[0.3; 2.7]$ while the range $]2.7; 5]$ is assumed to model extreme situations of abducted vocal folds. Figure 2.5 shows examples of the LFRd model for various Rd values.

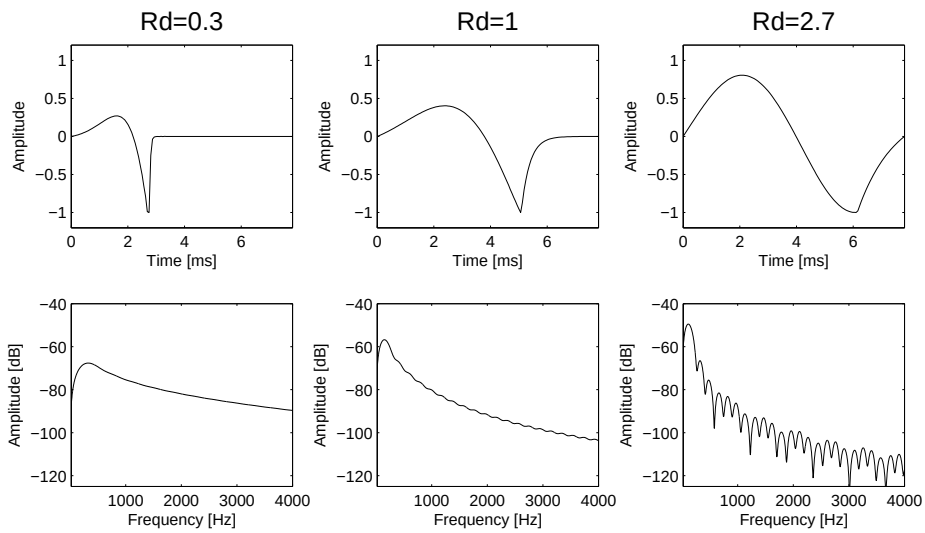


Figure 2.5: Examples of the LFRd model for various Rd values.

- **CALM** Causal-Anticausal Linear Model proposed by Henrich, Doval and d'Alessandro [DdH03, HDd99, DdD97]

  This is the only glottal model fully described in the spectral domain. The idea is the following. From the LF model, one can see that the open phase is a truncated impulse response of an anti-causal stable pole. In addition, the return phase is close to a truncated damping exponential. Therefore, defining the time origin at t_e, the open phase can be approximated using an anti-causal conjugate pole pair and the return phase can be approximated using a causal real pole. This model can thus be defined by two filters. The first anti-causal filter is:

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

with

$$a_1 = -2e^{-a_p/f_s}\cos(b_p/f_s) \qquad a_2 = e^{-2a_p/f_s} \qquad b_1 = \frac{\pi^2}{b_p^3}e^{-a_p/f_s}\sin(b_p/f_s)$$

$$a_p = -\frac{\pi}{O_q T_0 \tan(\pi\alpha_m)} \qquad b_p = \frac{\pi}{O_q T_0}$$

and the second causal filter $H_C(z)$ is equivalent to the low-pass filter of the KLGLOTT88 model which is used to control the spectral tilt in high frequencies. The CALM glottal model has thus the same shape parameters as the LF model except for the return phase which is replaced by a parameter TL in dB down at $3kHz$.

Note that, conversely to the LF model, synthesis parameters α and ϵ are no longer necessary in the CALM. Moreover, the anti-causal pole pair is independent of the TL parameter and the causal real pole is independent of the parameters O_q and α_m. However, in the LF model, there is no such separation between the parameters and the poles because the synthesis parameters α and ϵ are dependent on all of the 3 LF parameters and these two synthesis parameters influence all the poles.

Obviously, this list is far from exhaustive and other glottal models exist [SA10, Mil86, Hed84, Ana84].

In this study, the **Transformed-LF (LFRd)** model is used for the following reasons:

1) As argued in the introduction this study is not focused on the glottal models themselves but on methods to estimate their parameters. Therefore, we choose one relevant glottal model according to the following points.

2) The LF model is widely used and studied [RSP+08, Air08, van03, Hen01, DdH06, FL88]

3) Regarding to the ARX error minimization [Vin07] and regarding to the phase minimization criteria (see sec. 5.3), the parameters of the LF model are not independent in terms of estimation. Therefore, in order to avoid this issue and the possible ambiguities of estimation of multiple parameters, we chose to use only one shape parameter. The evaluation of the existing glottal models and the estimation of their multiple parameters should be investigated in a dedicated study.

4) The Rd parameter is the most effective parameter to describe voice qualities into a single value [Fan95].

2.5 Time and spectral characteristics of the glottal pulse

This section put together the elements which characterize the glottal source in terms of time and spectral properties.

Excitation amplitude (E_e) and *Shimmer*

The amplitude of the time-derivative of the glottal pulse at time t_e is termed E_e (see fig. 2.4). Since we use the LF model in this study, we prefer to characterize the amplitude excitation of the glottal model by this value instead of the voicing amplitude A_v (the maximum amplitude of the glottal pulse at time t_p). Note that in a natural voice, the pulse amplitude is never perfectly constant. Small variations of this value are termed *Shimmer*. Consequently, an amplitude modulation of the glottal source always exists. In this study, observing the speech signal through a window sufficiently short (≈4 periods), we assume that this modulation is negligible inside a single window.

Pulse duration, periodicity ($T_0 = 1/f_0$) and *jitter*

First, one can consider one glottal pulse alone with a given duration (T_0 in the definition of the glottal models). Then, repeating glottal pulses at regular interval creates a periodicity. Although a periodic source is necessary in singing and it is the most common case in speech, the pulses can be irregular when the lungs pressure slacken (see vocal fry in sec. 2.2). Moreover, such irregular behaviors can be also expected in transients with natural and healthy voices. In addition to these particular cases, the periodicity is not perfectly stable. A frequency modulation, termed *jitter*, always exists in natural voice. Note that the jitter has to be distinguished from vibrato. One one hand, the jitter is made of uncontrolled, small and high frequency modulations of the fundamental frequency. On the other hand, the vibrato is usually made of a periodic modulation of approximately $5-6\,\mathrm{Hz}$. Moreover, the amplitude of this frequency modulation is controlled. In this study, we assume that the periodicity is constant in a short window where a pitch tracking algorithm can be used to estimate this value. Numerous methods can be used to compute f_0 from the speech signal: *YIN* [dK02], *Swipep* [Cam07], subharmonic summation [Her88], using harmonic matching [YR04, DR93], using *STRAIGHT* [KdB+05]. The analysis window used in these methods is assumed to be short enough to model the fast variations of the fundamental frequency, like the jitter.

2.5.1 Time properties: glottal instants and shape parameters

The following two instants are often used to characterize the shape of the glottal pulse.

Glottal Opening Instant (GOI) (t_s in fig. 2.4)

This instant corresponds to the start of the pulse, when the glottal pulse start to increase compared to its minimal value.

Glottal Closure Instant (GCI) (t_e in fig. 2.4)

This instant corresponds to the minimum of the time derivative of the pulse. Note that, for low α_m values of the LF model, t_e does not corresponds to the minimum but is slightly moved backward. this instant is not symmetrical to the GOI. Therefore, the instant when the glottal pulse ends to decrease and reaches the minimum value of the pulse (t_c) is sometimes termed *effective closure instant*.

Shape parameters

Once the glottal model is normalized in duration and amplitude by the period and the excitation amplitude respectively, we termed *shape parameters* those which control the shape of the pulse. As shown in [DdH06], models using the scheme of fig. 2.2 can be represented by the following shape parameters:

O_q The *open quotient* is the duration from the GOI to the GCI normalized by the period: $O_q = t_e/T_0$. Even though the glottal pulse can be bigger than zero during the return phase, this phase is not considered in the open quotient definition. The sum of the return phase and the open phase is termed *effective open quotient*.

α_m The *asymmetry* represents the skewness of the pulse $\alpha_m = t_p/t_e$. The closer to 0.5, the more symmetric the pulse. Note that since the definition of the GOI and GCI are not symmetrical, $\alpha_m = 0.5$ does not mean that the pulse is perfectly symmetrical.

Q_a The *return phase duration* is the duration of the return phase normalized by the pulse duration. It represents how abrupt the closure is. The smaller this duration, the more abrupt the closure. $Q_a = t_a/T_0$

2.5.2 Spectral properties: glottal formant and spectral tilt

The *glottal formant* is a maximum which exists on the amplitude spectrum of the time-derivative of the glottal pulse (see fig. 2.4 lower right plot). This peak is termed glottal formant because of its similarity to the shape of the vocal-tract formants. One can consider two different ways to characterize this glottal formant:

F_g is the maximum value of the stylization of the amplitude spectrum of the LF model with a log frequency scale and a dB scale [DdH06, DdH03]. For the LF model, the determination of the F_g value is straightforward if the influence of the return phase duration is assumed to be negligible on the F_g value: $F_g = \frac{f_0}{2 \cdot O_q \cdot \alpha_m}$

F_{gm} We define F_{gm} the frequency of the maximum of the amplitude spectrum of the time derivative of the pulse, i.e $argmax|G(\omega) \cdot (1 - e^{-j \cdot \omega})|$. The determination of this value can be tricky depending on the analytical definition of the glottal model. In this study, the F_{gm} value of the LF model is estimated using a Brent's method [Bre73].

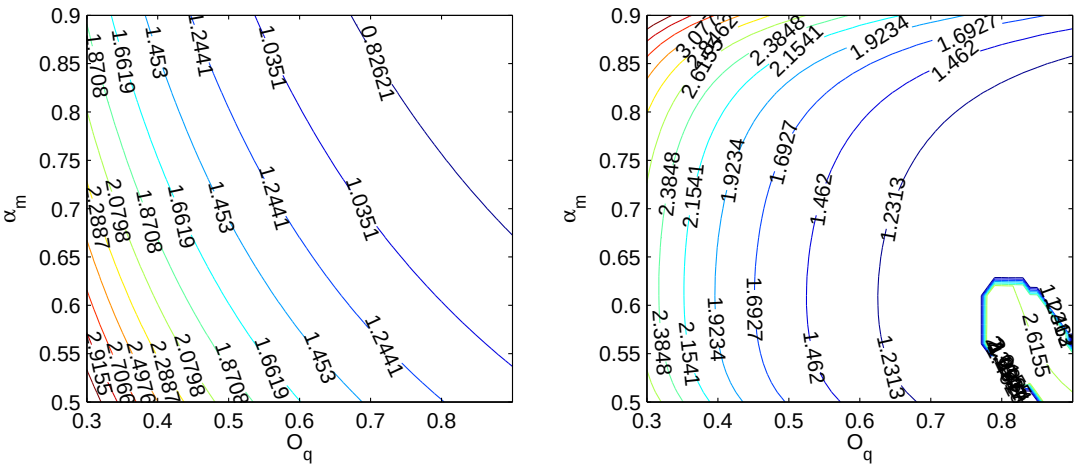


Figure 2.6: Lines of equal glottal formant: the glottal formant frequencies are noticed by the numbers (the values are normalized by f_0). F_g to the left and F_{gm} to the right.

Figure 2.6 shows lines of equal values of glottal formant for a fixed Q_a and different values of (O_q, α_m) One can see that the asymmetry and open quotient coefficients are highly dependent in terms of control of F_g and F_{gm}. For F_g, this dependency is visible on its formula. F_g is kept constant for all equal products $O_q \cdot \alpha_m$.

The *spectral tilt* is the slope of $|G(\omega)|$ for frequencies above F_a (see fig. 2.4). Although O_q and α_m influence this tilt, this latter is highly correlated to the return phase duration Q_a [Hen01]. In KLGLOTT88 and CALM model, the TL parameter is especially designed in the spectral domain to control the level of this tilt.

2.5.3 Mixed-phase property of the glottal source

The open phase of the LF model is made of an exponential modulated by a sinusoid. As shown by the CALM model, this part corresponds to a truncated anti-causal pole pair. Therefore, zeros exist outside of the unit circle in the z-transform of the LF pulse. Additionally, figure 2.7 show that such anti-causal zeros exist also in the Rosenberg model. Consequently, in this study, we assume that the glottal pulse is a mixed-phase signal. Bozkurt *et al.* [BDdD05] already used this assumption to develop a source/filter separation process using the Zeros of the Z-Transform (ZZT) of windowed speech segment. Additionally, separation methods by means of the complex cepstrum are based on this same assumption [DBD09, OSS68]. This assumption implies that the glottal pulse is not composed of a minimum-phase signal only (as it is postulated for the vocal-tract filter in section 3.1.2). Moreover, if we consider that the glottal pulse has no linear-phase component, the mixed-phase assumption implies that the pulse has energy in its anti-causal part. On figure 2.7, one can see that the Rosenberg model has only a maximum-phase component. Conversely, both LF and CALM models possesses a minimum-phase component due to the presence of their return phase.

2.5.4 Vocal folds asymmetry, pulse scattering and limited band glottal model

Around the closure of the glottis, the front and back of the vocal folds do not always close at the same time [MHT00, NMEH01]. It is thus difficult to determine only one closure instant. We propose to create a simple model of this asymmetry by sampling the vocal-fold length into N small parts i. Then, the contribution of each part can be summed in order to create a multi-pulse model $g_N(t)$:

$$g_N(t) = \frac{1}{N}\sum_{i=1}^{N} g_i(t-\gamma_i) \tag{2.2}$$

where $g_i(t)$ is a glottal model and γ_i is the duration between one reference t_e and the t_e of each part. Then, assuming the glottal pulse $g_i(t)$ has the same shape, duration and amplitude for all parts, one can isolate the effect of the delay γ_i

$$g_N(t) = g(t) \circledast \frac{1}{N}\sum_{i=1}^{N} \delta(t-\gamma_i) \tag{2.3}$$

where $\delta(t)$ is the Dirac delta function. When N tends to infinity, the sum of Dirac deltas tends obviously to the time distribution $P_\gamma(t)$ of γ_i:

$$\lim_{N\to\infty} g_N(t) = g(t) \circledast P_\gamma(t)$$

Finally, if we assume that this distribution is Gaussian, the scattering of the pulse by the asymmetry of the closure of the vocal folds is similar to a low-pass filtering. Note that the cutoff frequency of this filtering depends on the variance of the Gaussian distribution. The smaller the variance, the higher the cutoff. In conclusion, the deterministic part of the glottal pulse is band limited. The glottal source cannot produce energy up to an infinite frequency. Additionally, The more synchronous is the closure, the stronger the high frequencies. Note also that, the result of scattered pulses ca not be equivalent to a white noise because this latter is fully independent of the period whereas a scattered pulse is centered on an average closure instant.

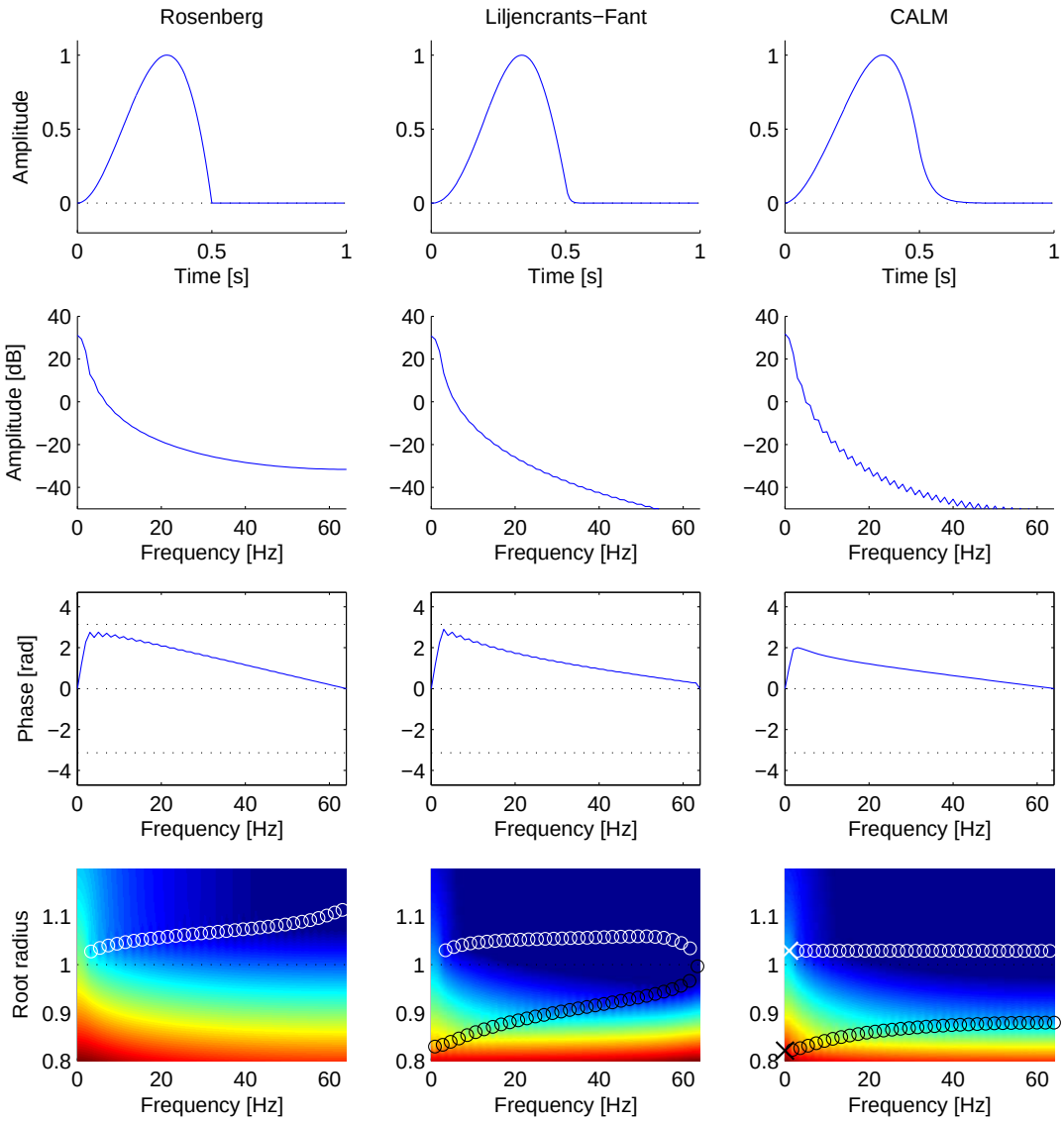


Figure 2.7: From the top to the bottom, this figure shows the pulses in the time domain, their amplitude and phase spectra and the last plots show the surface of the z-transform in dB (cold colors correspond to low amplitudes while hot colors correspond to high amplitude). Zeros related to the maximum-phase component are shown with white circles and zeros related to the minimum-phase component are shown with black circles. The poles used for the synthesis of the CALM are shown with crosses. For the phase spectra, the reference time is set to $t_e = O_q \cdot T_0$.

2.6 Aspiration noise

All the descriptions given above are related to the deterministic component of the glottal source. However, a random component exists. Although random variations can exist in the deterministic component (e.g. jitter and pulse scattering), aspiration noise is also generated in constricted areas of the vocal apparatus [Ste71, Fla72a]. The aspiration noise is the acoustic source of fricatives generated at the tongue level (termed *friction noise*) and the noise generated at the glottis level has an important role in the voice quality of vowels (this noise is thus termed *glottal noise*). In this study, since only voiced segments are discussed, we assume that aspiration noise can be created only at the glottis level.

The aspiration noise is usually described as follow [Fla72a, p.53-58],[Lil85, sec.5]: Considering a fluid passing through an opening of area A, the Reynold's number indicates the likelihood of a turbulence:

$$Re = \frac{2\,\rho\,U}{\mu\sqrt{\pi A}}$$

where U is the volume velocity of the flow passing through the area A, μ the air viscosity and ρ the air density. The noise sound pressure is then proportional to:

$$P_n \sim \begin{cases} Re^2 - Re_c^2 & \text{if } Re > Re_c \\ 0 & \text{else} \end{cases} \tag{2.4}$$

where Re_c is a critical Reynold's number related to the geometry of the area A. For a plastic model of the vocal folds, this value is around 1800 [Lil85].

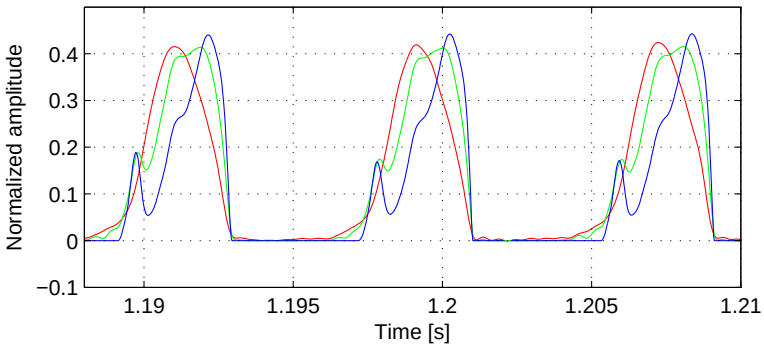


Figure 2.8: Glottal noise sound pressure: glottal area in red, glottal flow in green and noise sound pressure in blue. For the comparison, values are normalized to their maximum value.

According to figure 2.8, one can make the following remarks. Firstly, since the glottal flow is skewed to the right compared to the glottal area (sec. 2.3), the ratio $U/\sqrt{A}$ is always smaller at the glottal opening than at the closure. Consequently, the skewness of the noise sound pressure P_n is more important than the one of the glottal flow. Accordingly, a maximum of noise burst can thus be expected around the closure. Note that, the noise sound pressure is proportional to the squared of the volume velocity of the flow. Moreover, the glottal flow is purely laminar below Re_c and thus no noise is generated in that case.

In this study, we assume that the noise sound pressure P_n can be represented by an amplitude modulation of a Gaussian noise like in [dY08, MQ05, Lu02, Her91]. In terms of perception, the amplitude modulation is important [MQ05]. Indeed, without modulation, the random and deterministic components are perceived as two different sources. This modulated noise has also a spectral color which is dependent on the glottal area [Ste71]. Accordingly, in the proposed analysis/synthesis method, we will simply assume that the amplitude spectrum of the noise is weaker below a given cutoff frequency than above.

Conclusions

- The modeling and estimation of the glottal flow taking into account all the possible interactions between the source and the filter is important to explain and understand voice production. However, according to the purpose of this study, we assume that a non-interactive *glottal source* is sufficient for the manipulation of the voice in terms of perception.
- As argued in the introduction, the main point studied in this work is the estimation methods of the shape parameter of a glottal model. Therefore, the well known and studied Liljencrants-Fant (LF) glottal model has been selected from a list of known existing glottal models (see sec. 2.4). Additionally, the *Rd* parameter of the Transformed-LF model will be used instead of the full parametrization of the original LF model because of estimation issue of multiple parameters and the significance of this parameter to describe voice qualities [Fan95].
- As seen in section 2.5.3, the glottal pulse is a mixed-phase signal. Zeros of its z-transform exist outside of the unit circle.
- Due to imperfect symmetry of the vocal folds motion, the glottal pulse is band limited (sec. 2.5.4). Therefore, no energy should be generated above a given cutoff frequency by the deterministic component of the glottal source.

Chapter 3

Filtering elements and voice production model

In the first following section, the acoustic characteristics of the passive structures of the voice production are merged into a single filter termed Vocal-Tract Filter (VTF). In this filtering element, the following element are considered: the sub-glottal structures (bronchus and trachea), the pharynx, the buccal cavity and the nasal cavity. Then, the second section discusses the radiation of the acoustic pressure which leaves the vocal apparatus from the mouth and the nostrils. The last section describes the model used in this study which represent voice production based on the previous sections and chapter.

3.1 The Vocal-Tract Filter (VTF)

First, if we assume that the vocal-tract is a single tube with a length of 17 cm, a constant section, one closed boundary and one open boundary on other side, this tube have resonance frequencies at 500 Hz, 1500 Hz, 2500 Hz, etc. However, since the shape of the vocal-tract is obviously not a simple tube of constant section, the resonances are not equally spaced [FL71]. In the context of the voice, the *formants* are defined as the ordered resonances of the vocal-tract, from the lowest to the highest. Moreover, by controlling the vocal-tract shape with the articulators (tongue, jaw, velum and lips), the formants position can be controlled resulting in various sound timbres and consequently various phones. Besides the articulators, many structures and properties of the vocal-tract influence the voice timbre: the larynx size, the piriform sinuses [TAK$^+$06], the paranasal sinuses [LGS72], the material properties of the structures, etc.

3.1.1 Structures of the vocal-tract

Sub-glottal structures

Since the lungs are made of spongy tissues and numerous air sacs, they are assumed to be highly acoustically absorbent. However, the trachea and the bronchus, which are tubes of hard materials, have some effect on the resonances of the vocal apparatus [KK90, WF78, Fla72a].

The nasal cavity

The nasal cavity is connected to the buccal cavity at the nasopharynx level and terminated by the nostrils (see fig. 1.1). The opening of the nasal cavity is managed by the position of the velum and the uvula. When the airway to the nasal cavity is completely closed, it has been shown that the filtering effect of the vocal-tract can be modeled by an all-pole model (e.g. Linear Prediction (LP) [MG76] or the Discrete All-Pole (DAP) [EJM91]). However, it is known that the presence of the nasal cavity adds pole-zero pairs to the Vocal-Tract Filter (VTF) [NKv05] (see fig. 3.1). Note that although all-pole models are still widely used in speech analysis, those pole-zero pairs are necessary for nasalized sounds. Additionally, cavities exist in the skull bones which are termed paranasal sinuses (frontal, maxillary, ethmoid and sphenoidal). They are connected to the nasal cavity trough small orifices termed ostia. [Mae82b, LGS72] (dashed circles in figure 1.1). In addition to these side cavities, the nasal cavity itself has an important surface due to the presence of the nasal turbinates which heat and humidify the breathed air. Therefore, the resulting viscous friction and thermal loss should create resonances and anti-resonances with broader bandwidths than that of the non-nasalized sounds [RS78, p.78].

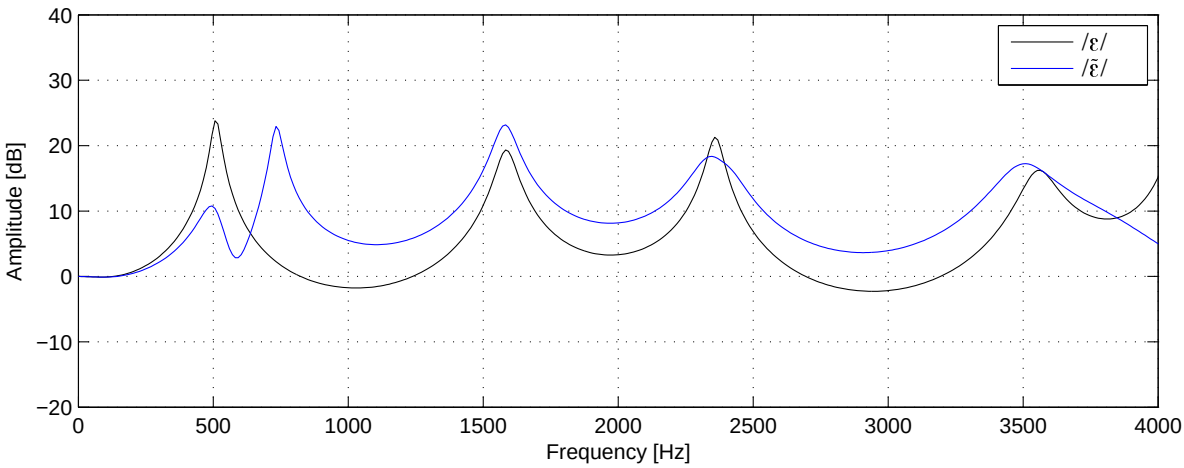


Figure 3.1: Amplitude spectrum of a /ɛ/ in black and a /ɛ̃/ in blue line using Maeda's synthesizer (see appendix D). A pole-zero pair is present in /ɛ̃/ around 500 Hz.

Vocal-tract length

Considering a single tube without nasal cavity, the formants position are inversely proportional to the vocal-tract length. The longer the vocal-tract, the lower the formants. The length of the pharynx can be slightly controlled by the larynx position. In presence of the nasal cavity, using *in vitro* measurements of impedance, it has been shown that by changing the pharynx length, the formants move whereas the pole-zero pair added by the nasalization does not move [Eli09].

Wall vibration, turbulent, viscous and heat conduction losses

The vocal-tract filter is affected by different elements: the viscous loss (the interaction between the wall and the air flow), the laminar or turbulent behavior of the DC flow, the heat conductance of the

wall materials, and finally the wall vibration of the structures of the vocal apparatus [Lil85]. These phenomenons imply mainly energy losses. They increase the formants bandwidth but also slightly shift the formants frequencies. In this study, whatever the importance of these phenomenons, only their causes on the VTF have to be modeled. The geometry of the structures of the vocal-tract is thus not considered. According to our objectives, modeling the VTF is sufficient.

Transverse modes and plane wave hypothesis

In voice analysis, only the waves which travel perpendicularly to the traveling axis are usually considered. Indeed, one can neglect the transverse waves if their half wave-length are large compared to the section of the cavities of the vocal-tract. For example, for a section of 4 cm, no transverse modes exists below $0.5 \cdot 340/0.04 = 4250$ Hz. Below this frequency the waves can be represented by reflexion line models like in Maeda's synthesizer [Mae82a, MG76]. To study the behavior of transverse waves in a multi-dimensional context, transmission line matrix can be used [EMPSB96]. In this study, in order to reduce the influence of the transverse modes on the proposed analysis methods, only the first 8 kHz are considered and the analyzed voiced signal is thus undersampled at 16 kHz.

Influence of the glottal opening on the vocal-tract filter

The glottal area has a direct influence on the VTF. Indeed, the impedance at the glottis level determines one boundary condition of the vocal-tract tubes. Additionally, this glottal impedance is highly dependent on the glottal area. When the glottal area increases the formants are slightly shifted [BdH07] Moreover, the formants bandwidth also increases since the glottal opening adds an additional acoustic loss [Fla72a]. Consequently, in a duration of a single pulse, the VTF is time-varying. However, in this study, the variation of the glottal impedance is assumed to be negligible in a single period. Therefore, a stationary filter can be estimated with an analysis window which cover a duration of a few periods (note that this latter duration has to be short enough to assume that the articulatory configuration is stationary).

3.1.2 Minimum-phase hypothesis

It has been shown that if the vocal-tract is approximated by a single tube, the VTF is a stable all-pole filter [MG76]. However, as already discussed in section 3.1.1, pole-zero pairs are added to the VTF in nasalized sounds. Whereas all the poles are obviously inside the unit circle whatever the nasalization, the position of the zeros are more tricky to establish. Lim et al. [LL93] have shown that if the vocal-tract is assumed to be lossless, these zeros are exactly on the unit circle. In this study, we postulate that the losses move these zeros inside the unit circle as the losses move the poles inside the unit circle. Note that the Maeda's synthesizer places also these zeros inside the unit circle (see appendix D). Consequently, using this postulate the impulse response of the VTF is minimum-phase. This hypothesis has already been used in the ZZT separation method [BDdD05] and complex cepstrum decomposition [DBD09, OSS68]. Using the minimum-phase hypothesis, the phase and amplitude spectrums are linked through the Hilbert transform [OS89]. The phase spectrum can be therefore retrieved from the amplitude spectrum using the real cepstrum (see appendix A).

3.2 Lips and nostrils radiation

In the literature, the radiation of the acoustic pressure at the mouth level is usually modeled by vibrating piston set in a baffle which is the head [Fla72a]. For reason of simplicity, the radius of this piston is assumed to be small compared to the radius of the head and this baffle is therefore represented by an infinite plane. Accordingly, the pressure $p(l)$ measured at a given distance l from the mouth can be expressed as [Pie81]:

$$p(l) = \frac{j\omega\rho}{2\pi l} \cdot e^{-j\omega l/c} \cdot U_L \tag{3.1}$$

where U_L is the flow at the lips level which is the glottal flow filtered by the VTF, c is the speed of sound and ρ is the air density. Expression (3.1) is valid only if the distance of the pressure measurement to the mouth is big compared to the radius of the piston a. More precisely, $l \gg a$ and $l \gg a^2\omega/c$. For an important mouth opening in speech, for example $a \approx 1.2\,\text{cm}$ (i.e. /a/), the condition is: $l \gg 0.23 \cdot 10^{-3} \cdot f$ (f is the frequency in Hz). Therefore, this expression is valid for low frequencies only. It is interesting to see that expression (3.1) is close to that obtained for a simple pulsating sphere [Fla72a]:

$$p(l) = \frac{j\omega\rho}{4\pi l} \cdot e^{-j\omega l/c} \cdot U_L \tag{3.2}$$

Indeed, the two expressions differ only in a factor of two. According to these two expressions (3.1) and (3.2), one can see that the pressure is proportional to a time-derivative of the lips flow. In our context, one can therefore make the following simplifications and derive a simple representation of the radiation from these models. First, one can assume that the positions of the speaker and that of the pressure measurement are fixed. Moreover, we are not interested in the absolute position of the speech signal. In the equations above, the delay term, which is constant, can thus be set aside. Then, the other constants, which only influence the absolute gain of the speech signal, can be ignored:

$$\tilde{p} = j\omega \cdot U_L \tag{3.3}$$

Consequently, a time-derivative is often used in speech analysis to model the lips radiation [Lju86, WMG79, RS78, MG76]:

$$L(\omega) = j \cdot \omega \tag{3.4}$$

Figure 3.2 shows the amplitude and phase spectrums of this model. In this study, for the sake of simplicity, this model is used despite of its lack of precision at high frequencies.

Nostrils radiation

As shown, the smaller the opening, the better the approximation of equation (3.4). Therefore, if this model is used for the mouth, this model is appropriate for the nostrils since the nostrils opening is smaller than the mouth opening. Additionally, as presented in the introduction, the source-filter model represents the voice production using convolved terms in the time domain (and thus multiplied in the frequency domain):

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot L(\omega)$$

where $G(\omega)$ is the glottal source, $C(\omega)$ the VTF and $L(\omega)$ represents the overall radiation of the acoustic pressure leaving the vocal apparatus. According to this model, it is also necessary to use the same model for both lips and nostrils radiation.

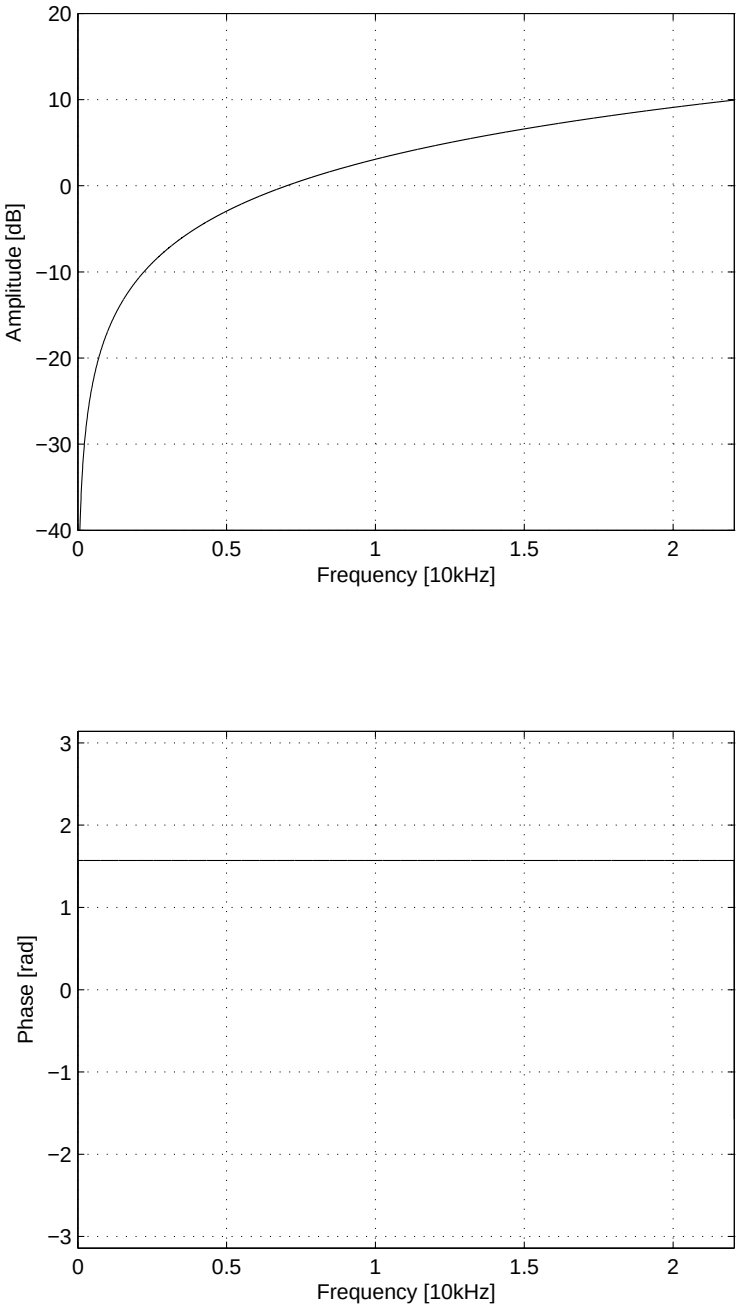


Figure 3.2: The radiation model $L(\omega) = j\omega$

3.3 The complete voice production model

As assumed in the introduction, the linear source-filter model is used in this study. According to the descriptions of the glottal source and the filtering elements given above, one can write the following voice production model in the spectral domain:

$$S(\omega) = \left[e^{j\omega\phi} \cdot H^{f_0}(\omega) \cdot G(\omega) + N^{\sigma_g}(\omega)\right] \cdot C_-(\omega) \cdot L(\omega) \tag{3.5}$$

where:

$G(\omega)$ is the spectrum representing the shape of the glottal pulse which is responsible of the deterministic component of the glottal source. In the following, $G^{\theta}(\omega)$ will represent a glottal model with its shape parametrized by θ.

$e^{j\omega\phi}$ is the linear-phase component of the glottal source which represents the position of the periodic glottal pulses relatively to a zero-time reference. For one pulse modeled by the LF model, the zero-time reference is the t_e instant.

$H^{f_0}(\omega)$ is the harmonic structure modeling an impulse train of fundamental frequency f_0. Therefore, $H^{f_0}(\omega) = \sum_{k\in\mathbb{Z}} e^{j\omega k/f_0}$

$N^{\sigma_g}(\omega)$ is the spectrum of the aspiration noise, the random component of the glottal source. In the analysis part, this noise is assumed to obey a Gaussian distribution of standard-deviation σ_g. For voice transformation and synthesis, this noise will be also modulated and colored according to section 2.6.

$C_-(\omega)$ is a minimum-phase filter corresponding to the Vocal-Tract Filter (VTF). The minimum-phase property is denoted by the negative sign.

$L(\omega)$ is the filter corresponding to the radiation at the lips and nostrils level. According to section 3.2, $L(\omega) = j\omega$.

Figure 3.3 illustrates the voice production model and its elements.

In this study, we assume that the spectrum of the glottal source can be split into a deterministic frequency band and a random frequency band using a Voiced/Unvoiced Frequency (VUF) (see fig. 3.3). Since the filtering elements are linear functions, the two parts of the speech spectrum can thus be expressed separately:

$$S(\omega) = \begin{cases} e^{j\omega\phi} \cdot H^{f_0}(\omega) \cdot G(\omega) \cdot C(\omega) \cdot L(\omega) & \text{for } \omega < \text{VUF} \\ N(\omega) \cdot C(\omega) \cdot L(\omega) & \text{for } \omega > \text{VUF} \end{cases} \tag{3.6}$$

In order to estimate the shape parameters of the deterministic component, the next part II will consider only the upper part of equation (3.6). The random component will be taken into account in the last part 8.3 to manage the synthesis of the glottal noise. We assume that the VUF is known *a priori* thanks to existing methods [KH07, Sty01]. This value is estimated by determination of voiced/unvoiced frequency bands [Sty01, p.3] by means of peak classification of the speech spectrum [ZRR08]. Compared to a multiband excitation model [GL88] or a Harmonic+Noise Model (HNM) [Sty96], this decomposition in only two separated frequency bands is an important simplification of voice production. However, in the last part of this study, we will see that such a simplification leads to a convenient estimation of the noise level of the glottal source.

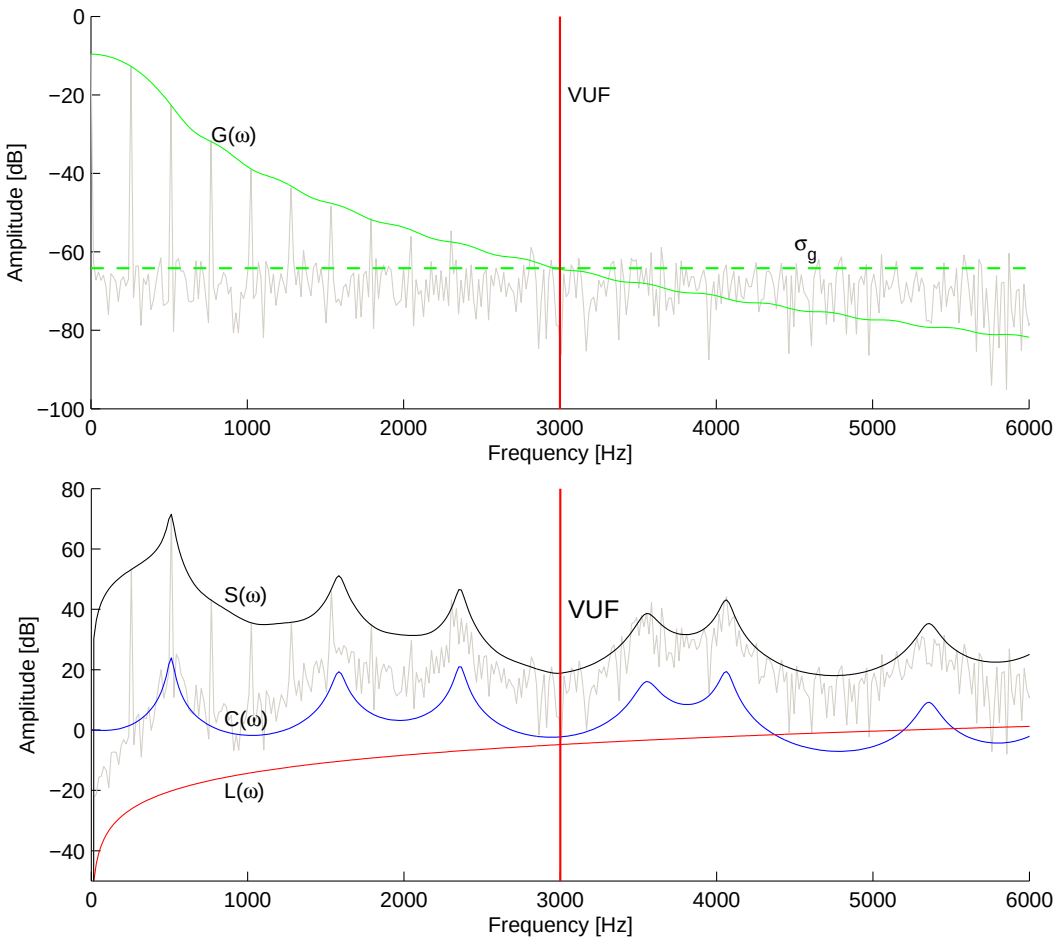


Figure 3.3: The model of the glottal source above and the model of voice production below. The spectrum of one speech period is shown in solid black line and the spectrum of multiple periods is shown in gray for each plot.

The radiated glottal source

Finally, since the voice production model is made of linear operators, the radiation filter can be moved in front of the glottal source:

$$S(\omega) = \underbrace{\left[e^{j\omega\phi} \cdot H^{f_0}(\omega) \cdot G(\omega) + N(\omega)\right] \cdot L(\omega)}_{\text{the radiated glottal source}} \cdot C(\omega) \tag{3.7}$$

In this study, the glottal source mixed with the lips radiation will be termed *the radiated glottal source* and $G(\omega) \cdot L(\omega) = G'(\omega)$ will be termed *the radiated glottal pulse*. Note that by schematizing the time-derivative of the glottal pulse, the LF model describes the shape of the radiated glottal pulse (sec. 2.4).

Conclusions

- The resonances of the VTF can be modeled by poles whereas the coupling with the nasal cavity introduce pole-zero pairs. The poles are obviously stable for reason of passivity of the vocal-tract. However, the position of the zeros is more tricky to establish. In this study, these zeros are assumed to be inside the unit circle. Accordingly, the VTF impulse response is a minimum-phase signal (sec. 3.1.2).
- From the point of view of the perception, the variations of the VTF due to the glottal area variations are assumed to be negligible. Therefore, we assume that the VTF is stationary in a duration where the articulatory configuration is assumed to be stationary (sec. 3.1.1).
- In this study, the model used for the lips radiation is the usual time-derivative $L(\omega) = j\omega$. Note that, this model is valid only for low frequencies and small mouth opening (sec. 3.2). This model is also used for the nostrils radiation since the opening is usually smaller than the one of the mouth. Both lips radiation and nostrils radiation can thus be merged into a single radiation filter.
- In voice analysis, due to the presence of transverse modes at frequencies over $\backsim$ 4 kHz (sec. 3.1.1) and the lack of precision of the radiation model over 5 kHz (sec. 3.2), the analysis is limited to a meaningful frequency band. In the following proposed methods the analyzed speech signal is undersampled at 16 kHz.
- One can assume that the deterministic and random components of the glottal source are disjoint with respect to frequency. Therefore, the model of the voiced spectrum can be expressed by two equations for the two frequency bands, one below and one above a Voiced/Unvoiced Frequency (VUF) (eq. 3.6).

Part II

Voice analysis

Chapter 4

Source-filter separation

In this chapter, we will first discuss the main problems met in voice source and vocal-tract filter separation. Using this framework, we will then present the state of the art.

4.1 General forms of the vocal-tract filter and the glottal source

This section first describes how the vocal-tract filter and the glottal source are related to each other through the voice production model. Although the following relations does not lead to a particular result, they will be useful for the next discussions.

The widely used inverse filtering technique [Mil59] allows to write the next two expressions from the linear source-filter model (3.5). Firstly, focusing on the deterministic component, the glottal source can be estimated from the VTF and the radiation by dividing in the frequency domain:

$$\tilde{G}(\omega) \cdot e^{j\omega\phi} \cdot H^{f_0}(\omega) = \frac{S(\omega)}{C(\omega) \cdot L(\omega)} \tag{4.1}$$

Reciprocally, the vocal-tract filter can be estimated from the glottal pulse and the radiation [1]:

$$\tilde{C}(\omega) = \mathcal{E}\left(\frac{S(\omega)}{G(\omega) \cdot L(\omega)}\right) \tag{4.2}$$

where $\mathcal{E}(.)$ is an estimate of a smooth envelope (LP, DAP, TE, see sec. 4.3). By replacing the observed spectrum $S(\omega)$ with its model in (4.2), one can examine the underlying result of the previous equation:

$$\tilde{C}(\omega) = \mathcal{E}\left(\frac{e^{j\omega\phi} \cdot H^{f_0}(\omega) \cdot G(\omega) \cdot C(\omega) \cdot L(\omega)}{G(\omega) \cdot L(\omega)}\right) = \mathcal{E}\left(H^{f_0}(\omega) \cdot C(\omega)\right) \tag{4.3}$$

Because the impulse response of the envelope estimate is usually assumed to have no delay. The linear-phase component of the source is not modeled by $\mathcal{E}(.)$ in (4.3). Therefore, the estimation of the envelope $\mathcal{E}(.)$ has to deal with the interpolation of the sampling by the harmonics of frequency response of the VTF. More generally, the reconstruction of the VTF in (4.3) is an inverse problem of the estimation of a filter frequency response considering the properties of its excitation source. In order to estimate such

[1] In part 8.3, in the context of voice transformation and speech synthesis, the estimation of the VTF will be extended to the random component of the glottal source.

a response, in system identification, a known white noise or a known sweep tone is usually used as an input of the unknown filter to span as much of the frequencies as possible [FL71]. However, in our case, the excitation source is also an unknown element of this estimation problem. Contrary to white noise or sweep tone, this source does not span all the frequencies uniformly:

- The periodic behavior of the glottal source creates zeros between the harmonic frequencies.
- The radiation creates a zero at the zero frequency in the z-transform of the speech signal.
- The glottal source is band limited (see sec. 2.5.4). One can find a frequency above which the VTF is not excited by the deterministic source.
- Above a given frequency, one can assume that the aspiration noise exceeds the deterministic source. Therefore, the properties of the source change according to a considered frequency band.

In addition, if we assume that the radiation is known, both equations (4.1) and (4.2) are related to each other through $G(\omega)$ and $C(\omega)$. Consequently, these two equations express a joint estimation problem of an unknown filter excited by a sparse source. In the current literature, on one hand either the source or the filter can be simplified in order to approximate the other one and on the other hand, joint estimation methods are also proposed.

Stationarity hypothesis

Considering this estimation problem, the duration of the observed signal can not be arbitrary. Indeed, the voiced signal is highly non-stationary from two different aspects. On the one hand, the glottal source can vary quickly from one period to the next (e.g. vocal fry). On the other hand, the articulators are rarely in a sustained position. They continuously move from one configuration to another, from one phone to the next. Consequently, if, for reasons of simplicity, the stationary hypothesis is used, the duration of the observed signal has to be no longer than a few periods. We should note that, within such a small window, a few elements of the voice production are not accessible using a Fourier transform. For example, it is not possible to observe the aspiration noise between the lowest harmonics, mainly because the main lobe of the harmonics masks this noise. Additionally, it is not possible to increase the harmonic-to-noise ratio just by increasing the window duration.

4.2 Estimation of glottal parameters

In this study, we will assume that the glottal pulse $G(\omega)$ obeys to a given glottal model $G^{\theta}(\omega)$ parametrized by θ. Firstly, one can try to estimate θ directly from the speech signal without separating the source and the filter. For example, the difference of the first two harmonics (H1-H2) have been shown to be highly correlated to the open quotient [STAV02, Han97, Han95]. Therefore, it has been proposed to approximate the open quotient by this difference. However, comparisons with EGG signals have shown a poor reliability of this estimate [HdD01]. Secondly, by means of the Complex Cepstrum (CC) [OSS68] or the zeros of the z-transform (ZZT) [Boz05], the speech signal can be decomposed into its minimum-phase and maximum-phase components. Then, a glottal model can be fitted on the resulting maximum-phase component which is assumed to be the glottal pulse. The IAIF separation method [Alk92] uses another separation principle by means of all-pole models. Similar to the CC and ZZT methods, the glottal parameters have to be estimated in a second step, by fitting the glottal model on the estimated glottal source (see Appendix B). However, in these methods, even though the glottal parameters are estimated

using a source separated from the filter, these two are not jointly estimated with the parameters of the VTF model. Thirdly, one can do this joint estimation of the glottal parameters and the VTF parameters according to the mutual dependence of equations (4.2) and (4.1). Methods based on ARX and ARMAX models aim to solve such a problem [Lju86, Hed84].

Necessary constraints of the VTF

In any context of estimation, this is necessary to ensure that the estimated θ parameters tends to the optimal parameter θ^* when minimizing a given error function. As shown in the following, this convergence property of the estimation methods is far from obvious and often implicitly postulated.

Given the context of joint estimation described above, one needs to ensure that the parameters of the VTF model can not compensate an error of the glottal parameters or inversely. From equation (4.2), the VTF expression related to a glottal parameter θ is:

$$\tilde{C}^{\theta}(\omega) = \mathcal{E}\left(\frac{S(\omega)}{G^{\theta}(\omega)\cdot L(\omega)}\right) \tag{4.4}$$

Then, if we assume that the real glottal pulse $G(\omega)$ can be fitted by the used glottal model $G^{\theta}(\omega)$, we can replace the observed speech spectrum $S(\omega)$ by its model (upper part of equation (3.6)):

$$\tilde{C}^{\theta}(\omega) = \mathcal{E}\left(\frac{e^{j\omega\phi}\cdot H^{f_0}(\omega)\cdot G^{\theta^*}(\omega)\cdot C(\omega)\cdot L(\omega)}{G^{\theta}(\omega)\cdot L(\omega)}\right) = \mathcal{E}\left(H^{f_0}(\omega)\cdot\frac{G^{\theta^*}(\omega)}{G^{\theta}(\omega)}\cdot C(\omega)\right) \tag{4.5}$$

In the following, the envelope estimation $\mathcal{E}(.)$ is assumed to perfectly fit the harmonic frequencies of its argument. Therefore,

$$\tilde{C}^{\theta}_h = \mathcal{E}\left(\frac{G^{\theta^*}_h}{G^{\theta}_h}\cdot C_h\right) \qquad \text{with } X_h = X(h\cdot f_0) \quad h\in\mathbb{N} \tag{4.6}$$

However, in order to avoid compensation effect between the VTF parameters and the glottal parameters, it is necessary that $\mathcal{E}(.)$ is not an identity function. Indeed, if $\mathcal{E}(.)$ has no constraints (ie. $\tilde{C}^{\theta}_h = C_h\cdot G^{\theta^*}_h/G^{\theta}_h$), one can show that the following modeling error is not sensitive to an error of the glottal parameters (in the following equations, to focus on glottal parameters, the position of the glottal pulse is assumed to be known *a priori*):

$$\begin{aligned} E^{\theta}_h &= S_h - G^{\theta}_h\cdot\tilde{C}_h\cdot L_h \\ &= e^{jh\phi^*}\cdot G^{\theta^*}_h\cdot C_h\cdot L_h - e^{jh\phi^*}\cdot G^{\theta}_h\cdot\tilde{C}^{\theta}_h\cdot L_h = e^{jh\phi^*}\cdot C_h\cdot L_h\cdot\left(G^{\theta^*}_h - G^{\theta}_h\cdot\frac{G^{\theta^*}_h}{G^{\theta}_h}\right) = 0 \quad \forall\theta \end{aligned}$$

Moreover, one can show the same by maximizing the whitening of the convolutive residual:

$$R^{\theta}_h = \frac{S_h}{e^{jh\phi^*}\cdot G^{\theta}_h\cdot\tilde{C}_h\cdot L_h} = \frac{e^{jh\phi^*}\cdot G^{\theta^*}_h\cdot C_h\cdot L_h}{e^{jh\phi^*}\cdot G^{\theta}_h\cdot\tilde{C}^{\theta}_h\cdot L_h} = \frac{G^{\theta^*}_h\cdot C_h}{G^{\theta}_h\cdot(G^{\theta^*}_h/G^{\theta}_h)\cdot C_h} = 1 \quad \forall\theta$$

Therefore, it is necessary to have some hypothesis on the VTF properties which are different to the hypothesis on the glottal pulse properties. The envelope estimate has thus to respect these hypothesis to insure that the estimation of the VTF parameters does not compensate an error of the glottal parameter or inversely. Mathews et al. [MMEED61] proposed that the VTF is an all-pole filter and thus the source is

made only of zeros. This all-pole hypothesis has been widely used and the Linear Prediction (LP) method using autocorrelation satisfies this hypothesis [MG76]. More recently, some constraints on formants have been used to avoid any modeling of the source spectral characteristics by the VTF [SKA09]. However, using ARMA models and the covariance methods which gives mixed-phase estimate of the VTF, it is not clear which constraint ensures the convergence of the estimated parameters of the voice model to the optimal ones.

4.3 Spectral envelope models

Using the expression of the VTF in equation (4.2), the main problem is to estimate an envelope using $\mathcal{E}(.)$ which has to deal with the sampling of the VTF by the harmonic structure as seen in equation (4.3). In this section, we will enumerate the known means to estimate such an envelope.

- AutoRegressive models (AR)

  If the nasalization is not considered, it is possible to model the VTF with an all-pole filter as it is proposed by the following methods. Two different methods have been proposed to estimate an AR model by **Linear Prediction (LP)**. The autoregressive coefficients can be computed using the Levinson-Durbin recursion on the **autocorrelation** function or these coefficients can be computed by using a matrix decomposition method on the **covariance** matrix [MG76, AH71]. Note that the envelope given by the autocorrelation is always stable because of the Levinson-Durbin recursion. Conversely, the covariance solution is sensitive to the instability of the signal (e.g. see also closed-phase covariance methods below). An envelope using LP-covariance can be unstable which does not correspond to the passivity assumption of the VTF [Mil86]. Therefore, in terms of signal properties, the difference between the VTF and the glottal model is not clear. The LP-autocorrelation minimize the Itakura-Saito distance between the analyzed spectrum and the envelope while the LP-covariance minimize the mean squared prediction error [MG76]. None of these two solution align the estimated envelope on the peaks of the harmonics. However, because of the stability of the LP-autocorrelation and its low computational cost, this method is the most used envelope in voice analysis. Indeed, this envelope is widely used in current telecommunications by modeling its convolutive residual with codebooks [SA85, AR82]. In order to reduce the lack of precision of the LP, the **Discrete All-Pole (DAP)** fits only the harmonic frequencies of an observed spectrum [RVR07, EJM91]. We should also notice that time-varying AR solutions have been proposed to deal with the non-stationarity of the vocal-tract filter [SL09, SL08].

- AutoRegressive Moving Average models (ARMA)

  Since zeros can exist in the VTF due to the nasalization, it is interesting to use a mixture of poles and zeros to model the VTF. However, the linear solutions for AR modeling can not be easily extended to pole-zero modeling because this latter is nonlinear [Ste77]. Nevertheless, by inversion of the speech spectrum, a few techniques exist to use AR solutions to model the valleys made by the zeros as it was formants [KOT77, Mak75]. Moreover, by using a slightly different error function, the problem can be linearized (see ARMAX models in sec. 4.4.4). Konvalinka & Matausek [KM79] therefore proposed an ITerative Inverse Filtering method (ITIF) to estimate both AR and MA models. Note that, in this method, the filters coefficients are estimated by solving linear equations using a Cholesky decomposition. The proposed solution is thus a covariance-like method and the position of the zeros of the filters are not constrained to the interior of the unit circle like in a stable and minimum-phase filter. Like for AR solutions, a time-varying solution of the ARMA model has been proposed [MMN86]

- Cepstral Envelopes

  As discussed above, the main goal of the envelope estimation is to eliminate the harmonic structure of the speech spectrum. Therefore, these fine structures can be removed by truncating the real cepstrum of a speech signal in order to obtain a smoothed log-magnitude spectrum. However, the cepstral envelope is sensitive to the spectral content between the harmonics like the linear prediction. To overcome this problem, the following two variations of the cepstral envelope exist. Additionally, it is important to notice that, conversely to AR and ARMA models, the optimal order of the cepstral envelopes is known and equal to $0.5 \cdot f_s/f_0$ [RVR07] (f_s is the sampling frequency).

  True-Envelope (TE): Initially proposed by Imai and Abe in a Japanese publication [IA79], this method reduces iteratively the distance between the estimated envelope and the partials peak. After convergence, the final envelope lies on the summit of the harmonics. This process is known to be relatively slow due to the computation of the discrete cosine transform at each iteration. Nevertheless, the computational time can be significantly reduced to make this estimation process nearly as fast as the LP-autocorrelation [RR05].

  Discrete Cepstrum Envelope (DCE): In the same way the discrete all-pole method has been proposed for all-pole modeling to fit a given set of spectral amplitudes at given frequencies, the DCE has been proposed for the cepstral envelope [GR90].

  A causal and time-limited impulse response of the VTF can be retrieved through the estimation of a cepstral envelope. Then, the estimation of an AR or ARMA model of this latter can be reduced to a classical system identification which can be solved by different methods (e.g. Kalman, Shanks) [Jac89, Ste77]. The method using **True-Envelope-Linear-Prediction (TE-LP)** has been proposed to retrieve a stable all-pole model of the TE envelope [VRR06]. One may argue that a cepstral envelope is difficult to interpret in terms of formant and acoustic properties of the vocal-tract. Indeed, the AR models provide a way to obtain formant frequencies and bandwidths through their poles as well as an approximation of the vocal-tract area function [Den05, MG76]. However, as shown by the TE-LP method, the TE can be used to properly eliminate the harmonic structure of the speech spectrum using the optimal order given above. Then, considering the all-pole hypothesis, the formant properties can be estimated using the linear prediction methods.

- Spectral Envelope Estimation VOCoder (SEEVOC)

  The interpolation problem of the harmonics can be explicitly managed using the existing interpolation techniques. Splines can be used instead of the periodic functions used by the cepstral envelopes [KNZ97, Pau81]. Note that this type of interpolation is fast enough to be used in many applications [Bon08, Bon04].

- Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT)

  If a segment of speech signal is windowed (ie. time modulated), the harmonics of the speech spectrum are convolved by the spectral representation of the window. The idea proposed in STRAIGHT is to use a particular window which interpolate the harmonic peaks of the speech spectrum by means of the summation of the main lobes of the window [KMKd99, Kaw97]. Using this idea along the frequency axis, as well as for smoothing the time variation of the envelope, an envelope of the amplitude spectrum can thus be built using a simple DFT. The minimum-phase spectrum of this envelope is then retrieved from the amplitude spectrum of the latter [OS78]. This method has been reformulated using two analysis windows (Tandem-STRAIGHT) [KMT+08].

4.4 The state of the art of source-filter separation and estimation methods of voice production models

Different measurements have been proposed to separate the vocal-tract properties of the glottal source properties. The oral flow can be measured by a mask [Rot73]. Note that conversely to most of the other measures, this mask allows to obtain the DC component of the flow. The electroglottography is mainly used to detect instant of closure of the glottis [TN09, HdDC04] but can also be used to estimate the open quotient [SdD06, HdDC04]. Obviously, in this study, the acoustic pressure waveform is used in order to transform a given voice recording or synthesize a voice from a large database of this same measurement.

In the following sections, the existing separation methods are presented to the best of the author's knowledge.

4.4.1 Analysis-by-Synthesis

The *Analysis-by-Synthesis* process has been initially proposed by Stevens [Ste60, BFH$^+$61]. The initial idea is to store a codebook of 6 source spectra and 24 resonance spectra and to find the best combination which minimizes the mean squared log amplitude difference according to an observed amplitude spectrum. Stevens built these codebooks from synthetic spectra or spectra obtained by manual inverse filtering (see sec. 4.4.3). Conversely, using the same log difference, Mathews et al. [MMEED61] proposed to estimate the parameters of an ARMA model assuming the source and the filter are made of zeros and poles respectively. Once the glottal source is estimated by closed-phase analysis (see below), Oliveira also proposed to estimate the parameters of a glottal model by minimization of the amplitude difference of harmonic models [Oli93]. More recently, Shue et al. [SKA09] proposed to use a codebook only for the source. For each element of the codebook, the VTF is estimated according to equation (4.2) using constraints on formants frequencies and bandwidths [HM95]. Note that, in all of these methods, only the amplitude spectrum of the observed signal is considered to estimate both source and filter. However, the source is not assumed to be constant contrarily to the next approach.

4.4.2 Pre-emphasis

In a first approximation, since the glottal pulse is mainly low-pass, its amplitude spectrum can be represented with a $-12dB/\text{octave}$ [MG76]. Therefore, its spectral amplitudes can be modeled with a double real pole:

$$G(\omega) = \frac{1}{(1-\mu e^{-j\omega})^2}$$

with μ close to unity. Additionally, the radiation can be approximated with a real zero ν close to unity (if $\nu = 1$ this model is an approximation of the time-derivative model defined in the frequency domain (sec. 3.2)):

$$L(\omega) = 1 - \nu e^{-j\omega}$$

Consequently, if we assume $\nu = \mu$, one can cancel both contributions of the glottal source and the radiation by simply pre-emphasizing the speech signal. From equation (4.2), the VTF estimate can thus be obtained by the following equation:

$$\tilde{C}(\omega) = \mathcal{E}\left(\frac{S(\omega)\cdot(1-\mu e^{-j\omega})^2}{1-\nu e^{-j\omega}}\right) = \mathcal{E}(S(\omega)\cdot(1-\mu e^{-j\omega})) \tag{4.7}$$

The LP-autocorrelation method was initially proposed to compute $\mathcal{E}(.)$ [MG76]. This pre-emphasis process is widely used in voice analysis because of its simplicity.

In order to estimate the glottal source properties, according to equation (4.1), the glottal source, can be estimated:

$$\tilde{G}(\omega) = \frac{S(\omega)}{\tilde{C}(\omega) \cdot (1 - \nu e^{-j\omega})} \tag{4.8}$$

Then, using a second order linear prediction model, a complex pair of poles can be fitted on $\tilde{G}(\omega)$ to estimate the glottal formant [HDd99, DdD97]. Moreover, using the analytical definition of the glottal formant of the KLGLOTT88 glottal model, the open quotient O_q can be estimated.

Minimum curvature criterion

It is interesting to see that Milenkovic [Mil86] proposed to use the pre-emphasis with a very different argumentation. Indeed, the additional zero of the pre-emphasis can be merged with the one of the radiation:

$$S(\omega) \cdot (1 - \nu e^{-j\omega}) = \left[G(\omega) \cdot L(\omega) \cdot (1 - \nu e^{-j\omega})\right] \cdot C(\omega) = \underbrace{\left[G(\omega) \cdot (1 - \nu e^{-j\omega})^2\right]}_{Excitation} \cdot C(\omega)$$

By minimizing the energy of the *excitation* given above, the 2nd order time derivative of the source is minimized. Therefore, it's a way of assuming that the estimated glottal source is of a minimum curvature. According to [Mil86], the AR coefficients and the parameters of a glottal model are optimized with a covariance-like LP method in order to minimize this energy.

Iterative Adaptive Inverse Filtering (IAIF/PSIAIF)

Alku [ATN99, Alk92] proposed to first obtain a rough envelope of the glottal source using a first order all-pole model of the speech signal. By inverse filtering the speech signal with this source estimate, the VTF is then estimated. Consequently, these first steps are equivalent to a pre-accentuation with a ν coefficient adapted to the observed speech signal. Follows, a more precise estimate of the source which is obtained using an all-pole model of order 2 or 4 of the inverse filtered speech signal by the VTF estimate. Finally, this process can be repeated in order to iteratively improve the glottal source envelope and the VTF estimate. Although the VTF was initially estimated using linear prediction, Alku and Vilkman [AV94] proposed to use the DAP method.

Adaptive Estimation of the Vocal-Tract (AEVT)

One of the main characteristics of the glottal pulse is the glottal formant (see sec. 2.5.2). In a first approximation, this formant can be assumed to be disjoint from the vocal-tract formants. Consequently, a high-pass filter can be used to eliminate the influence of the glottal formant on the VTF estimate [AM05]. One can consider that this method is close to the procedures using pre-emphasis, because the time-derivative is a first order high-pass filter. Therefore, the AEVT method improves the pre-emphasis process using a more accurate filter design and an adapted cutoff frequency.

4.4.3 Closed-phase analysis

From acoustic pressure measurements, one of the first techniques of source-filter separation consists of formants canceling using an all-zero analog filter ($\approx 1/C(\omega)$). The parameters of this filter was manually tuned to flatten the closed-phase of the glottal pulse and minimize the formant ripples on the open-phase [Mil59]. Then, many methods followed this approach trying to make this process automatic. The main idea in closed-phase analysis is to estimate the VTF in the time segment where the glottis is assumed to be closed because such an analysis is assumed to provide more precise and stable analysis of the VTF than using the whole glottal cycle. However, the counterpart is the localization problem of this closed phase in a cycle. Consequently, many different criteria have been proposed to localize this closed phase: by detecting stationarity segments of formants modulation [PQR99], using the Hilbert envelope of the LP error [AY79], or using GCI and GOI estimation methods (either from electroglottographic signals [VB85] or from the speech signal directly (see chapter 7).

Wong and Markel initially assumed that during the closed phase, the speech signal represents the vocal-tract filter impulse response without any influence of the glottal pulse and the radiation [WMG79]. Therefore, the LP-covariance method is used to identify poles representing the free oscillations. However, if a given time segment represents only the vocal-tract filter, it implies that the excitation source is equivalent to a Dirac delta, or at least, the excitation source has a flat amplitude spectrum if we assume that the envelope method disregards any phase distortion. Consequently, as argued in [WMG79], the amplitude spectra of the glottal pulse and the radiation seem not to be explicitly taken into account. However, using an analysis window smaller than a period, the envelope method can be significantly influenced by the DC component of the window content. Accordingly, in closed-phase analysis, any real pole of the estimated VTF are usually removed and the VTF is thus emphasized. Therefore, both contribution of the glottal source and the radiation can be implicitly removed in the VTF estimate.

Note that, Plumpe et al. [PQR99] changed the argumentation of the closed-phase analysis by assuming that the source-tract interactions are the most negligible in the closed phase. In other terms, the validity of the source-filter hypothesis is maximized in this time segment [MT08].

4.4.4 Pole or pole-zero models with exogenous input (ARX/ARMAX)

The **ARX** model (AutoRegressive with eXogenous input) assumes that the VTF is an AR system ($C(z) = 1/A(z)$). Then, the input is given by a glottal model parametrized by a set of parameters θ (conversely to LP methods wherein a white noise or an impulse train is assumed):

$$S(z) = G^{\theta}(z) \cdot \frac{1}{A(z)}$$

To estimate the autoregressive coefficients a_i of $A(z)$, the following error is minimized:

$$E_{\text{ARX}}(z) = A(z) \cdot S(z) - G^{\theta}(z) \tag{4.9}$$

A matrix representation and a QR decomposition can be used to compute the optimum coefficients a_i minimizing the energy of $E_{\text{ARX}}(z)$ [VRC05a, Hed84]. Therefore, this method is similar to the LP-covariance method but using a given input vector [BRW10, Mil86, Lju86]. Generally, the glottal parameters θ can not be estimated by the QR decomposition because their are in a nonlinear relation with the temporal representation of the glottis. Therefore, the glottal parameters have to be optimized using a minimum search algorithm (e.g. interior-reflective Newton method [FM06, CL93], simplex [VRC05a, Vin07], SUMT

[PB05, Lu02], *hill-climbing*[2] [FL86, Lju86], gradient descent [Hed84]). In order to estimate the glottal parameters with the autoregressive coefficients during the QR decomposition, Milenkovic proposed to use a glottal model which is male of a weighted sum of fixed glottal shapes [Mil86].

Note that, similar to the ARX method where the mean squared error is minimized in the QR decomposition, Frohlich et al. [FMS01] proposed to minimize the Itakura-Saito error of the DAP envelope method to estimate glottal parameters (leading to the method: Simultaneous Inverse filtering and Model matching (SIM)). However, conversely to the ARX approach, the solution is computed in frequency domain taking into account only the amplitude spectrum of the speech signal. Therefore, once the shape parameters are estimated, the time-synchronization is finally estimated in the time domain using the mean squared error.

In order to model the zeros due to nasalization, an **ARMAX** model (AutoRegressive Moving Average with eXogenous input) can be used:

$$S(z) = G^{\theta}(z) \cdot \frac{B(z)}{A(z)}$$

To express the ARMAX solution in a linear form, the following error has to be minimized [Lju86]:

$$E_{\text{ARMAX}}(z) = A(z) \cdot S(z) - B(z) \cdot G^{\theta}(z) \tag{4.10}$$

One last method using exogenous input has to be mentioned. The **Glottal-ARMAX** model is made of a mixed excitation composed of a deterministic and a random part, a KLGLOTT88 model and a white noise respectively [FMT99]:

$$S(z) = \left(G^{\theta}(z) \cdot D(z) + N^{\sigma}(z) \cdot B(z)\right) \cdot \frac{1}{A(z)}$$

where $N^{\sigma}(z)$ is the white noise spectrum of standard-deviation σ, $D(z)$ and $B(z)$ are MA filters and $1/A(z)$ is an AR filter. For each hypothetical glottal parameters, the MIS method (Model Identification System) is used to compute the filter coefficients [MMN86]. Then, the glottal parameters are optimized using a hybrid local search algorithm using both a genetic algorithm and simulated annealing. We should note that, using a mixed input of deterministic and random components, this method jointly estimates the voicing (voiced time segments and VUF) with the other model parameters.

Comments about ARX/ARMAX methods

Among the existing methods, the weighting of the error function can be very different. Indeed, the equation (4.9) shows that the error is weighted by the source amplitude spectrum whereas the error of the ARMAX model is weighted by the source and the MA filter (eq. 4.10). Since the glottal model emphasizes mainly the low frequencies, in order to partly compensate this nonuniform weighting, it has been proposed to pre-emphasize the observed signal and its model [FMT99, Lju86].

The orders of the AR and MA models are unknown as in the case of methods based on linear prediction. Moreover, their optimal values depend on the underlying filter to model. Therefore, the orders of the VTF model are additional variables which can be optimized during the estimation of the voice model parameters [VRC05a].

[2] Although *hill-climbing* is a class of algorithms, the exact method is not described in [Lju86] but the implementation is available in the Appendix of Ljunqvist's Thesis.

Last but not least, the solutions given for ARX/ARMAX models are usually similar to a solution using LP-covariance [BRW10, Mil86, Lju86]. Therefore, the estimation of the VTF have no constraints on the positions of the poles and zeros regarding the unit circle. Consequently, although error functions can be shown [FM06, Lju86], theoretically speaking, it is not clear to what extent the parameters of the VTF model can compensate the glottal parameters (see sec. 4.2). Note that reflecting the unstable poles inside the unit circle during the estimation process can be a way of constraining the representation of the VTF [JS05].

4.4.5 Minimum/Maximum-phase decomposition, complex cepstrum and ZZT

The basic idea of this approach is to separate the maximum-phase and minimum-phase components of the speech signal [PAD10, DBD09, Boz05, OS78, OSS68]. According to sections 2.5.3 and 3.1.2, these two components can be attributed to the glottal source and the VTF respectively if the glottal pulse is assumed to have no minimum-phase component (e.g. Rosenberg glottal model). Compared to the previous approaches, the main advantage of this approach is that a glottal model is not necessary. Indeed, this separation is a natural decomposition of any signal which is meaningful in the case of a speech signal. This separation can be obtained using different methods. The two main approaches are: through the separation of the Zeros of the Z-Transform (ZZT) [Boz05] and through the complex cepstrum [DBD09, OSS68], which can be computed in many different ways [OS78]. Once the anti-causal part of the complex cepstrum is retrieved (i.e. the maximum-phase component of the speech signal), it is possible to estimate different characteristics of the source [SdD06].

4.4.6 Inverse filtering quality assessment

As mentioned in the introduction, the evaluation and validation of the separation processes is a tricky problem because of the lack of precise ground truth. Therefore, one of the recent idea is to use signal properties, either in time or spectral domain, to evaluate the resulting estimations. For example, Gray and Markel [GM74] proposed a *spectral flatness measure* to evaluate the quality of the inverse filtering using linear prediction. Additionally, recent studies proposed various assessment measures termed Glottal waveform Quality Measure (GQM) [BALA05, AABP05] and evaluation of their reliability [MT08, LAB+07]. In the context of estimation of glottal parameters, any GQM becomes a potential way of defining a new error function, which can be minimized with a local search algorithm in order to estimate glottal parameters (like the Itakura-Saito error of the DAP method [FMS01], the covariance error of ARX/ARMAX models [FMT99, Lju86, Hed84], the minimum curvature criterion [Mil86], etc.).

Conclusions and chosen approach

- The voice production model implies a mutual dependency of the estimation of the VTF and of the glottal source (see equations 4.1 and 4.2).
- In this study, conversely to Analysis-by-Synthesis, pre-emphasis, IAIF and AEVT methods, the phase spectrum will be used in the proposed separation methods, like in the ARX/ARMAX based methods and the minimum/maximum-phase decomposition methods (complex cepstrum decomposition and ZZT).
- As proposed in the introduction of this study, in order to separate the source from the filter, the real glottal pulse is assumed to obey a given glottal model, like assumed with ARX/ARMAX methods and not with the minimum/maximum-phase decomposition methods.
- Due to the mutual dependency of the estimation of the VTF and the one of the glottal pulse, a constraint is necessary on the VTF or the glottal pulse to avoid any compensation effect between the parameters of the VTF model and the parameters of the glottal model. Using the stable all-pole model, this condition is satisfied with the LP-autocorrelation because this method is unable to model the mixed-phase property of the glottal pulse (but the zeros of the VTF occurring during nasalization cannot be modeled). In the case of ARX and ARMAX models solved using a covariance-like method, the constraints on the vocal elements are not clear. In this study, like in complex cepstrum decomposition and ZZT, we assume that the glottal pulse and the VTF are mixed-phase and minimum-phase respectively (see 2.5.3 and 3.1.2).
- Finally, one of the main problems of the source-filter separation is the estimation of the VTF by means of interpolation of the harmonic frequencies (eq. 4.3). Accordingly, a harmonic model will be considered for voice analysis, and the True-Envelope (TE) will be used for voice transformation and synthesis.

Chapter 5

Joint estimation of the pulse shape and its position by minimization of both phase distortion and slope

This chapter describes new methods to estimate the shape parameters of a glottal model.In order to make the innovative theoretical ideas as clear as possible about the Mean Squared Phase (MSP), the first section discusses the estimation process and the mathematical derivations without taking into account the details related to the realization. The conditions of convergence will be also discussed at the end of this theoretical section. Next, the realization of the methods using the Mean Squared Phase and the method using the phase difference operator are described. The last section discusses the estimation of multiple shape parameters.

5.1 Phase minimization

One can consider that the proposed idea is close to methods minimizing the phase slope or the group-delay to estimate GCIs [NKGB07, SY95]. The inverse filtering quality measure using the group-delay is also related to this approach [AABP05]. This idea is the following: to estimate the parameters of a glottal model (the shape θ and the position ϕ), the goal is to minimize an error between the observed spectrum $S(\omega)$ and its model $M(\omega)$ parametrized by (θ, ϕ). In order to focus on the phase properties of the speech signal, the phase of the convolutive residual is minimized. First the convolutive residual is expressed as:

$$R(\omega) = \frac{S(\omega)}{M^{(\theta,\phi)}(\omega)}$$

Thus, if a given model has to tend to the observed spectrum, the convolutive residual has to tend to 1:

$$M^{(\theta,\phi)}(\omega) \to S(\omega) \quad \Leftrightarrow \quad R(\omega) = \frac{S(\omega)}{M^{(\theta,\phi)}(\omega)} \to 1 \ \forall\omega$$

which implies that $R(\omega)$ tends to a unit amplitude spectrum and a zero-phase signal:

$$|R(\omega)| \to 1 \quad \text{and} \quad \angle R(\omega) \to 0 \quad \forall\omega$$

Finally, if one can ensure in the optimization method that $|R(\omega)| = 1\ \forall\omega$, then the minimization of the phase is sufficient to make the model tend to the observed spectrum:

$$\angle R(\omega) \to 0 \quad \Rightarrow \quad M^{(\theta,\phi)}(\omega) \to S(\omega) \quad \forall\omega \tag{5.1}$$

To illustrate this phase minimization criterion, the source model will be first considered without filtering elements. For example, one can write the following if the observed signal is assumed to be a linear-phase only (i.e. a Dirac delta with a position ϕ^* in time domain):

$$S(\omega) = e^{j\omega\phi^*} \quad \text{and} \quad M^{\phi}(\omega) = e^{j\omega\phi}$$

where $\phi = \phi^* + \Delta\phi$ is an arbitrary position with an error $\Delta\phi$. Accordingly, the computation of the convolutive residual implies:

$$R^{\phi}(\omega) = \frac{S(\omega)}{M^{\phi}(\omega)} = \frac{e^{j\omega\phi^*}}{e^{j\omega\phi}} = \frac{e^{j\omega\phi^*}}{e^{j\omega(\phi^*+\Delta\phi)}} = e^{-j\omega\Delta\phi}$$

In that case, the minimization of the slope of the convolutive residual ensure that the model tends to the observed spectrum. Using the group delay (or similarly the phase-slope), this idea has been already used for GCI estimation [NKGB07, SY95]. However, in this study, conversely to these methods, the shape parameter of a glottal model has to be estimated, and not only its time position. Therefore, the idea is to add a glottal model to the previous simplified signal model:

$$S(\omega) = e^{j\omega\phi^*} \cdot G^{\theta^*}(\omega) \quad \text{and} \quad M^{(\theta,\phi)}(\omega) = e^{j\omega\phi} \cdot G^{\theta}(\omega)$$

and

$$R^{(\theta,\phi)}(\omega) = \frac{S(\omega)}{M^{(\theta,\phi)}(\omega)} = \frac{e^{j\omega\phi^*} \cdot G^{\theta^*}(\omega)}{e^{j\omega\phi} \cdot G^{\theta}(\omega)} = e^{-j\omega\Delta\phi} \cdot \frac{G^{\theta^*}(\omega)}{G^{\theta^*+\Delta\theta}(\omega)}$$

Therefore, the phase of the residual has a slope which is proportional to the position error $\Delta\phi$ of the glottal pulse and a phase distortion around this slope which represents the shape parameter error $\Delta\theta$. Consequently, by minimization of the phase of the convolutive residual, both slope and the phase distortion around this slope should be minimized together with both position and shape error.

In order to complete the voice production model, the filtering elements can be added. Moreover, within a given window, the speech signal is periodic with a fundamental frequency f_0. Therefore, one can build a discrete spectrum S_h which possesses the complex value of each h-harmonic retrieved in this window. Using this single period representation and according to the filtering elements described in chapter 3, the voice production model of the deterministic component can thus be expressed as:

$$S_h = e^{jh\phi} \cdot G_h \cdot C_h \cdot jh \quad \text{and} \quad M^{(\theta,\phi)}(\omega) = e^{jh\phi} \cdot G_h^{\theta} \cdot C_{h-} \cdot jh \tag{5.2}$$

where the term jh stands for the radiation in the harmonic model[1]. Firstly, from the general form of the VTF estimation (eq. 4.2) and according to the minimum-phase hypothesis of the VTF, one can write:

$$\tilde{C}_{h-}^{\theta} = \mathcal{E}_{-}\left(\frac{S_h}{G_h^{\theta} \cdot jh}\right) \tag{5.3}$$

[1] Although this harmonic notation is unconventional, it will make the notation simpler and easier to read.

where the operator $\mathcal{E}_-(.)$ is computed using the real cepstrum (see [OS78] or appendix A). Secondly, this VTF expression can be replaced in the voice production model (5.2) to derive the convolutive residual $R_h^{(\theta,\phi)}$:

$$R_h^{(\theta,\phi)} = \frac{S_h}{e^{jh\phi} \cdot G_h^\theta \cdot \mathcal{E}_-(S_h/G_h^\theta \cdot jh) \cdot jh} \tag{5.4}$$

In the first method proposed in this chapter, the Mean Squared Phase (MSP) of this convolutive residual is minimized to obtain the optimal parameters which best fit the observed spectrum:

$$\mathrm{MSP}(\theta, \phi, N) = \frac{1}{N} \sum_{h=1}^{N} \left(\angle R_h^{(\theta,\phi)} \right)^2 \tag{5.5}$$

It is interesting to investigate the computation of the residual to show that the conditions of the phase minimization criteria are ensured. $\mathcal{E}_-(.)$ is multiplicative (i.e. $\mathcal{E}_-(A \cdot B) = \mathcal{E}_-(A) \cdot \mathcal{E}_-(B)\ \forall |A|, |B| \geq 0$, see also appendix A). Therefore equation (5.4) can be rewritten as:

$$R_h^{(\theta,\phi)} = e^{-jh\phi} \cdot \frac{S_h}{\mathcal{E}_-(S_h)} \cdot \frac{\mathcal{E}_-(G_h^\theta)}{G_h^\theta} \cdot \frac{\mathcal{E}_-(jh)}{jh} \tag{5.6}$$

One can see that the computation of the residual flattens the amplitude spectrum of S_h, G_h^θ and jh by their respective minimum-phase version. $R_h^{(\theta,\phi)}$ has a unit amplitude spectrum whatever the chosen glottal model and its parameters: $|R_h^{(\theta,\phi)}| = 1\ \forall k\ \forall \theta\ \forall \phi$. Moreover, any error of the parameters changes only the phase of $R_h^{(\theta,\phi)}$. The necessary condition for the phase minimization criterion is thus satisfied (eq. 5.1). Additionally, the MSP error function is not sensitive to the excitation amplitude of the glottal model. The shape parameter can thus be estimated independently of the amplitude parameter.

Using the LF^{Rd} glottal model, figure 5.1 shows an example of $\mathrm{MSP}(Rd, \phi, 12)$ computed on a synthetic speech signal (see chapter 8 for more details on the synthesis).

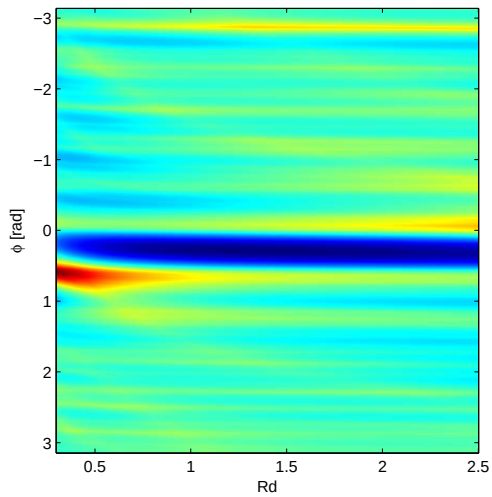


Figure 5.1: Example of $\mathrm{MSP}(Rd, \phi, 12)$ computed on a synthetic signal. The colder the color, the smaller the mean squared phase. The optimal parameters are $Rd \approx 1.5$ and $\phi \approx 0.3$.

5.1.1 Conditions of convergence

In this section, in order to discuss the convergence conditions of the optimization process, the shape of the real glottal pulse G_h is assumed to be correctly represented by the chosen glottal model $G_h^{\theta^*}$ with an optimal parameter θ^*. Indeed, it is important to know which properties of the glottal model are necessary to ensure the convergence of (θ, ϕ) to the optimal parameters (θ^*, ϕ^*) when the MSP error function is minimized. Therefore, in the computation of the convolutive residual (5.4), the observed spectrum S_h can be replaced by the voice production model with the optimal parameters:

$$R_h^{(\theta,\phi)} = e^{jh(\phi^*-\phi)} \cdot \frac{G_h^{\theta^*} \cdot C_{h-}^*}{G_h^{\theta} \cdot \mathcal{E}_-(G_h^{\theta^*} \cdot C_{h-}^* / G_h^{\theta})} \tag{5.7}$$

Then, by distributing $\mathcal{E}_-(.)$ to the terms of its argument, the VTF terms cancel from the previous equation because one can assume $\mathcal{E}_-(C_{k-}^*) = C_{k-}^*$. Therefore, equation (5.7) reduces to

$$R_h^{(\theta,\phi)} = \underbrace{e^{jk(\phi^*-\phi)}}_{\textit{position error}} \cdot \underbrace{\frac{G_h^{\theta^*} \cdot \mathcal{E}_-(G_h^{\theta})}{G_h^{\theta} \cdot \mathcal{E}_-(G_h^{\theta^*})}}_{\textit{shape error}} \tag{5.8}$$

First, according to equation (5.8), note that the error function of equation (5.5) is periodic with respect to ϕ since the *position error* term is periodic. Therefore, looking for an optimal position in the interval $[-\pi; \pi[$ is sufficient.

Secondly, a condition has to be expressed which, if satisfied, ensures that the shape parameter influences the *shape error*: The zeros inside the unit circle in $G_h^{\theta^*}$ and G_h^{θ} are always canceled by their corresponding $\mathcal{E}_-(.)$ expressions. However, a zero outside of the unit circle in $G_h^{\theta^*}$ can be canceled only by G_h^{θ}. Consequently: *θ influences $R_h^{(\theta,\phi)}$ if θ influences at least one zero outside of the unit circle in G_h^{θ}.*

Finally, a condition has to be expressed which, if satisfied, ensures that the shape and the position do not offset each other, at least theoretically: it is sufficient to ensure that the *shape error* has no linear-phase component. G_h^{θ} has a linear-phase which depends on the zero-time reference given by the definition of the glottal model. Therefore, if θ influences that linear-phase component, a residual linear-phase exists in $G_h^{\theta^*}/G_h^{\theta}$ which biases the position error. To avoid the offset effect, the condition is: *θ does not influence the linear-phase component of the glottal model.* Note that using the Liljencrants-Fant model, this condition is satisfied if the zero-time reference is set to the t_e instant (see sec. 2.4).

5.1.2 Measure of confidence

The maximum of the Mean Squared Phase (MSP) is bounded to π^2. Indeed, this function cannot exceed this value because it is assumed to be an average of squared values in $[-\pi; \pi[$. Therefore, a measure of confidence can be proposed to evaluate how close the estimated model is to the observed signal.

$$\psi(\theta, \phi, N) = 1 - \frac{\sqrt{\mathrm{MSP}(\theta, \phi, N)}}{\pi} \in [0; 1] \qquad \text{with } \angle(.) \in [-\pi; \pi[\tag{5.9}$$

Accordingly, it is possible to evaluate how erroneous is the parameter estimate due to the presence of noise or due to a difference between the glottal model and the real glottal pulse.

5.1.3 Polarity

Regarding the MSP, the polarity of the analyzed signal is important. Indeed, if the observed signal is multiplied by -1, it implies a rotation of π of its phase spectrum. Therefore, whereas the estimation of the VTF is independent on the polarity of the observed signal (equation (5.3) is computed only from the amplitude spectrum of its argument), the polarity of the observed signal has to correspond to the polarity of the glottal model in order to estimate the parameters of this latter. Moreover, depending on the recording device, one can not ensure that the polarity of the observed signal is always the same. At least, one can assume that the polarity does not change in a single recording.

There are two different ways to deal with this issue. First, one can define a new angle function whose absolute value wrap above $\pi/2$ in order to create an MSP function which is insensitive to the signal polarity:

$$\bar{\angle}(X) = \begin{cases} \angle(X) & \text{if } |\angle(X)| < \pi/2 \\ \angle(-\overline{X}) & \text{if } |\angle(X)| > \pi/2 \end{cases} \qquad \text{with } \angle(X) \in [-\pi; \pi[\tag{5.10}$$

Consequently, the function $\bar{\angle}(X)$ tends to zero if $\angle(X)$ tends to either zero or π. However, this solution has a drawback. Using this new angle function, the behavior of the MSP is the same for a sinusoidal component of the glottal model which is fully out of phase or perfectly in phase compared to the observed spectrum. Roughly speaking, a perfect match or a perfect mismatch of a component of the model is equally evaluated by the MSP. Therefore, if $\bar{\angle}(.)$ is used, it is assumed that the maximum of the phase error of the glottal model is $\pi/2$. In other terms the glottal model is assumed to be twice closer to the real glottal pulse than using the usual angle function $\angle(.)$.

A method estimating the polarity of a given recording would be a second solution. For example, a method estimating glottal parameters using the MSP and the usual $\angle(.)$ function can be run on both the original recording and its reversed version. Then, the estimates with highest average confidence can be kept. In the following, the polarity will be assumed to be known *a priori*. In case of doubt, the second solution is used.

5.1.4 The iterative algorithm using MSP

In this section, a method is described which uses the Mean Squared Phase (MSP) to estimate the parameter of the LFRd model. First, the spectrum of a voiced segment is computed with a blackman window and the Discrete Fourier Transform (DFT). A window of only one period would estimate the complex coefficients S_h directly. However such a duration is not suitable since the convolutive effect of the window in the spectral domain has to be negligible compared to the harmonic amplitudes and phases of the underlying signal we need to represent. Therefore, 4 periods are used and a harmonic model is built from the DFT of these periods [MQ86]. The amplitude and phase of the hth-harmonic are obtained using the amplitudes and phases of the neighbor bins of $h \cdot f_0$ in the DFT. A parabola is fitted to the amplitudes of the bins to estimate the harmonic amplitude and the harmonic phase is obtained by linear interpolation. Finally, the LFRd glottal model synthesizes directly both terms $G_h^\theta \cdot jh$ in equation (5.4),

To minimize MSP(Rd, ϕ, N), since only two variables are estimated, only a small number of harmonics should be necessary to find the global minimum. However, the glottal model do not perfectly correspond to the real glottal pulse. Therefore, an average solution with the different contributions of all the available harmonics is preferable. N is therefore set to $\lfloor f_{lim}/f_0 \rfloor$, where f_{lim} is the Nyquist frequency, the Voiced/Unvoiced Frequency (VUF) or any other meaningful frequency. As one can see in figure 5.1, the error function corresponding to a linear-phase deviation is a deep and narrow valley embedded in a noisy neighborhood. In such a context, the search for the global minimum is difficult. However, the high frequency behavior of the error function comes from the high frequencies of the convolutive residual. Therefore, to smooth down the error function, the MSP can be first limited to the lowest harmonics (e.g. $N = 3$ seems a good compromise). Then, a Preconditioned Conjugate Gradient (PCG) algorithm [CL96, CL94] is used to find the nearest minimum of the error function from starting values. Then, N is increased one harmonic at a time up to its maximum value while using the PCG algorithm at each incrementation to refine (Rd, ϕ) obtained at the preceding step (see Algorithm 1 and figure 5.2). Note that, since initial values are necessary to start this optimization method, the results of this estimation depend on the choice of these values.

Algorithm 1 Iterative algorithm using MSP(Rd, ϕ, N)

```
Build S_h using a sinusoidal model
Initiate Rd and φ with rough estimates
for N = 3 to ⌊f_lim/f_0⌋ do
  repeat
    Synthesize G_h^Rd · jh with LF model and Rd
    Compute the VTF C_h-^Rd with eq. (5.3)
    Compute convolutive residual R_h^(Rd,φ) with eq. (5.4)
    Compute MSP(Rd, φ, N) with eq. (5.5)
  until PCG algorithm find a minimum of MSP(Rd, φ, N)
end for
```

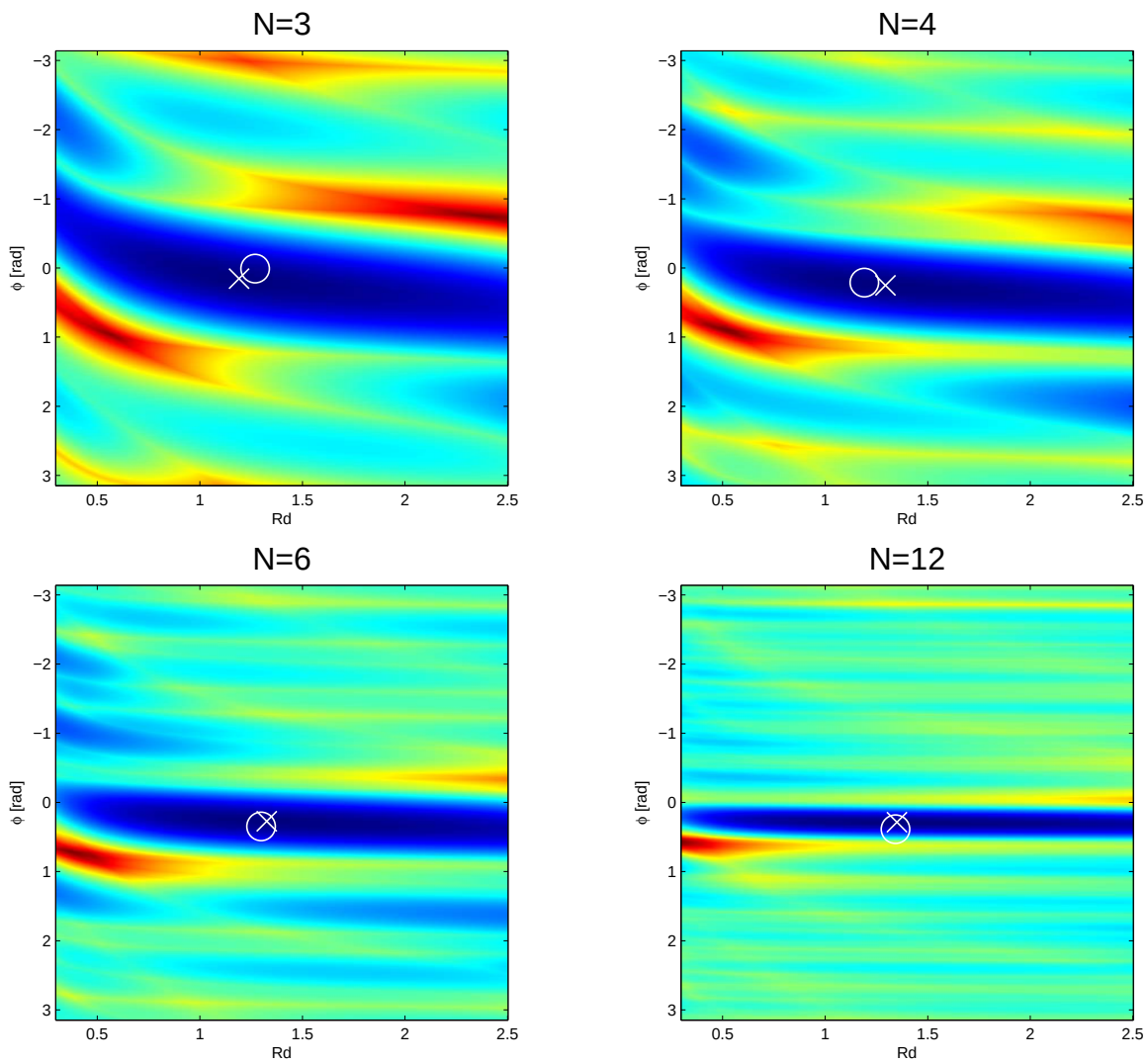


Figure 5.2: MSP(Rd, ϕ, N) surface while increasing N. The colder the color, the smaller the mean squared phase. Starting values of each step are indicated with a circle and the final steps of Preconditioned Conjugate Gradient are shown with a cross.

5.2 Difference operator for phase distortion measure

Instead of the phase spectrum, it has been shown that the group-delay (or the phase-slope) can be used to estimate GCIs [NKGB07, SY95]. Similarly, in this section, the phase derivative is used to estimate the shape parameter of the glottal model. Since only harmonics are used in equation (5.4), the difference operator with respect to harmonic phase is used to approximate the frequency derivative of the phase:

$$\Delta\angle X_h = \angle X_{h+1} - \angle X_h$$

Therefore, the corresponding error function to minimize is the Mean Squared Phase Difference (MSPD):

$$\mathrm{MSPD}(\theta, \phi, N) = \frac{1}{N}\sum_{h=1}^{N}\left(\Delta\angle R_h^{(\theta,\phi)}\right)^2 \tag{5.11}$$

Consequently, applying the difference operator to equation (5.8) leads to:

$$\Delta\angle R_h^{(\theta,\phi)} = (\phi^* - \phi) + \Delta\angle\left(\frac{G_h^{\theta^*}\cdot\mathcal{E}_-(G_h^{\theta})}{G_h^{\theta}\cdot\mathcal{E}_-(G_h^{\theta^*})}\right) \tag{5.12}$$

Compared to (5.8), one can see that the linear-phase error is no longer weighted by the harmonic number h. Moreover, this conditioning is also promising in order to estimate the shape parameter because it represents linearly the time shifting of a given frequency. Using the LFRd model, figure 5.3 shows an example of MSPD$(Rd, \phi, 12)$. Although the influence of Rd and ϕ seems better balanced compared to figure 5.1, the two parameters are actually highly dependent. Indeed, the position error in equation (5.12) can fit the average value of the phase distortion of the shape error. Without the difference operator, the harmonic number h weights the MSP error function and constrains ϕ to its ideal value. In the example of figure 5.1, one can see that the optimal ϕ value is not affected by Rd, a straight horizontal trench is visible at $\phi \approx 0.3$.

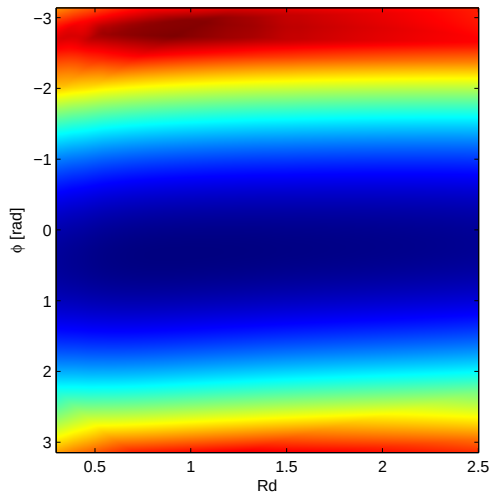


Figure 5.3: Example of MSPD$(Rd, \phi, 12)$ computed on a synthetic signal. Compared to fig. 5.1, the influence of each parameters Rd and ϕ is better balanced.

5.2.1 The method using MSPD

First, to avoid any problem with the phase wrapping in a limited range (e.g. $[-\pi;\pi[$), the phase difference operation of equation (5.11) can be computed in the complex plane:

$$\Delta\angle X_h = \angle\left(\frac{X_{h+1}}{X_h}\right)$$

Then, conversely to MSP, the function MSPD(Rd, ϕ, N) has always only one minimum from our experiments with synthetic signals (although more minima can exist with real signals since the glottal model does not always correspond to the real glottal pulse). Therefore, Algorithm 1 is not used to find the global minimum of MSPD(Rd, ϕ, N). Instead, a regular Preconditioned Conjugate Gradient method is used with $N = \lfloor f_{lim}/f_0 \rfloor$.

5.3 Estimation of multiple shape parameters

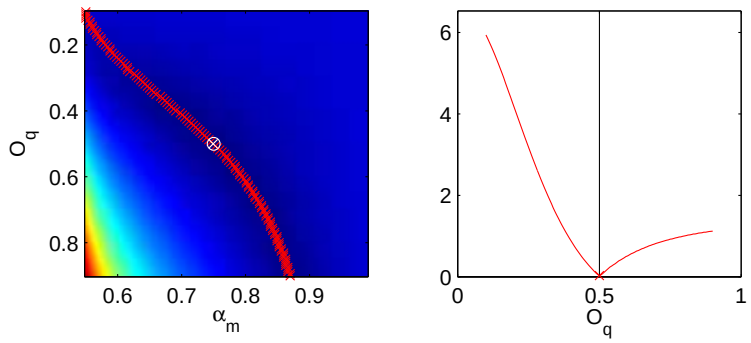
Current glottal models have from 1 to 6 shape parameters (see sec. 2.4). As mentioned in the introduction, although this study focuses on the estimation of only one shape parameter, it is interesting to investigate, at least briefly, the estimation of multiple shape parameters. By increasing the number of shape parameters of a given glottal model, the shape space covered by this model increases. Conversely, the risk to increase the dependency between the shape parameters is more important (i.e. the risk that a parameter moves the model in the same direction as another one). Using dependent parameters, one can compensate another one or, more generally, a set of parameters can compensate another set. Consequently, in terms of estimation of multiple parameters, one can expect offset effects between dependent parameters. Moreover, these effects are dependent on the error function that is used. Indeed, an error function projects the shape space into another new space where parameters can become dependent. In this section the dependency of the full shape parameters set (O_q, α_m, Q_a) of the LF model is investigated.

More technically, to focus on the dependency of the shape parameters, in the following the position of the glottal model is assumed to be known. However, as seen in the previous sections since the estimation of the position is dependent on the shape parameter and vice versa, a more complete study would include ϕ among the studied parameters. Figures 5.4, 5.5 and 5.6 show on the left column error surfaces for a given error function and a given synthetic signal whereas the right column shows the function of minimum error for each line of the corresponding error surface. For each figure, (a,b) are computed without the filtering elements whereas (c,d) show the same error using the filtering elements (the synthetic VTF correspond to an /a/). Additionally, (a,c) show the Root Mean Square (RMS) of the additive error (the difference between the LF model and itself). (b,d) show the square root of the MSP which is equivalent to the RMS of the phase. The optimal parameters are shown with circles and the minimum of the error surfaces are shown using a cross.

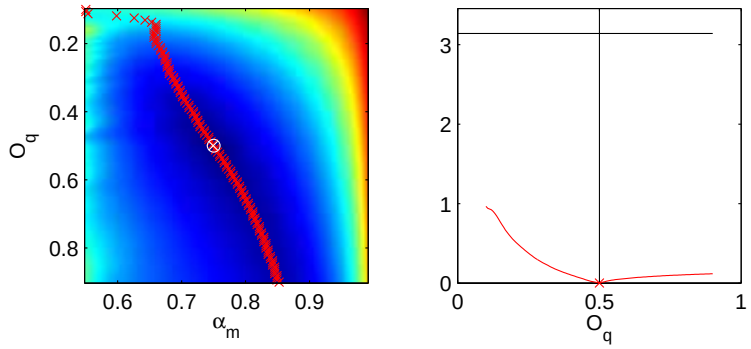
From figures 5.4, 5.5 and 5.6, one can make the following observations:

- All the error surfaces of all figures are never symmetric around the optimal point. Therefore, no pair of parameters is perfectly independent. Note that for the ARX error, such a dependency between O_q and α_m has been already reported [Vin07].

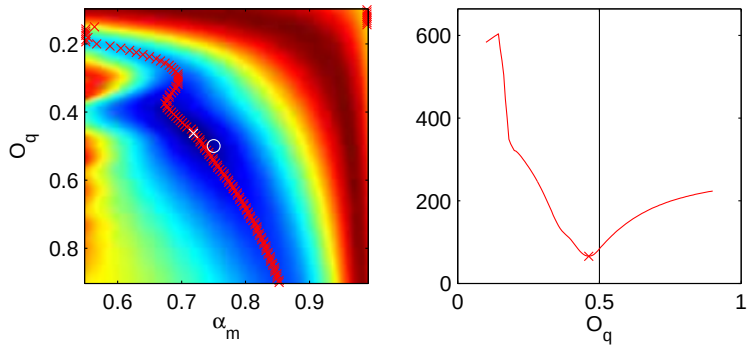
- However, by scaling and rotating the error surfaces of figures 5.5(a) and 5.6(a) (i.e. by using a linear transformation of the error function), it seems possible to obtain an error surface close to a symmetric paraboloid.



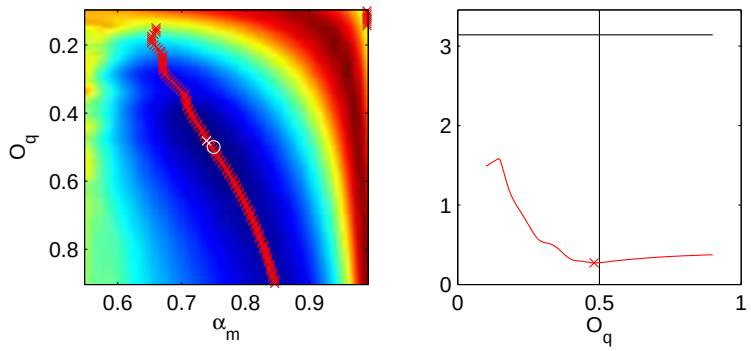
(a) $\mathrm{RMS}_{12}\big(G_h^{(0.5,0.75,0.01)} - G_h^{(O_q,\alpha_m,0.01)}\big)$



(b) $\mathrm{RMS}_{12}\big(\angle(G_h^{(0.5,0.75,0.01)}/G_h^{(O_q,\alpha_m,0.01)})\big)$

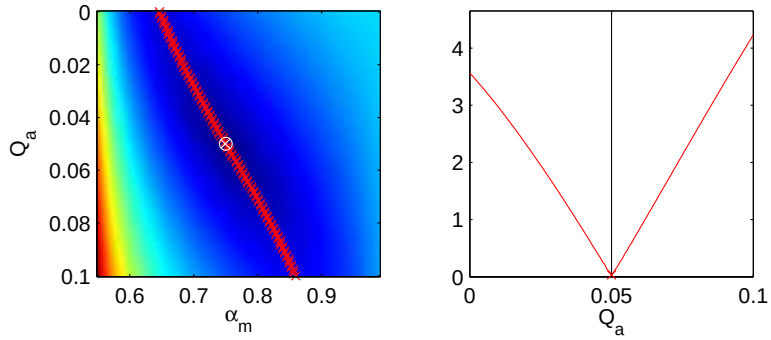


(c) $\mathrm{RMS}_{12}\big(S_h - M_h^{(O_q,\alpha_m,0.01)}\big)$ where $M_h^{(O_q,\alpha_m,0.01)} = e^{jh\phi^*} \cdot G_h^{(O_q,\alpha_m,0.01)} \cdot \tilde{C}_{h-}^{(O_q,\alpha_m,0.01)} \cdot jh$

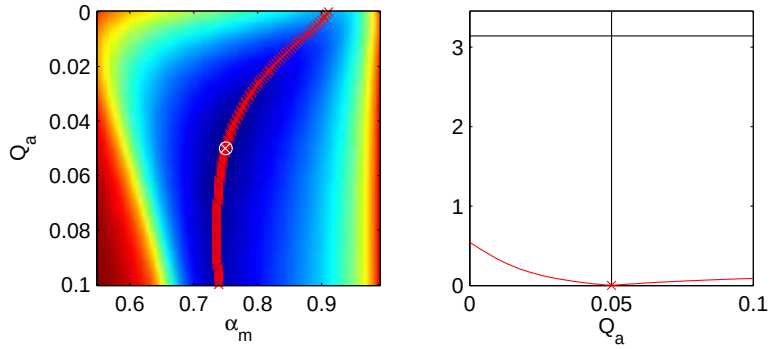


(d) $\mathrm{RMS}_{12}\big(\angle(S_h/M_h^{(O_q,\alpha_m,0.01)})\big)$

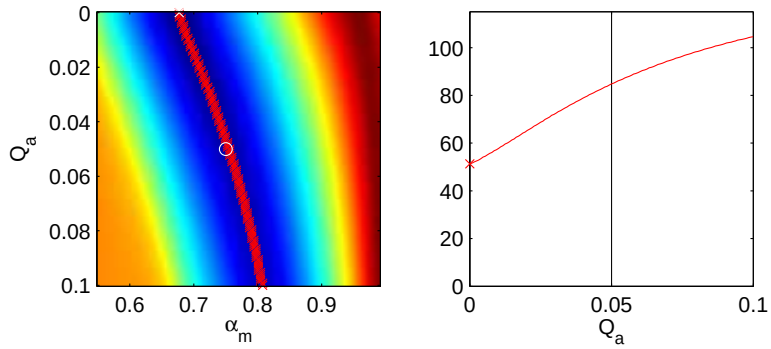
Figure 5.4: RMS of the first 12 harmonics for (α_m, O_q) using the LF model $(O_q^* = 0.5, \alpha_m^* = 0.75, Q_a^* = 0.01)$. The RMS is shown to the left with colors (the colder the color, the smaller the RMS) and to the right by the ordinate. In red line, the right column shows the function of minimum error for each line of the error surface. Horizontal an vertical lines show respectively the optimal value and the maximum error.



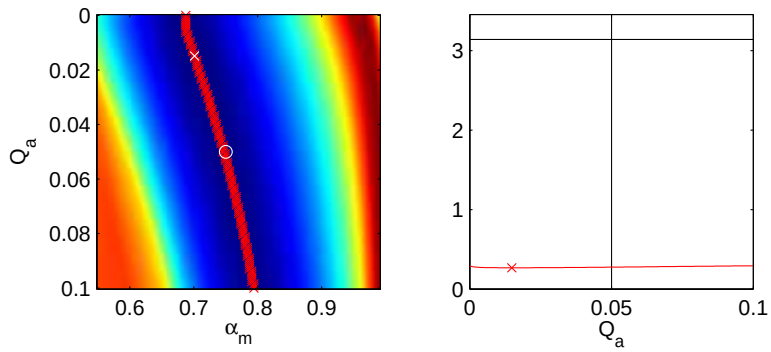
(a) $\mathrm{RMS}_{12}\left(G_h^{(0.5,0.75,0.05)} - G_h^{(0.5,\alpha_m,Q_a)}\right)$



(b) $\mathrm{RMS}_{12}\left(\angle(G_h^{(0.5,0.75,0.05)}/G_h^{(0.5,\alpha_m,Q_a)})\right)$

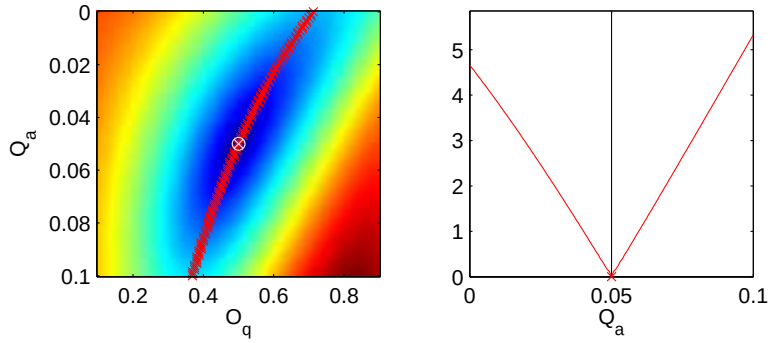


(c) $\mathrm{RMS}_{12}\left(S_h - M_h^{(0.5,\alpha_m,Q_a)}\right)$ where $M_h^{(0.5,\alpha_m,Q_a)} = e^{jh\phi^*} \cdot G_h^{(0.5,\alpha_m,Q_a)} \cdot \tilde{C}_{h-}^{(0.5,\alpha_m,Q_a)} \cdot jh$

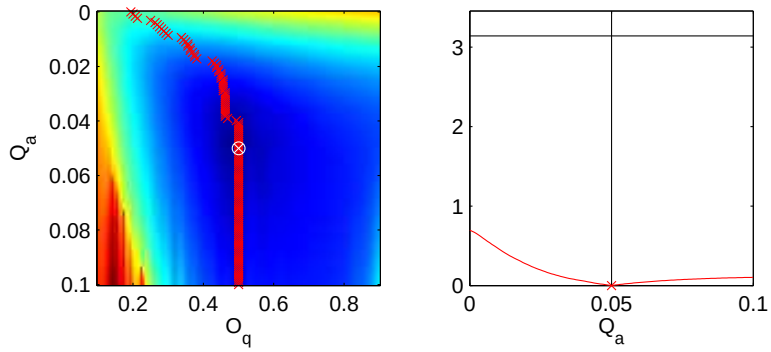


(d) $\mathrm{RMS}_{12}\left(\angle(S_h/M_h^{(0.5,\alpha_m,Q_a)})\right)$

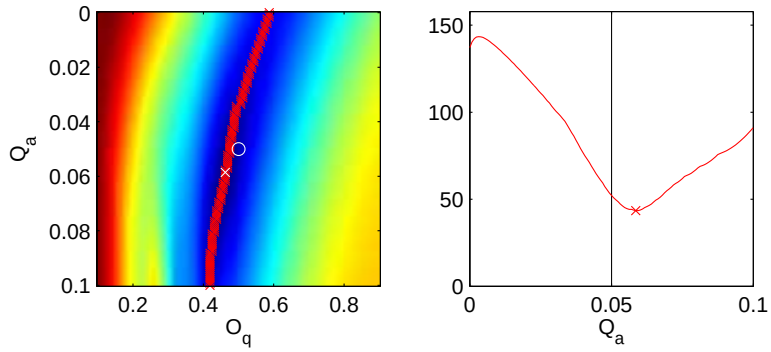
Figure 5.5: RMS of the first 12 harmonics for (α_m, Q_a) using the LF model $(O_q^* = 0.5, \alpha_m^* = 0.75, Q_a^* = 0.05)$



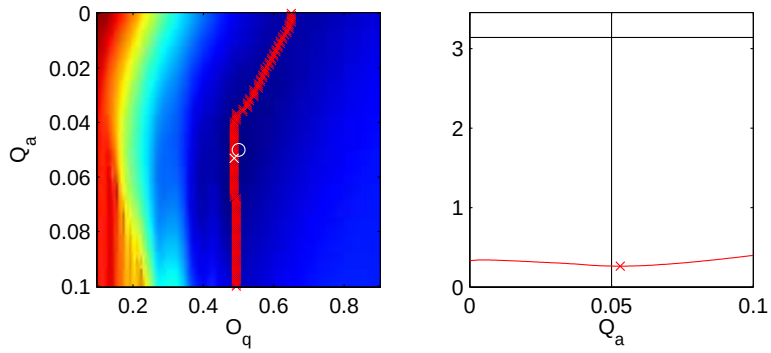
(a) $\mathrm{RMS}_{12}\left(G_h^{(0.5,0.75,0.05)} - G_h^{(O_q,0.75,Q_a)}\right)$



(b) $\mathrm{RMS}_{12}\left(\angle(G_h^{(0.5,0.75,0.05)}/G_h^{(O_q,0.75,Q_a)})\right)$



(c) $\mathrm{RMS}_{12}\left(S_h - M_h^{(O_q,0.75,Q_a)}\right)$ where $M_h^{(O_q,0.75,Q_a)} = e^{jh\phi^*} \cdot G_h^{(O_q,0.75,Q_a)} \cdot \tilde{C}_{h-}^{(O_q,0.75,Q_a)} \cdot jh$



(d) $\mathrm{RMS}_{12}\left(\angle(S_h/M_h^{(O_q,0.75,Q_a)})\right)$

Figure 5.6: RMS of the first 12 harmonics for (O_q, Q_a) using the LF model $(O_q^* = 0.5, \alpha_m^* = 0.75, Q_a^* = 0.05)$

- On left plots, one can see that the functions of minimum error (red crosses) are far from straight lines. This means that a non-linear relation exists between parameters regarding the error function (e.g. figure 5.4(a))
- With the filtering elements (plots (c,d)), the minimum of the error surface does not correspond to the optimal parameters. Whatever the used algorithm to find the parameters of minimum error, the solution is biased.
- In plots 5.6(d) and 5.5(d), one can see that the curve of minimum error is almost constant for any Q_a value (the same is visible for 5.4(d) for O_q values bigger than 0.5). Therefore, in presence of any disturbance (e.g. noise or an unexpected pulse shape), parameter pairs on these curves can have the same error and the optimum of the error surfaces can be not unique. Therefore, according to these figures, the estimation of the return phase of the LF model seems to be difficult. Note that, because the additive error is dependent on the glottal model amplitude, such observations are less obvious to discuss.

Comments on the weighting of the error function

The parameters of a glottal model influence frequency bands with different weighting [Dd99, DdD97]. Therefore, if the weighting of the error function is not uniform among the parameters, the estimation of one parameter (e.g. Q_a) can be flooded into the estimation error of another one (e.g. O_q) [FMT99]. Note that, although the MSP has an uniform weighting in spectral domain compared to the additive error, the size of the frequency bands influenced by each parameter is not uniform. Therefore, a weighting function should be used in order to obtain an equal influence of each parameter on the final error value.

5.3.1 O_q/α_m vs. I_q/A_q

This section illustrates the offset effect which can appear in the estimation of multiple parameters of a glottal model. In the following, the O_q and α_m parameters of the LF model are estimated with a grid search algorithm using the error function $\mathrm{MSP}((O_q, \alpha_m), \phi, 12)$. In this method, the optimal parameter triplet (O_q, α_m, ϕ) is simply the one giving the smallest error among a grid of 128 O_q values and 128 α_m values and 128 delays. Conversely to an iterative method which can stop in a local minimum, this method can find the global minimum of the error hyper-surface. On the other hand, the computation cost of such a method is obviously unrealistic in real applications.

In figure 5.7, one can see that the O_q value and the α_m values are particularly irregular around $t = 2.46$. Nevertheless, using the LF definition and these estimated parameters, one can compute $I_q = (t_e - t_i)/T_0 = O_q(1 - \alpha_m)$ (see figure 2.4). It is the duration between the extrema of the glottal pulse derivative. In addition, one can compute the ratio between the corresponding amplitudes of these extrema $A_q = \dot{g}(t_e)/\dot{g}(t_i)$ (which is independent of the excitation amplitude of the glottal pulse E_e). Finally, from these two new measures shown in figure 5.7, one can see that O_q and α_m offset each other in order to keep a relatively smooth time evolution of the I_q measure whereas A_q represents the underlying irregularities which create the uncertainty on the solutions of (O_q, α_m).

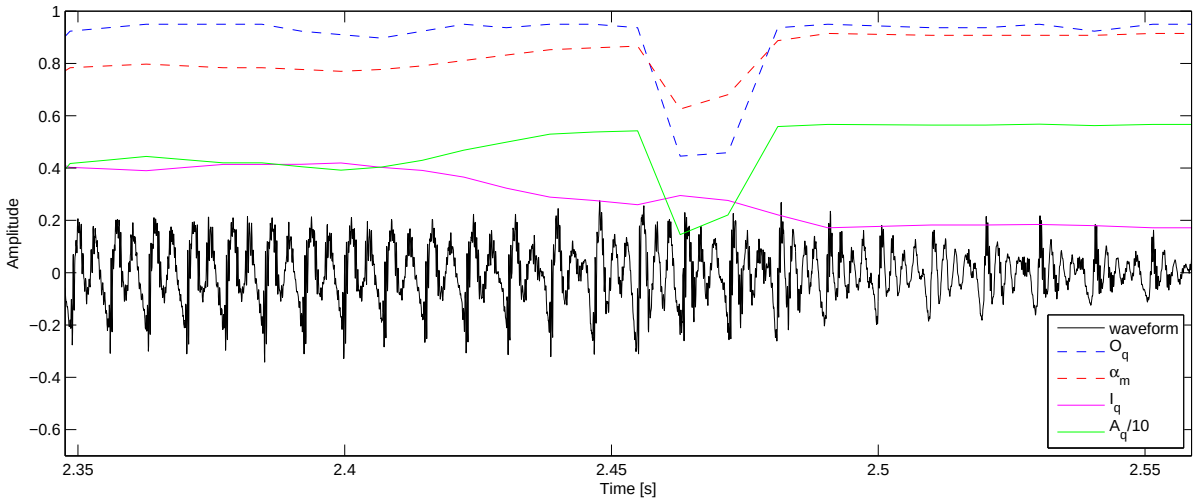


Figure 5.7: The parameters (I_q, A_q) from the estimation of (O_q, α_m) using MSP on a real speech segment. The waveform in plain black line and the parameters in color according to the legend. One can see that O_q and α_m offset each other in order to keep a relatively smooth time evolution of I_q whereas A_q represents the underlying irregularities.

Conclusions

- According to the minimum-phase and mixed-phase properties of the VTF and the glottal pulse respectively, the shape parameters of a glottal model can be estimated using the phase minimization criteria already proposed for GCI estimation [NKGB07, SY95].
- The proposed error function is the minimization of the Mean Squared Phase (MSP) of the convolutive residual (the spectral division of the observed spectrum by its model). Note that this function is independent on the excitation amplitude of the glottal model.
- Moreover, using the Root Mean Square (RMS) of the residual phase spectrum, one can create a measure of confidence (eq. 5.9). The quality of the proposed estimation methods can thus by quantitatively evaluated (again, this evaluation is independent on the amplitude excitation of the glottal model).
- Using MSP, it has been shown that the the following conditions have to be satisfied to attain convergence of the estimated shape parameters towards the optimal ones:
  1. A shape parameter θ influences the MSP if θ influences at least one zero outside of the unit circle in the glottal model.
  2. To avoid that a shape parameter θ offsets the estimated pulse position, the zero time reference of the glottal model has to be chosen such that θ does not influence the linear-phase component of the glottal model (e.g. t_e for the LF model).
- A first method minimizing the MSP has been proposed to jointly estimate the shape parameter and the time position of a glottal model. An iterative process is used to minimize the risk that the search method stop in a local minimum of the error surface (sec. 5.1.4).
- According to the frequency derivative used in GCI estimation methods, a second method has been proposed to better balance the influence of the position error compared to the influence of the shape error.
- The estimation of multiple parameters has been briefly discussed. Section 5.3 shows that strong dependencies exist between the LF parameters. Additionally, figure 5.7 shows that the O_q and α_m parameters can offset each other in order to keep a smooth time evolution of the duration between the extrema of the glottal pulse derivative I_q.

Chapter 6

Estimation of the shape parameter without pulse position

This chapter shows that the shape of a glottal model can be estimated independently of its time position. A first method is proposed which takes advantage of the phase difference operator. Finally, a quasi closed-form expression of the shape parameter is developed from the observed spectrum.

6.1 Parameter estimation using the 2nd order phase difference

As seen in section 5.2, the 1st order phase difference can be used to remove the weighting of the harmonic number on the error function. Consequently, using the the 2nd order phase difference (Δ^2), one can see that the position parameter ϕ can be completely removed from the convolutive residual:

$$\Delta^2 \angle R_h^\theta = \Delta^2 \angle \left(\frac{G_h^{\theta^*} \cdot \mathcal{E}_-(G_h^\theta)}{G_h^\theta \cdot \mathcal{E}_-(G_h^{\theta^*})} \right) \tag{6.1}$$

However, to recover the representation of the first order frequency derivative which emphasis the phase distortion by the shape error, the anti-difference operator Δ^{-1} is used which is the discrete version of the integral. Accordingly, to estimate the shape parameter θ, the corresponding error function to minimize is:

$$\mathrm{MSPD}^2(\theta, N) = \frac{1}{N} \sum_{h=1}^{N} \left(\Delta^{-1} \Delta^2 \angle R_h^\theta \right)^2 \tag{6.2}$$

where R_h^θ is computed using equation (5.4) without the linear-phase term. Figure 6.1 shows an examples of $\mathrm{MSPD}^2(Rd, 12)$.

6.1.1 The method based on MSPD^2

In order to obtain an estimate of the shape parameter, equation (6.2) is minimized with respect to θ. Therefore, a proper algorithm has to be chosen regarding to the shape of this error function. Figure 6.1 shows MSPD^2 error functions computed on a voiced signal synthesized with the phoneme /e/ and a fundamental frequency of 128 Hz for three different Rd parameters. Figure 6.2 shows the same for

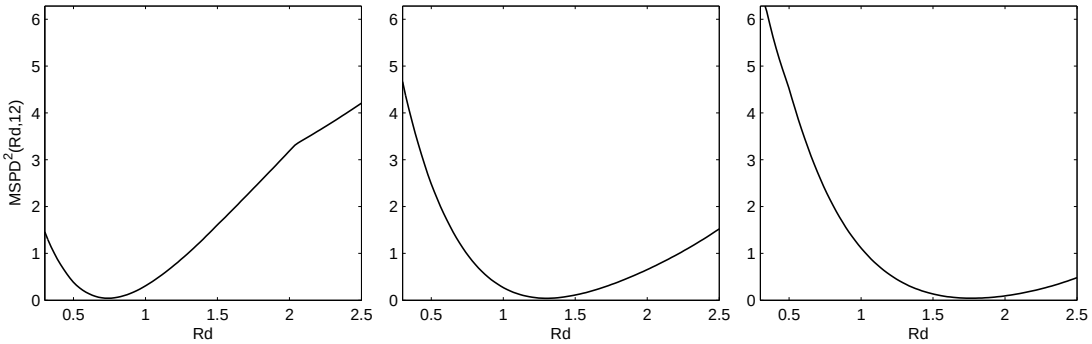


Figure 6.1: Examples of $\mathrm{MSPD}^2(Rd, 12)$ error functions computed on a voiced signal synthesized with the phoneme /e/ and $Rd^* = 0.6$, 1.4 and 2.2.

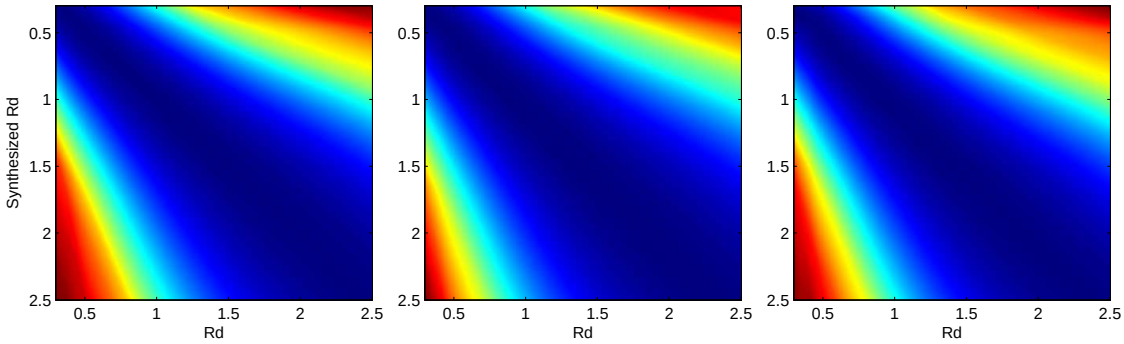


Figure 6.2: error function $\mathrm{MSPD}^2(Rd, 12)$ with respect to the Rd value used in the synthesized voiced signal for the phonemes /a/, /i/ and /u/.

Rd synthesis values in $[0.3; 2.5]$ and the tree phonemes defining the vocalic triangle /a/, /i/ and /u/. Consequently, since the MSPD^2 error function seems to have always only minimum, a simple Brent's algorithm [Bre73] is used to find the global minimum of this function. Note that, conversely to the preconditioned conjugate gradient algorithm, no initial value is necessary for this optimization method.

Finally, like for the MSPD based method, to avoid any problem with the wrapping of the phase in a limited range, the 2nd order phase difference centered on each k-harmonic is first computed in the complex plane:

$$\Delta^2 \angle X_h = \angle \frac{X_{h+1} \cdot X_{h-1}}{X_h^2}$$

Then, applying the anti-difference operation, the previous equation leads to:

$$\Delta^{-1}\Delta^2 \angle X_h = \angle \prod_{n=1}^{k} \frac{X_{n+1} \cdot X_{n-1}}{X_n^2} \tag{6.3}$$

which is used to compute the MSPD^2 error function defined by equation (6.2).

6.2 Parameter estimation using function of phase-distortion

This section presents a last method to estimate the shape parameter of a glottal model. Below, the Function of Phase-Distortion (FPD) is first presented. Then, section 6.2.1 shows that the FPD of a glottal model can be used to estimate a unique shape parameter. Finally, section 6.2.2 will show that the properties of the FPD of a glottal model can be used to evaluate *a priori* to which extent a shape parameter can be estimated using the methods based on mean squared phase.

For a given spectrum, the Function of Phase-Distortion (FPD) expresses the components of the phase spectrum which are neither related to the linear-phase nor to the minimum-phase component of the spectrum. For any harmonic spectrum X_h, the FPD is thus formalized as follows:

$$\Phi_h(X) = \Delta^{-1}\Delta^2\angle\left(\frac{X_h}{X_{h-}}\right) \tag{6.4}$$

where the difference operators are computed using equation (6.3). Consequently, the FPD is a generalization of the phase distortion measure of the convolutive residual which is used in the method based on MSPD^2. As an example, the left plot of figure 6.3 shows the first three harmonics of the FPD of the LF model with respect to the *Rd* shape parameter.

6.2.1 The method FPD^{-1} based on FPD inversion

Below, a means to estimate the shape parameter of a glottal model is described which expresses the shape parameter in a quasi closed-form of an observed spectrum.

First, the voice production model is assumed to perfectly represent the observed spectrum (conversely to previous methods, no residual term is taken into account):

$$S_h = M_h^{(\theta,\phi)} = e^{jh\phi}\cdot G_h^\theta\cdot\mathcal{E}_-\left(\frac{S_h}{G_h^\theta\cdot jh}\right)\cdot jh$$

Then, $\mathcal{E}(.)$ is assumed to be multiplicative. Thus, it can be distributed to the elements of its argument:

$$S_h = e^{jh\phi}\cdot G_h^\theta\cdot\frac{\mathcal{E}_-(S_h)}{\mathcal{E}_-(G_h^\theta\cdot jh)}\cdot jh \tag{6.5}$$

and therefore, one can put the observed data and the models on each side of the equality:

$$\frac{S_h}{\mathcal{E}_-(S_h)} = e^{jh\phi}\cdot\frac{G_h^\theta\cdot jh}{\mathcal{E}_-(G_h^\theta\cdot jh)}$$

For the sake of simplicity, one can write $X_{h-} = \mathcal{E}_-(X_h)$ and merge the radiation into the glottal model ($G_h'^\theta = G_h^\theta\cdot jh$ (see sec. 3.3)):

$$\frac{S_h}{S_{h-}} = e^{jh\phi}\cdot\frac{G_h'^\theta}{G_{h-}'^\theta} \tag{6.6}$$

In terms of phase-distortion, if both sides of equation (6.6) are equal, their respective FPDs are also equal:

$$\Phi_h(S) = \Phi_h(G'^\theta) \tag{6.7}$$

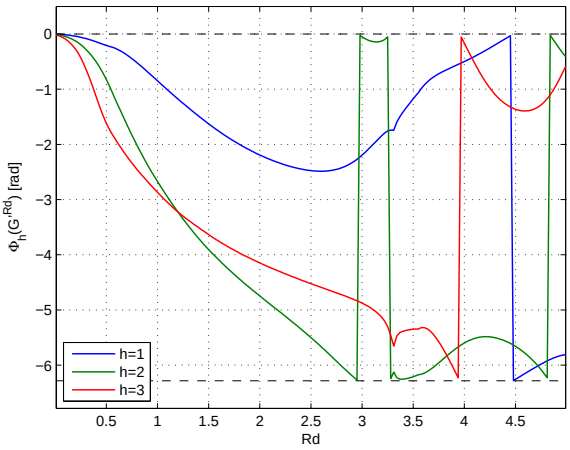


Figure 6.3: The first three harmonics of $\Phi_h(G'^{Rd})$ with respect to Rd.

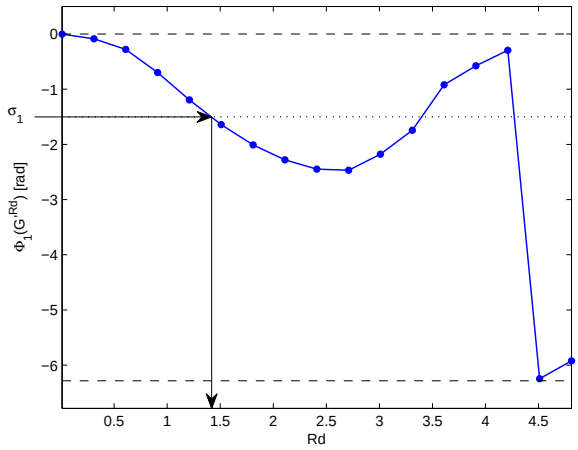


Figure 6.4: Inversion of $\Phi_1(G'^{Rd})$ using a lookup table.

Using equation (6.4), the FPD of the observed spectrum can be measured in equation (6.7). Therefore, to estimate a unique shape parameter of a glottal model, it is sufficient to inverse $\Phi_h(G'^{\theta})$ with respect to the shape parameter for a given harmonic h.

$$\text{given the observation } \sigma_h = \Phi_h(S) \qquad \text{find } \theta \ : \Phi_h(G'^{\theta}) = \sigma_h \tag{6.8}$$

However, this inversion is far from straightforward for the existing glottal models. For example, the shape of the LFRd model is defined using synthesis parameters which are not closed-form expressions of the shape parameters (see sec. 2.4). However, the analytic inversion can be approximated numerically. Indeed, $\Phi_h(G'^{\theta})$ can be sampled for each harmonic to create a lookup table whose elements are used to predict θ from the observed σ_h values (see figure 6.4 for the LFRd model).

According to figure 6.4, an observed σ_h value can cross $\Phi_h(G'^{\theta})$ at multiple abscissa. Therefore, using only one harmonic, the shape parameter can be estimated only in an interval where $\Phi_h(G'^{\theta})$ is monotonic (e.g. $[0; 2.7]$ for $\Phi_1(G'^{Rd})$). Additionally, the interval of the observed value σ_h is limited due to the wrapping of the angle function. Therefore, a proper interval has to be defined where the unwrapped functions of the lookup table have the less number of discontinuities. According to figures 6.5, 6.6 and 6.7, the interval $[0; -2\pi]$ is used in the following. Finally, once a θ value is predicted from each observed σ_h value from each harmonic, an average value over different harmonics can be retrieved. Here, the mean value of the predicted θ values is used. Algorithm 2 summarizes the whole method.

Algorithm 2 The method FPD^{-1}

Compute the lookup table $\gamma_h(\theta) = \Phi_h(G'^{\theta})$
for each analysis time in the speech recording **do**
  Build a harmonic model S_h on a window of ≈ 4 periods
  Compute $\sigma_h = \Phi_h(S)$ according to equation (6.4) using equation (6.3)
  Predict θ_h from σ_h for each harmonic h using the lookup table $\gamma_h(\theta)$
  Obtain an average $\bar{\theta}$ value from the mean value of the θ_h values
end for

Compared to the methods minimizing a mean squared phase (MSP, MSPD and MSPD^2), a few differences exist. Firstly, it is not clear how this new approach could be used to estimate multiple parameters. Secondly, whereas a harmonic representation of the VTF is explicitly computed in the other methods (through equations (5.4) and (5.3)), this new approach separate the terms of the VTF (see eq. 6.5). As a consequence, although the $\mathcal{E}_-(.)$ function is assumed to be multiplicative in theory, a few differences arise in practice (e.g. the DC value in $\mathcal{E}_-(.)$ is extrapolated on different terms depending on the method). In conclusion, even for the estimation of a single shape parameter, one can expect different results between methods based on mean squared phase minimization and the method FPD^{-1}.

6.2.2 Conditioning of the FPD inversion

Given a glottal model and its parametrization, it is interesting to evaluate *a priori* the reliability of the method FPD^{-1}. In this section, the conditioning of the inversion of the function $\Phi_h(G'^{\theta})$ is evaluated.

One can evaluate to which extent an error of the observed σ_h value can influence the estimation of the shape parameter θ. Indeed, according to figure 6.4, the steeper the slope of the $\Phi_h(G'^{\theta})$ functions, the more robust will the estimate of θ be. The condition number κ, which is mostly used for matrices, can be generalized to functions. Therefore, the following expression is studied:

$$\kappa\left(\Phi_h(G'^{\theta})\right) = \left|\frac{d\tilde{\Phi}_h(G'^{\theta})}{d\theta}\right|^{-1} \tag{6.9}$$

where $\tilde{\Phi}_h(G'^{\theta})$ returns the unwrapped phase of the function $\Phi_h(G'^{\theta})$ with respect to θ. Since the analytical derivative of the previous equation is far from straightforward, a discrete approximation is used:

$$\kappa\left(\Phi_h(G'^{\theta})\right) \approx \left|\frac{\Delta\theta}{\tilde{\Phi}_h(G'^{\theta+\Delta\theta}) - \tilde{\Phi}_h(G'^{\theta})}\right| \quad \text{with} \quad |\Delta\theta| < \epsilon$$

and ϵ is an arbitrary small positive value. Finally, the inversion of $\Phi_h(G'^{\theta})$ is well-conditioned if κ is low. Thus, the smaller the κ, the more reliable the method FPD^{-1}.

Figures 6.5,6.6,6.7 show, for different harmonics, the Φ and κ functions for the glottal models: LFRd, CALMRd and Rosenberg^{O_q}. Additionally, the mean value of the first N harmonics of $\kappa\left(\Phi_h(G'^{\theta})\right)$ is shown. This value should be relevant according to the mean value of the shape parameters estimated from each harmonics (see end of Algorithm 2). According to these figures, one can make the following remarks. Firstly, for each glottal model and its parametrization, a confidence interval of the shape parameter can be defined where the κ value is significantly low. In this interval, one can ensure that the shape parameter can be estimated (approximately $[0; 2.5]$ for the LFRd and CALMRd models and $[0.5; 1]$ for the Rosenberg^{O_q} model). Secondly, a link exists between the FPD of a glottal model and the error surfaces of the methods based on mean squared phase. For example, the error function MSPD^2 is nothing more than the mean squared value of the FPD of the convolutive residual. Therefore, if two different abscissa of $\Phi_h(G'^{\theta})$ for a given harmonic h have the same ordinate (see figure 6.4, $\Phi_h(G'^{1.4}) = \Phi_h(G'^{3.3}) = -1.5$), one can expect two local minima in the MSPD^2 error function. However, since this error function does not use a unique harmonic, one can expect an averaging effect similar to the mean value used in Algorithm 2.

In conclusion, one can enumerate the following properties which have to be satisfied by a glottal model and its parametrization to ensure that its parameter can be properly estimated by the proposed method FPD^{-1}.

- $\Phi_h(G'^{\theta})$ has to be injective with respect to θ. Note that, if two values of the shape parameter imply equivalent phase-distortion, the parameter defines the same position of a given harmonic h for two different shape parameter values.
- $\Phi_h(G'^{\theta})$ has to be continuous. First, this condition is necessary to interpolate the elements of the lookup table. Then, a discontinuity would imply a control of the glottal pulse shape which is not continuous.
- According to the condition number defined by equation (6.9), the smaller the κ number of each harmonic, the more reliable the estimate of the shape parameter.

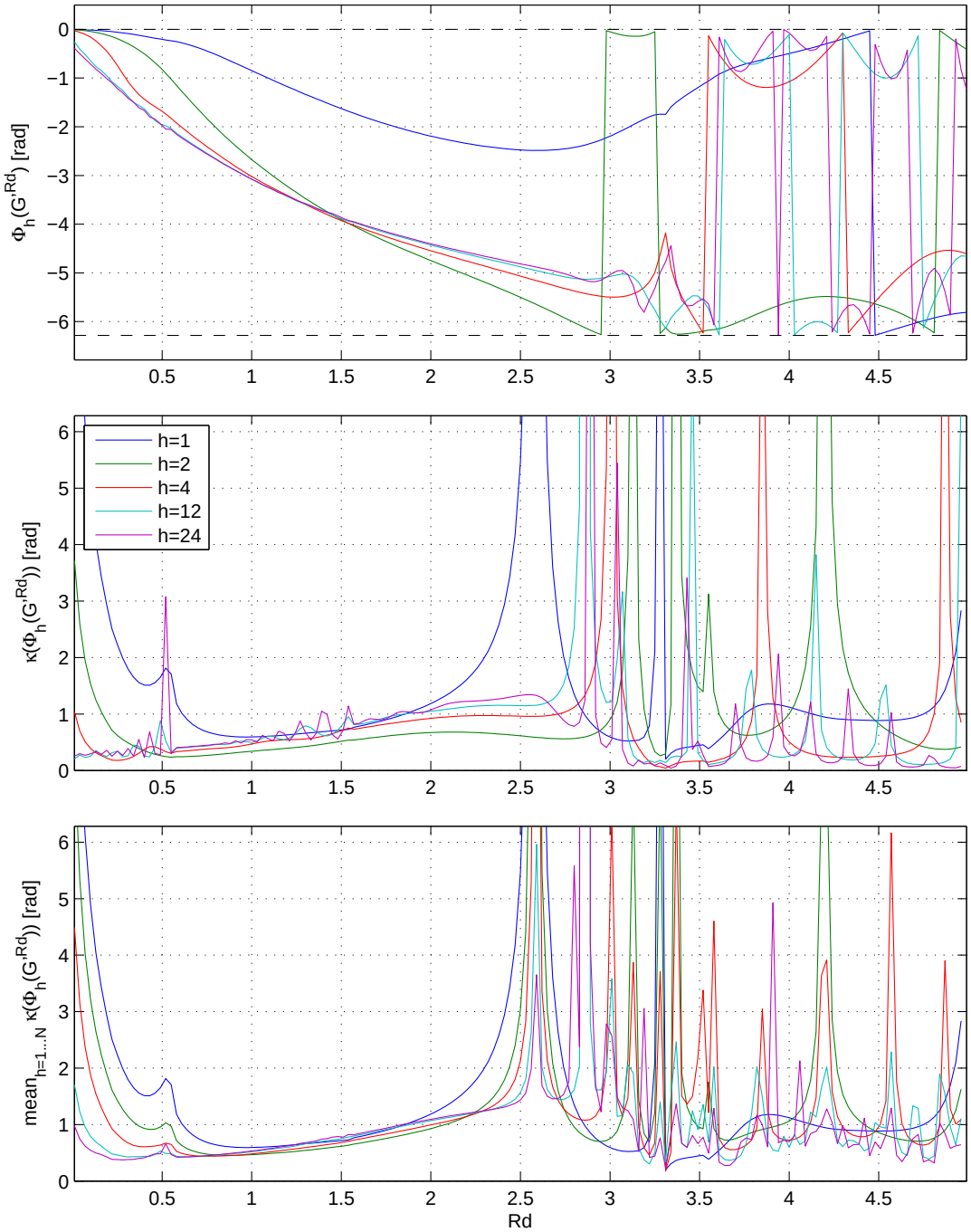


Figure 6.5: For the LFRd glottal model: its Functions of Phase Distortion (FPD) Φ_h to the top. The conditioning measure κ in the middle and the mean conditioning measure related to the method FPD^{-1} to the bottom. From the conditioning measures, one can see that the shape parameter can be estimated only in the interval $[0; 2.5]$.

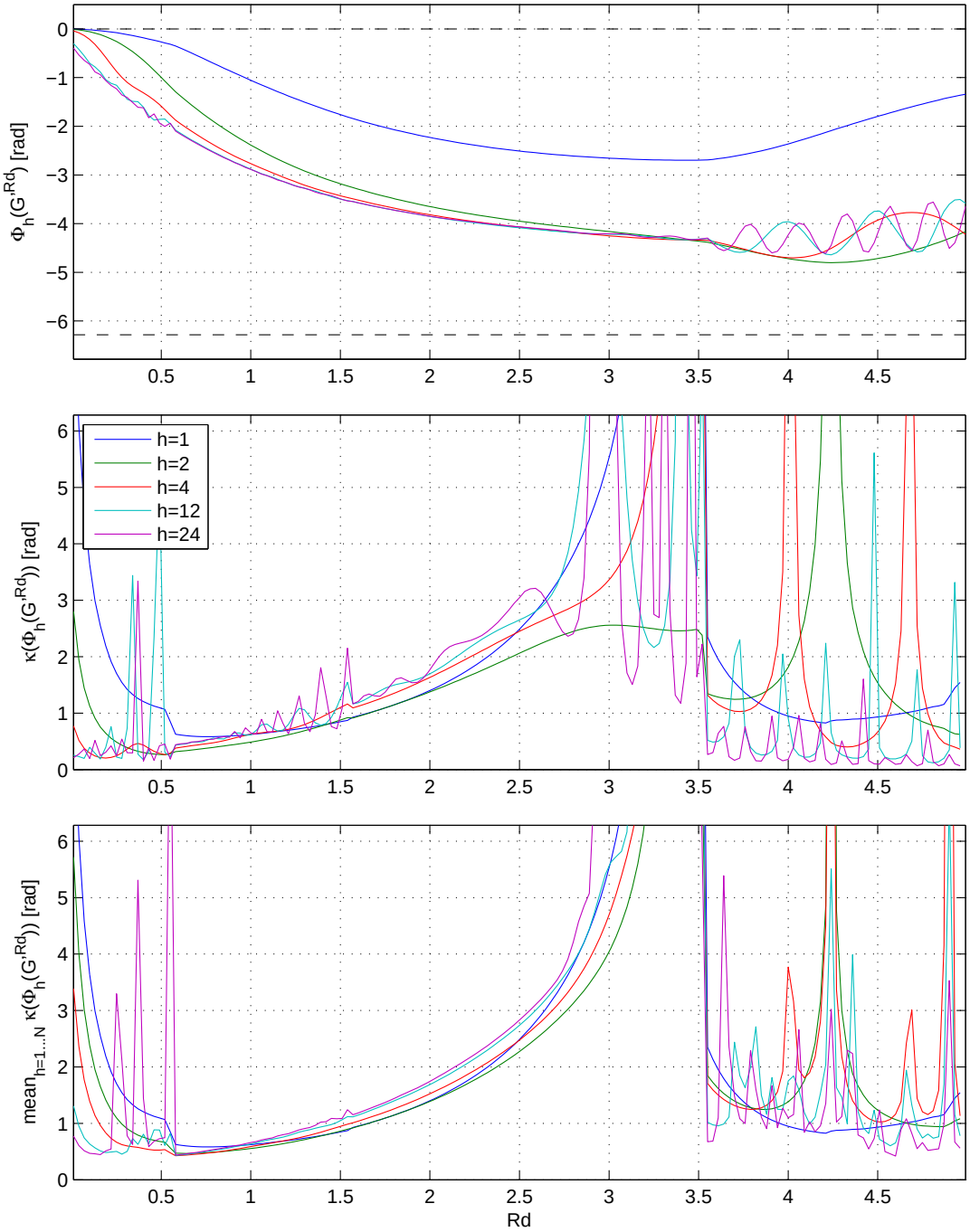


Figure 6.6: FPD Φ_h, conditioning measure κ and mean conditioning measure for the CALMRd glottal model.

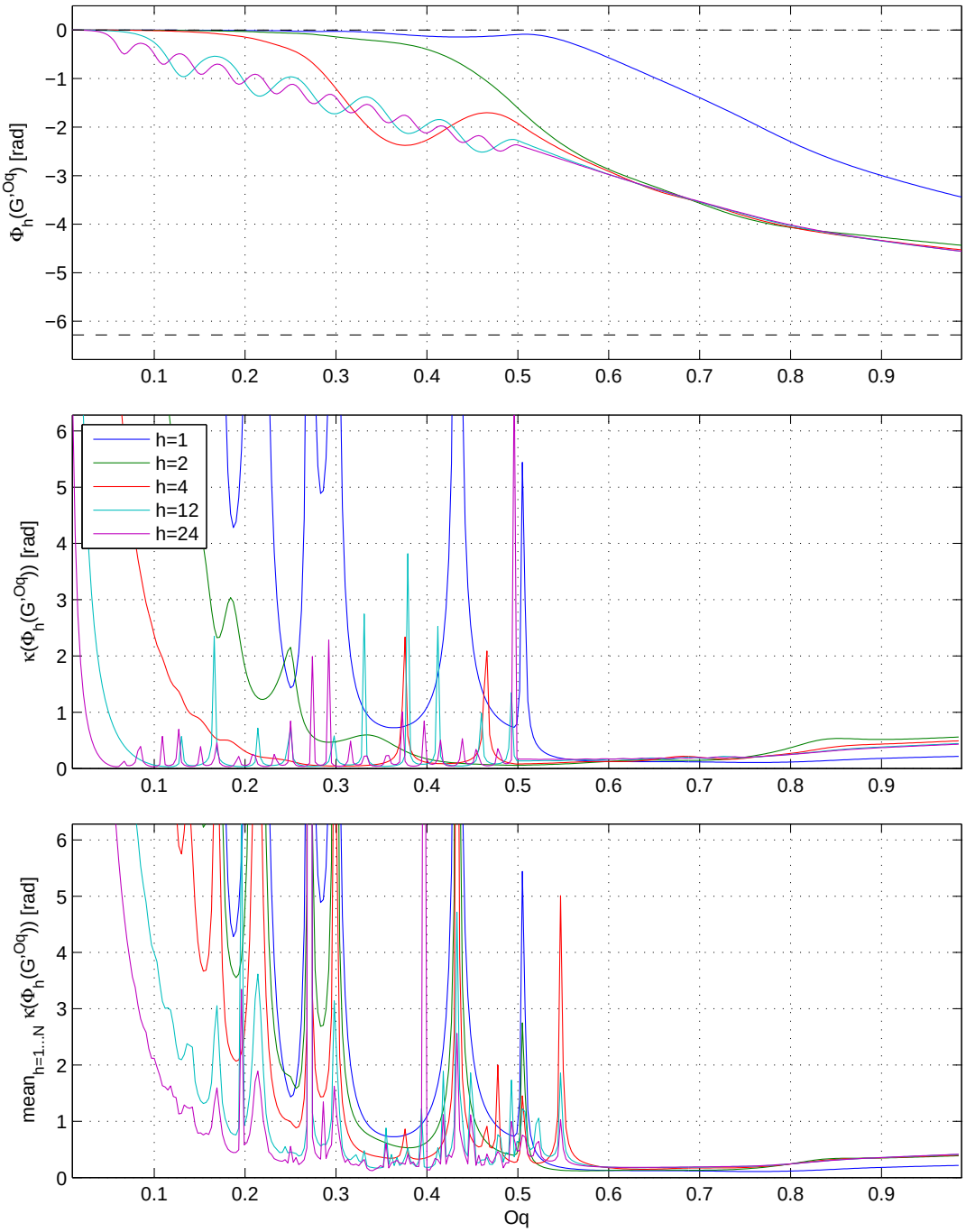


Figure 6.7: FPD Φ_h, conditioning measure κ and mean conditioning measure for the Rosenberg^{O_q} glottal model.

Conclusions

- Using the 2nd order phase difference and the Mean Squared Phase (the error function MSPD^2 in equation (6.2)), it has been shown that the estimation of the shape parameters of a glottal model can be independent of the position and the amplitude of the pulse.
- Accordingly, in order to estimate the Rd parameter of the LF^{Rd} glottal model, it is sufficient to minimize the MSPD^2 using a simple Brent's method.
- Based on the Function of Phase-Distortion (FPD) (eq. 6.4), a second method termed FPD^{-1} has been proposed in this chapter to estimate a unique shape parameter of a glottal model. Note that, if the FPD of a glottal model is analytically invertible, it would be possible to obtain a closed-form expression of the shape parameter from the observed spectrum. Since this inversion is not possible with the LF^{Rd} model, a lookup table has been used in the proposed algorithm.
- For a given glottal model and its parametrization, the study of the FPD also allows to evaluate *a priori* the reliability of the methods minimizing a mean squared phase and the method FPD^{-1}.
- The FPD of a glottal model $\Phi_h(G'^{\theta})$ has the following properties:
  - It has no linear-phase component (ie. insensitive to the glottal pulse position).
  - It is insensitive to the amplitude of the glottal model.
  - Since it is defined on harmonic frequencies, it is normalized with respect to the pulse duration.
  - From the three points above, the FPD is only related to the shape of the glottal pulse.
- Finally, the following two elements have to be taken into account in order to maximize the estimation reliability of a shape parameter:
  - For each harmonic h, $\Phi_h(G'^{\theta})$ should be injective and continuous with respect to the shape parameter θ (and thus monotonic).
  - The reliability is maximal if the condition number define in equation (6.9) is minimal.

Chapter 7

Estimation of the Glottal Closure Instant

In this chapter, a method is proposed to find the instant of maximum excitation energy of the vocal-tract filter within a speech signal period. This instant is usually termed Glottal Closure Instant (GCI) (see sec. 2.5.1). This problem has received many attention since decades [SdR09, Lob01, PQR99, SY95, CO89, AY79, AH71]. The usual approach is to model the glottal source with an impulse train where each impulse is placed at the maximum of VTF excitation. Accordingly, the following criteria have been proposed to estimates the GCIs: maximum of the linear prediction error [AH71], the normalized prediction error [WMG79], the Frobenius norm [MKW94], the Log determinant of the covariance matrix [Str74], the instants of formant modulation [PQR99], the minimum of phase or group-delay [NKGB07, SY95], using wavelet transform [Lob01] and Lines Of Maximum Amplitudes (LOMA) [SdR09], error of an ARX model [VRC06] and finally using a nonlinear voice production model [SL07].

Zero-phase or minimum-phase pulse

In a first approximation, one can assume that the source is sufficiently concentrated in time to assume it is a symmetric signal and thus also a zero-phase signal. Therefore, over one period, a Dirac delta function can be used to represent the radiated glottal pulse. The model of the source is thus reduced to a linear-phase term:

$$S(\omega) = e^{j\omega\phi} \cdot G(\omega) \cdot C_-(\omega) \cdot L(\omega) \qquad e^{j\omega\phi} \cdot G(\omega) \cdot L(\omega) = e^{j\omega\phi} \quad \Rightarrow \quad S(\omega) = e^{j\omega\phi} \cdot C_-(\omega) \tag{7.1}$$

Therefore, according to the last term of this equation, an estimation of a minimum-phase envelope $\mathcal{E}_-(.)$ of the signal can be used to retrieve the linear-phase term. First, the spectral envelope is computed:

$$\tilde{C}(\omega) = \mathcal{E}_-\big(S(\omega)\big) = \mathcal{E}_-\big(e^{j\omega\phi} \cdot C_-(\omega)\big) = \mathcal{E}_-\big(C_-(\omega)\big) \approx C_-(\omega) \tag{7.2}$$

Then, the estimated glottal source is:

$$\tilde{G}'(\omega) = \frac{S(\omega)}{\tilde{C}_-(\omega)} = \frac{e^{j\omega\phi} \cdot C_-(\omega)}{\tilde{C}_-(\omega)} \approx e^{j\omega\phi} \tag{7.3}$$

Therefore, to estimate ϕ, one can locate the maximum of energy of the radiated glottal source or minimize its phase-slope (or the group-delay). Note that the same can be concluded for a minimum-phase pulse. Indeed, the minimum-phase component can be modeled by the envelope $\mathcal{E}_-(.)$. Then, this component is removed in equation (7.3) and only the linear-phase term remains.

Maximum-phase or mixed-phase pulse

Using a more complete glottal model, one can assume that the glottal pulse is a maximum-phase signal. Similar to the last remark above, the same conclusion holds for a mixed-phase pulse compared to a maximum-phase pulse. Therefore, given the voice production model:

$$S(\omega) = e^{j\omega\phi} \cdot G(\omega) \cdot C_-(\omega) \cdot L(\omega) \qquad e^{j\omega\phi} \cdot G_+(\omega) \cdot L(\omega) = e^{j\omega\phi} \cdot G'_+(\omega) \quad \Rightarrow \quad S(\omega) = e^{j\omega\phi} \cdot G'_+(\omega) \cdot C_-(\omega) \tag{7.4}$$

One can compute the VTF estimate $\tilde{C}(\omega)$:

$$\tilde{C}(\omega) = \mathcal{E}_-\big(S(\omega)\big) = \mathcal{E}_-\left(e^{j\omega\phi} \cdot G'_+(\omega) \cdot C_-(\omega)\right) = \mathcal{E}_-\big(G'_+(\omega) \cdot C_-(\omega)\big) \tag{7.5}$$

Then, one can assume that the envelope computation is multiplicative and therefore,

$$\tilde{C}_-(\omega) \approx G'_-(\omega) \cdot C_-(\omega)$$

And, as in equation (7.3), the estimated radiated glottal pulse is:

$$\tilde{G}'(\omega) = \frac{S(\omega)}{\tilde{C}_-(\omega)} \approx \frac{e^{j\omega\phi} \cdot G'_+(\omega) \cdot C_-(\omega)}{G'_-(\omega) \cdot C_-(\omega)} = e^{j\omega\phi} \frac{G'_+(\omega)}{G'_-(\omega)} \tag{7.6}$$

This equation can be expressed in terms of phase:

$$\angle\tilde{G}'(\omega) \approx \omega \cdot \phi + \angle G'_+(\omega) - \angle G'_-(\omega)$$

Moreover, since $\angle G_-(\omega) = -\angle G_+(\omega)$ (see appendix A):

$$\angle\tilde{G}'(\omega) \approx \omega \cdot \phi + 2\angle G'_+(\omega) \tag{7.7}$$

In conclusion, the distortion made by the glottal pulse $G'_+(\omega)$ in equation (7.7) is far from linear in the low frequencies (see phase plots in sec. 2.5.3). Therefore, if the glottal pulse is not considered, this distortion can have a serious impact on the estimation of the linear-phase term. Accordingly, most of the existing methods usually remove an estimation of the spectrum of the radiated glottal source when computing the VTF estimate (e.g. using pre-emphasis). Accordingly, equation (7.5) can be changed:

$$\tilde{C}_-(\omega) = \mathcal{E}_-\left(e^{j\omega\phi} \cdot G'_+(\omega) \cdot C_-(\omega) \cdot (1 - \mu e^{j\omega})\right)$$

If the amplitude spectrum of the pre-emphasis is sufficiently close to the one of the inverse of the radiated glottal pulse, one can assume that $\tilde{C}_-(\omega) \approx C_-(\omega)$. Finally, the estimated radiated glottal pulse is:

$$\tilde{G}'(\omega) \approx \frac{e^{j\omega\phi} \cdot G'_+(\omega) \cdot C_-(\omega)}{C_-(\omega)} = e^{j\omega\phi} \cdot G'_+(\omega)$$

The bias made by the maximum-phase component of the pulse is thus divided by 2 compared to eq. (7.6).

The next section will present the new theoretical ideas used in this context and the following sections will describe and discussed the proposed method.

7.1 The minimum of the radiated glottal source

According to the discussion above, the main two ideas used in the proposed method are: 1) the VTF envelope will be estimated using a glottal model (instead of the usual a pre-emphasis). 2) Since the minimum of the radiated glottal pulse seems the most significant instant according to most glottal models (see fig 2.4), the minimum of the radiated glottal pulse will be estimated. Accordingly, this method is termed GCIGS (GCI estimation method using a Glottal Shape). Following the same processes as above and the idea 1), the VTF is first estimated using:

$$\tilde{C}_-(\omega) = \mathcal{E}_- \left(\frac{S(\omega)}{G^\theta(\omega) \cdot L(\omega)} \right) \tag{7.8}$$

where $G^\theta(\omega)$ is a glottal model parametrized by θ and $L(\omega)$ is modeled by the time-derivative $L(\omega) = j\omega$. Therefore, one can expect the following:

$$\tilde{C}_-(\omega) = \mathcal{E}_- \left(\frac{e^{j\omega\phi} \cdot G(\omega) \cdot C_-(\omega) \cdot L(\omega)}{G^\theta(\omega) \cdot L(\omega)} \right) \approx \frac{G_-(\omega)}{G^\theta_-(\omega)} \cdot C_-(\omega) \tag{7.9}$$

And the estimation of the radiated glottal source is:

$$\tilde{G}'(\omega) = \frac{S(\omega)}{\tilde{C}_-(\omega)} \approx \frac{e^{j\omega\phi} \cdot G(\omega) \cdot C_-(\omega) \cdot L(\omega)}{C_-(\omega) \cdot G_-(\omega)/G^\theta_-(\omega)} \tag{7.10}$$

$$\tilde{G}'(\omega) \approx e^{j\omega\phi} \cdot \frac{G(\omega) \cdot G^\theta_-(\omega)}{G_-(\omega)} \cdot L(\omega) \tag{7.11}$$

In this equation, one can see the following results:

- The pure linear phase term $e^{j\omega\phi}$ is the one of the real glottal source. Accordingly, the position of the glottal pulse is kept in (7.11). Obviously, this condition has to be satisfied for GCI estimation.
- The real glottal pulse $G(\omega)$ and its minimum-phase version $G_-(\omega)$ have exactly the sample amplitude. Their division is thus an all-pass filter. Consequently, the amplitude of the glottal pulse in (7.11) is only defined by the glottal model $G^\theta_-(\omega)$.
- The phase spectrum of a minimum-phase signal is linked to its amplitude spectrum. Therefore, $|G^\theta_-(\omega)| = |G_-(\omega)|$ implies that $G^\theta_-(\omega) = G_-(\omega)$. Consequently, if the amplitude spectrum of the glottal model $|G^\theta_-(\omega)|$ is equal to the amplitude spectrum of the real glottal pulse $|G_-(\omega)|$, these two terms cancel each other in equation (7.11). Consequently, to retrieve the real radiated glottal source, only the amplitude spectrum has to obey to the glottal model (the phase spectrum of the glottal model does not matter).

Then, most glottal models assume that a strong negative impulse exist on their time-derivative (figure 2.4). Therefore, according to idea 2), the proposed idea is to find this impulse in $\tilde{G}'(\omega)$ (fig. 7.1).

$$t_{GCI} = \underset{t}{\operatorname{argmin}} \left(\mathcal{F}^{-1} \left(\tilde{G}'(\omega) \right) \right)$$

Polarity

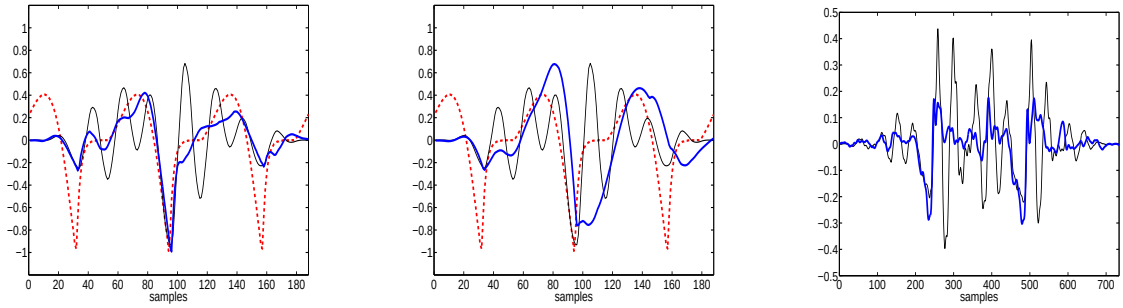
The polarity of the signal has to be known. Indeed, in the time domain, the minimum of the time-derivative of the glottal pulse is assumed to correspond to the GCI. If the polarity is false, the proposed method will be confused with positive peaks.

Effect of the shape parameter

Finally, one can study the effect of an error $\Delta\theta$ of the shape parameter θ of the glottal model. In order to study this effect, the glottal model is assumed to be able to fit the real glottal pulse. Therefore, from equation (7.11), one can write:

$$\tilde{G}'(\omega) \approx e^{j\omega\phi} \cdot \frac{G^{\theta^*}(\omega) \cdot G_-^{\theta^*+\Delta\theta}(\omega)}{G_-^{\theta^*}(\omega)} \cdot L(\omega)$$

where θ^* is the optimal parameter of the real glottal pulse and $\theta = \theta^* + \Delta\theta$ is the estimated shape parameter. One can see that the term representing the error of the shape parameter is minimum-phase. Therefore, this term can only delay the energy of the estimated glottal pulse. Figure 7.1 shows examples of $\tilde{G}'(\omega)$ for two synthetic signals and a real signal. The first synthetic signal (a) is computed without parametrization error, $\Delta Rd = 0 \Rightarrow G_-^{\theta^*+\Delta\theta}(\omega)/G_-^{\theta^*}(\omega) = 1$. The second one (b) is computed with a parametrization error corresponding to $\approx 50\%$ of the parameter range. A few of the theoretical elements



(a) Synthetic signal (b) Synthetic signal with an error ΔRd (c) Real signal

Figure 7.1: Examples of radiated glottal source $\tilde{G}'(\omega)$ in the time domain: (a,b) Synthetic signals using the LFRd glottal model with $f_0 = 128\,\text{Hz}$ $Rd = 1$ and a /e/: the waveform in thin black line, the synthetic source in dashed red line and $\tilde{G}'(\omega)$ in thick blue line. (c) Real signals: the waveform in thin black line and $\tilde{G}'(\omega)$ in thick blue line.

can be seen in this figure:

- In the synthetic example (a): $C_-(\omega)$ is not perfectly reconstructed because the vocal-tract filter response is sampled by the harmonic structure induced by the periodicity. Consequently, ripples appears all along $\tilde{G}'(\omega)$. However, the negative peak of the GCI exceed clearly these ripples.
- In the synthetic example with the parametrization error (b): the negative peak of $\tilde{G}'(\omega)$ is still prominent but slightly blurred. This result is very important for GCI detection because this peak position is hardly contested by other ripples. Consequently, a rough estimate of θ should be sufficient for such an estimate.

- In the real example (c): The negative peaks, which are assumed to correspond to GCIs, are prominent and clearly distinct from the positive peaks.

7.2 The method using a glottal shape estimate

Using the theoretical results of the previous section, this section presents the proposed method with a few technical details. The GCI estimation process is made of two levels: First, for a given instant, the strongest GCI is estimated among a small speech segment (3 periods). Then, a subdivision algorithm is used to recover all GCIs in a speech utterance. In the proposed implementation, the LFRd glottal model is used. Firstly, the speech spectrum $S(\omega)$ is computed using the DFT of a speech segment of 3 periods windowed by a hanning function. Secondly, in order to estimate the VTF $\tilde{C}_-(\omega)$ using equation (7.8), an estimate of Rd is necessary. In the following, the method minimizing the error function MSPD^2 is used [1]. To compute the minimum-phase envelope $\mathcal{E}_-(.)$, the TE envelope is used, mainly because of its precision and its ability to model zeros of the VTF. Finally, once the VTF estimate is retrieved, the estimation of the radiated glottal source is computed using equations 7.10.

Effect of the analysis window

The speech spectrum $S(\omega)$ is computed using a window function. Therefore, the radiated glottal source is estimated according to:

$$\tilde{G}'(\omega) = \frac{W(\omega) \circledast S(\omega)}{\tilde{C}_-(\omega)} \tag{7.12}$$

However, if the main lobe of the window decreases fast enough, one can assume:

$$\frac{W(\omega) \circledast S(\omega)}{\tilde{C}_-(\omega)} \approx W(\omega) \circledast \frac{S(\omega)}{\tilde{C}_-(\omega)} \tag{7.13}$$

Note that, the bigger the variation of the amplitude spectrum of $\tilde{C}_-(\omega)$, the bigger the difference between the two sides of this equation. However, since $\tilde{C}_-(\omega)$ has a relatively smooth amplitude spectrum, this assumption should not introduce a significant error. Consequently, in the following, the effect of the window function is assumed to remain in the final estimated radiated glottal source as in equation (7.13). Figure 7.1 shows that the effect of the window function is visible on $\tilde{G}'(\omega)$.

7.2.1 Estimation of a GCI in one period

The effect of the window can move the minimum of $\tilde{G}'(\omega)$. Therefore, from an arbitrary starting position, an iterative method is proposed to converge to the closest GCI (Algorithm 3). From our experiments, this algorithm usually stops after 3 or 4 iterations.

7.2.2 Estimation of GCIs in a complete utterance

A method estimating the GCIs of a complete utterance has to take care of different aspects: 1) No GCI should be missed. 2) To minimize computation time, one GCI should be estimated only once. 3) An error of one GCI detection should not be propagated to detection of other GCIs. The main algorithmic idea presented here is the following. A segment is recursively subdivided into two smaller segments and

[1] In the original presentation of this work [DRR09a], a rough estimate was used [DRR09b]

Algorithm 3 Estimation of a GCI in one period

From a given initial time t_{GCI}
Estimate Rd and synthesize $G^{Rd}(\omega)$
repeat
  Compute $S(\omega)$, the DFT of 3 windowed periods of $s[n]$ around t_{GCI}
  Compute $\tilde{C}_-(\omega) = \mathcal{E}_-\big(S(\omega)/G^{Rd}(\omega) \cdot j\omega\big)$
  Compute $\tilde{G}'(\omega) = S(\omega)/\tilde{C}_-(\omega)$
  Locate $t'_{GCI} = \text{argmin}(\mathcal{F}^{-1}(\tilde{G}'(\omega)))$
until $|t'_{GCI} - t_{GCI}|$ is smaller than a sample duration. Then $t_{GCI} = t'_{GCI}$

this recursion stops when the segment duration is shorter than a period. To define the shortest segment, the fundamental period $T_0(t)$ is assumed to be known *a priori* at any instants t of the utterance. The whole procedure is summarized in Algorithm 4.

Algorithm 4 Estimation of GCIs in a complete utterance

Create a pair $[t_s, t_e]$ with the start and end time of the speech utterance. Add this pair to a list
Chose a threshold α (as discussed below)
while the list is not empty **do**
  Take a pair $[t_s, t_e]$ from the list
  Obtain the closest GCI t_{GCI} from the middle position $t_m = (t_s + t_e)/2$, using Algorithm 3
  If $\alpha \cdot (t_{GCI} - t_s) > T_0$ at time t_{GCI}, add the time segment $[t_s, t_{GCI}]$ to the list
  If $\alpha \cdot (t_e - t_{GCI}) > T_0$ at time t_{GCI}, add the time segment $[t_{GCI}, t_e]$ to the list
end while

The α parameter controls the minimum segment duration where a GCI is assumed to exist. Ideally, for a constant f_0, α should be equal to 1. However, the f_0 variations are not negligible in speech. Therefore, a tolerance on the f_0 estimate should be considered. If the T_0 estimate is smaller than $(1-\alpha) \cdot T_0^*$, the method creates a false alarm. Conversely, if the T_0 estimate is bigger than $\alpha \cdot T_0^*$, the method misses a GCI. Note that, it is more convenient to create a false alarm than miss a GCI because duplicated GCIs can be removed at the end of the estimation procedure. By evaluation of the method results using various α values between 0.5 and 1 (see evaluation procedure in sec. 8.2.1), $\alpha = 2/3$ has been used in the following. Additionally, one can make the following remarks. Firstly, the Algorithm 4 subdivides the initial interval into sub-intervals smaller than a period. Thus, no GCI should be missed. Secondly, search intervals are between two GCIs. Thus, no GCI should be detected twice. Thirdly, the subdivision process uses two different GCIs t_s and t_e and both have to be erroneous to maximize the probability that an error is propagated inside the time segment $[t_s; t_e]$.

Diagram of figure 7.2 summarizes the procedure and figure 7.3 shows examples of GCIs estimation. An evaluation of the efficiency of this method will be presented in the next chapter.

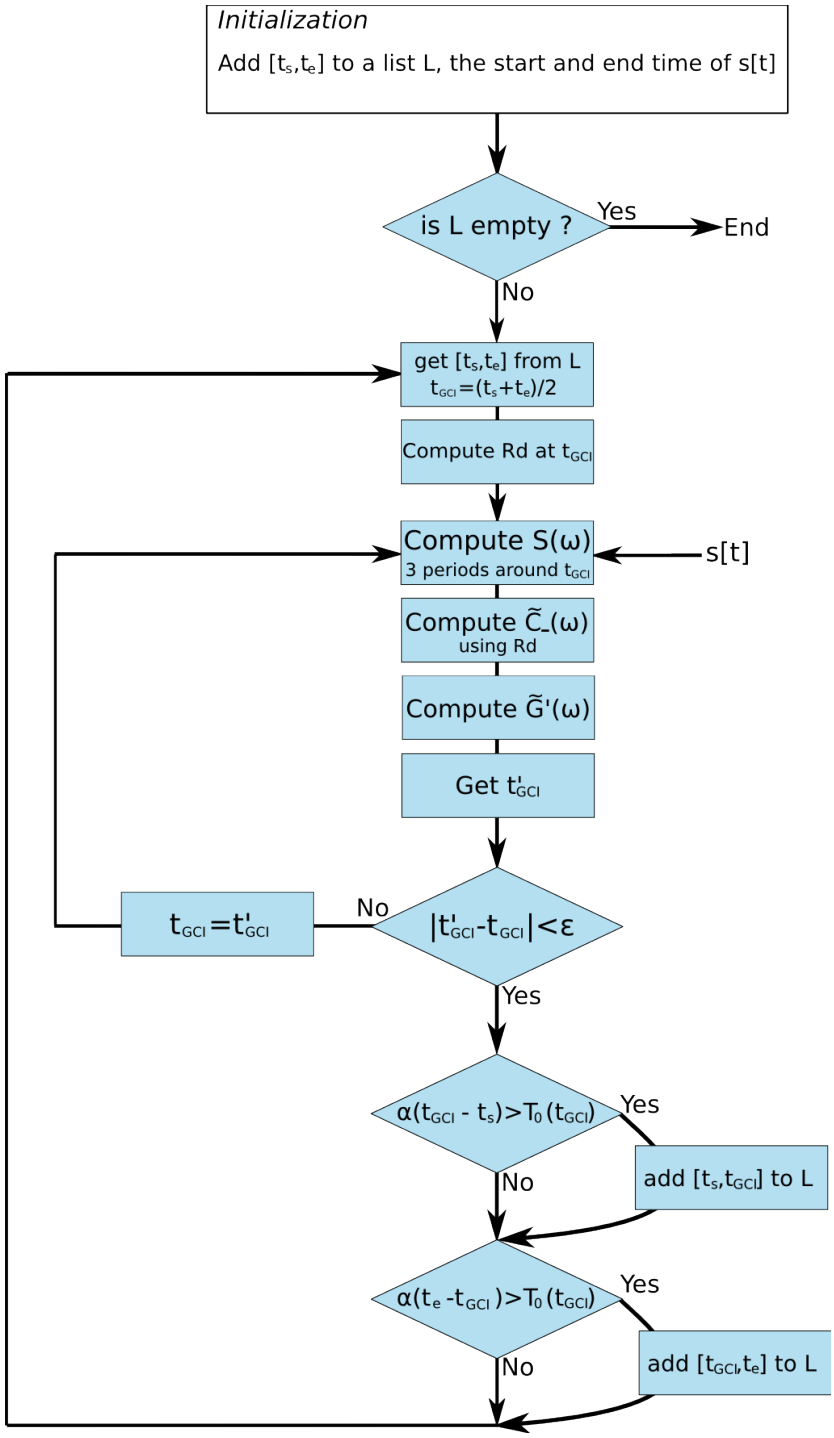


Figure 7.2: Diagram of the method GCIGS

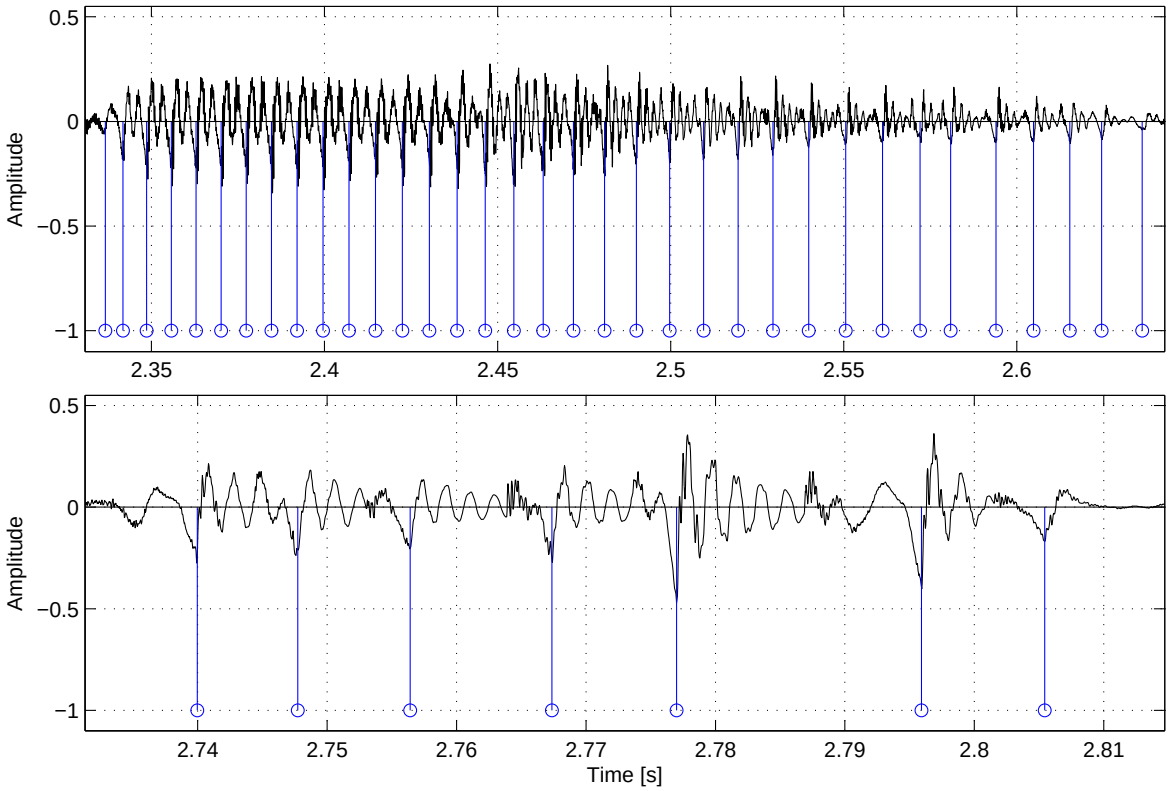


Figure 7.3: GCI estimation on a sustained phone and a segment of vocal fry.

7.3 Evaluation of the error related to the shape parameter

For many reasons, the Rd estimate can be not accurate. Therefore, it is interesting to evaluate the error of the GCI estimator related to an error of the shape parameter. The following evaluation is carried out using a synthetic signal with a known shape parameter. This reference signal (eq. 7.14) is controlled by the LFRd glottal model, a random GCI t^*_{GCI}, a fixed fundamental frequency $f_0 = 128Hz$. The Maeda's synthesizer is used to generate 13 different VTFs $C^p_-(\omega)$ of phoneme p covering the vocalic triangle (see appendix D).

$$S(\omega) = \left(e^{j\omega t^*_{GCI}} \cdot G^{Rd^*}(\omega) \cdot \left[\sum_{l \in \mathbb{N}} e^{j\omega l / f_0} \right] \right) \cdot C^p_-(\omega) \cdot j\omega \tag{7.14}$$

To estimate the bias and standard-deviation of the GCI estimator among different voice qualities, the shape parameter Rd varies in $[0.3; 2.5]$. Moreover, the error is computed from 8 signals with different positions t^*_{GCI}, to obtain a valuable statistical evaluation. Finally, to evaluate the estimator reliability, a biased shape parameter $\Delta Rd + Rd^*$ is used as input of the GCIGS method. The results are shown in figure 7.4. Note that a systematic bias exists which is almost linear between -1 and 0.3. In this same interval, the error standard-deviation is almost negligible. However, as shown in figure 7.1, ripples exist on the radiated glottal source which are created by a miss cancellation of the formants. Therefore, one can expect that outside of the interval $[-1; 0.3]$, such ripples are confused with the peak of the GCI.

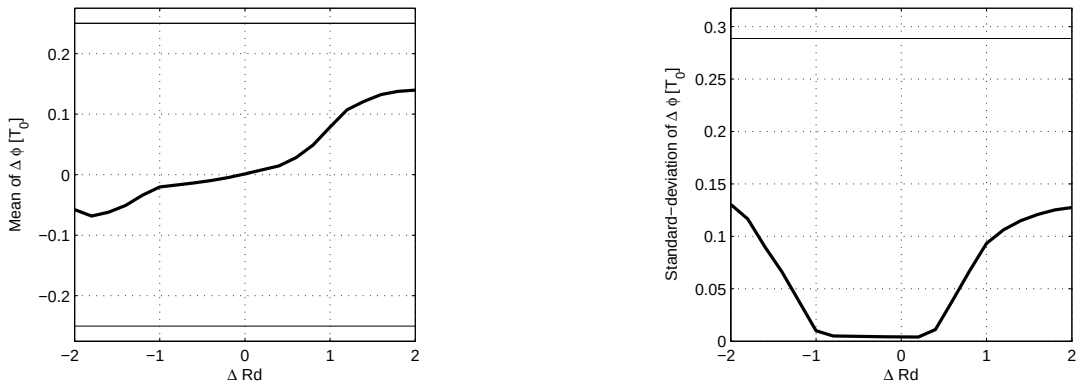


Figure 7.4: In thick line, the error of the GCIGS estimator related to an error of the shape parameter Rd. In thin line, the mean estimator (this estimator return a random delay in $[0.5/f_0; 0.5/f_0]$) for the plot of the standard-deviation. For the plot of the mean error, the absolute value of the position error is used $|t_{GCI} - t^*_{GCI}|$).

Conclusions

- Theoretically, to estimate a maximum of energy of the glottal source, either in the time domain or the spectral domain, the amplitude spectrum of the glottal source has to be taken into account. Otherwise, a significant bias is made by the phase spectrum of the glottal pulse (see eq. 7.7). In current literature, to remove the amplitude spectrum of the glottal source, the speech signal is usually pre-emphasized.
- Two ideas are used in the proposed method GCIGS (GCI using a Glottal Shape estimate): 1) Instead of the pre-emphasis, a glottal model is used to estimate the VTF. 2) The minimum of the radiated glottal source is estimated instead of a global maximum of energy. Moreover, in this proposed method, one can see that:
  - The amplitude spectrum of the estimated radiated glottal source is the one of the used glottal model
  - Moreover, if the amplitude spectrum of the glottal model obeys to the real glottal pulse, the phase spectrum of the estimated radiated glottal source is the one of the real glottal source. Therefore, the phase spectrum of the used glottal model has no impact on the method results.
  - Using the LF^{Rd} glottal model, an error of the Rd parameter in the interval $[-1; 0.3]$ implies a negligible error of the GCI estimation. A very precise estimation of this parameter is thus not necessary for the proposed method.

Chapter 8

Evaluation of the proposed estimation methods

In the previous chapters, four methods have been proposed to estimate the shape parameter of a glottal model. A first method is based on the Mean Squared Phase (MSP eq. 5.5). A second one, closely related to the first one, uses the discrete approximation of the frequency derivative (MSPD eq. 5.11). A third method removes the influence of the pulse position on the estimation of its shape (MSPD^2 eq. 6.2). A last method FPD^{-1} expresses the shape parameter in a quasi closed-form of the observed spectrum (based on the Functions of Phase-Distortion (FPD based on eq. 6.8)). Additionally, a GCI estimation method has been proposed which takes advantage of a Glottal model Shape (GCIGS sec 7). In this chapter, the efficiency of these proposed methods is evaluated with synthetic signals and ElectroGlottoGraphic (EGG) signals. A brief qualitative evaluation concludes this chapter using a few examples on real speech utterances.

8.1 Evaluation with synthetic signals

Similar to the evaluation of the GCIGS method, the synthetic signal of equation (8.1) is generated with a 44.1 kHz sampling frequency. This synthetic voice is controlled using the Rd^* shape parameter of the LF^{Rd} glottal model. A known fundamental frequency f_0 is used and the position of the pulses are delayed by a random ϕ^*. Then, Gaussian noise $n^{\sigma_g}[n]$ of standard deviation σ_g called *glottal noise*, is added to the glottal source and one Gaussian noise $n^{\sigma_a}[n]$ called *environment noise* is added to the voiced signal. Filters $C_-^p(\omega)$ are synthesized using the Maeda's simulator to model 13 different voiced phonemes p covering the vocalic triangle (see appendix D). Among these phonemes, 4 are nasalized. In comparison to estimated frequencies and bandwidths of AR models on real speech signals, an acoustic model allows to simulate the frequency response of the vocal-tract considering a constant impedance at the glottis level [Mae82a].

$$\begin{aligned} E(\omega) &= e^{j\omega\phi^*} \cdot G^{Rd^*}(\omega) \cdot \big[\sum_{l\in\mathbb{N}} e^{j\omega l/f_0}\big] + \mathcal{F}(n^{\sigma_g}[n]) \\ s[n] &= \mathcal{F}^{-1}\big(E(\omega) \cdot C_-^p(\omega) \cdot j\omega\big) + n^{\sigma_a}[n] \end{aligned} \tag{8.1}$$

The amplitude of the Gaussian noise is set so as to control the Signal to Noise Ratio (SNR) with either the glottal source or the voiced signal. As discussed in section 4.1, the deterministic model of the voice can be irrelevant at high frequencies. Therefore, for all of the compared methods, the analyzed signal is resampled to $16kHz$ and the error measure is limited to a Voiced/Unvoiced Frequency fixed to $2kHz$. Using this latter, the estimation of the parameters is mainly influenced by the lowest frequencies of the glottal source, mainly by the glottal formant and less by the spectral tilt. This value is kept constant in order to have all of the compared methods equally affected by this limit. The influence of this value on the results of the methods will be discussed for real signals in section 8.2.2. Using this synthesis, the methods efficiency related to the fundamental frequency, the glottal noise and the environment noise is evaluated. Additionally, in the following comparison figures, theoretical limits are given through the *mean* estimator. This method returns the mean value of the *Rd* parameter range, without taking into account the observed signal. For the plots about the mean error, the absolute value of the error is used instead of its relative value.

In addition to the proposed methods, three other methods are added to the comparison: the Iterative Adaptive Inverse Filtering (IAIF) and two Minimum/Maximum-phase decomposition methods, the Complex Cepstrum (CC) and the ZZT. Conversely to the proposed methods, these three last methods first estimate the glottal source. Then, a glottal model is fitted on one period of the glottal source. See Appendix B for more details on the estimation of shape parameters using these methods.

8.1.1 Error related to the fundamental frequency

The reconstruction of the VTF is related to the harmonics which depend on the fundamental frequency f_0. It is thus interesting to known to which extent the fundamental frequency influences the estimators results. In figure 8.1, for each f_0 value, the estimation error of the compared methods is computed for the 13 VTFs and the random delay ϕ^*. For the methods based on MSP and MSPD, the initial shape value is given by the method based on MSPD^2. The initial position is given by the ideal value ϕ^* delayed by a random number in $[-0.1/f_0; 0.1/f_0]$ to simulate an initial error of position. The error is computed 16 times with different initial positions in order to obtain a sufficiently high confidence on the estimate of the mean and standard-deviation of the error. Finally, to focus on the influence of f_0, the noise signals are set to zero in equation (8.1). Figure 8.1 shows the mean and the standard-deviation of the estimation error, in terms of position error and shape parameter error.

As expected, the variance of the estimators increases with f_0 since the sampling of the VTF by f_0 does not provide enough information to reconstruct the VTF perfectly. Concerning the methods based on mean squared phase: The MSPD is the worst of the proposed methods because the position parameter can offset the shape error as discussed in section 5.2 (This method is thus discarded in the evaluations using real signals). Additionally, the 2nd order phase difference of the MSPD^2 removes the information provided by the average phase spectrum of the glottal model. Therefore, the method based on MSP is more precise than MSPD^2 (moreover, using the phase difference, the phase frequency derivative is approximated using discrete frequencies). Finally, compared to the theoretical limits, the position of the pulse is significantly more precise than the estimation of the shape parameter. Concerning the method FPD^{-1}: The estimation of the VTF is implicit in equation (6.5), whereas the same computation is explicit for MSP (eq. 5.3). Therefore, the computation of the minimum-phase component of the speech signal is not the same between these two classes of methods. For example, the extrapolation of the DC component is not the same between equation (6.5) and equation (5.3).

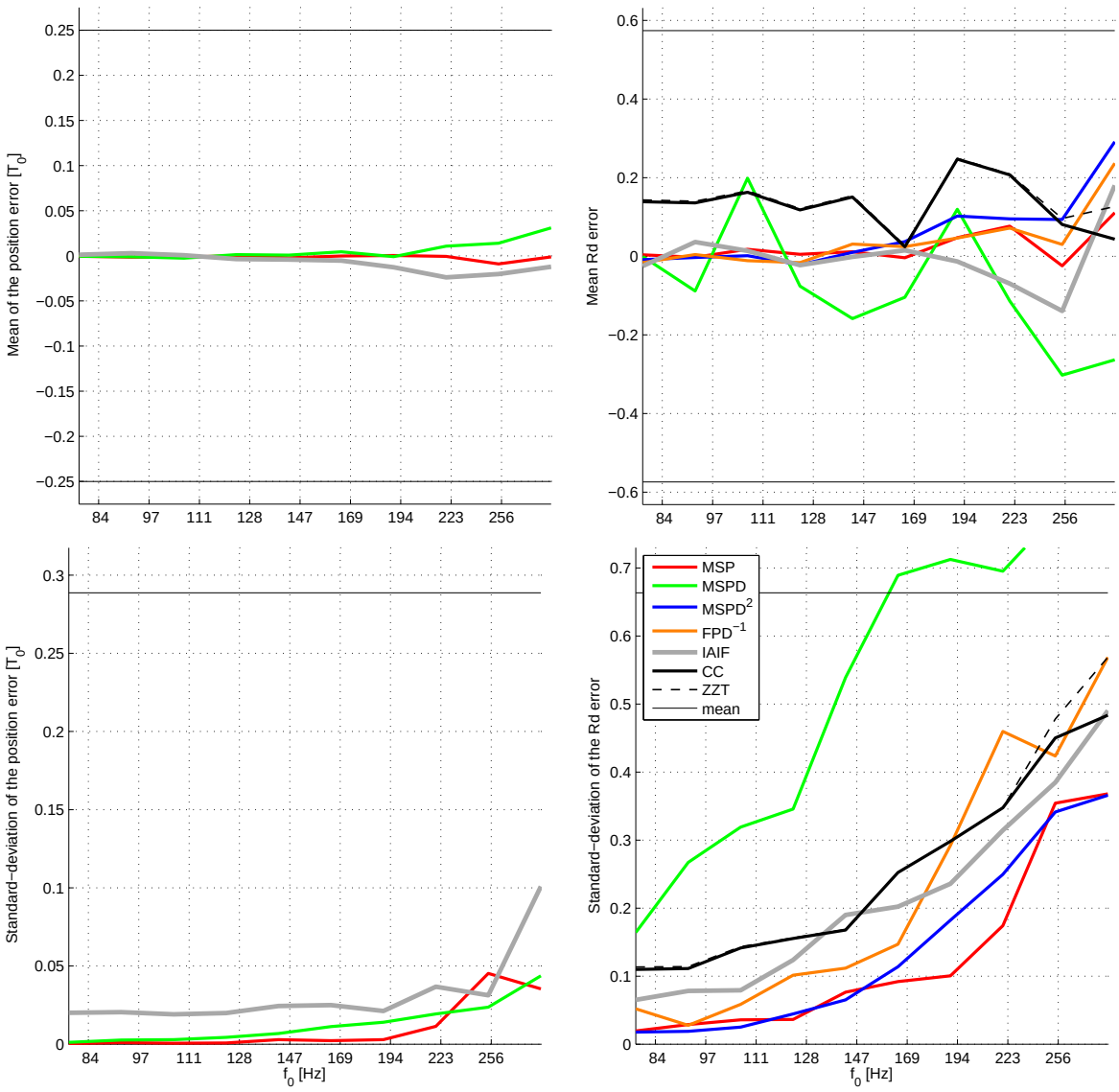


Figure 8.1: Mean and standard-deviation of the Rd error related to the fundamental frequency f_0.

8.1.2 Error related to the noise levels

This second test evaluates the influence of the noise levels σ_g and σ_e on the compared methods. To obtain a satisfying confidence in this evaluation, the error is computed 16 times for each σ value with the 13 different VTFs and a random position ϕ^*. To focus on the influence of the noises, f_0 is fixed to $128Hz$. In addition, when one noise is tested, the other one is set to zero. The results are shown in figures 8.2 and 8.3.

For equivalent SNR, one can see that the efficiency of the estimators are less disturbed by environment noise than by glottal noise. Moreover, the efficiency of all of the methods decrease rapidly when the glottal noise level increases. This can raise a serious issue in the presence of aspiration noise in breathy vowels. For low noise levels, the most reliable methods are the proposed ones. However, although the IAIF efficiency reduces significantly for glottal noise, this method seems the most robust for environment noise. The results of the CC and ZZT methods are close to each other. These methods are outperformed by the other methods in low noise conditions whereas, in high noise conditions, their efficiency are between those of the methods based on MSP and MSPD^2. For low noise levels, like in the test with f_0 variations, the methods based on MSP and MSPD^2 seem more efficient than the method FPD^{-1}.

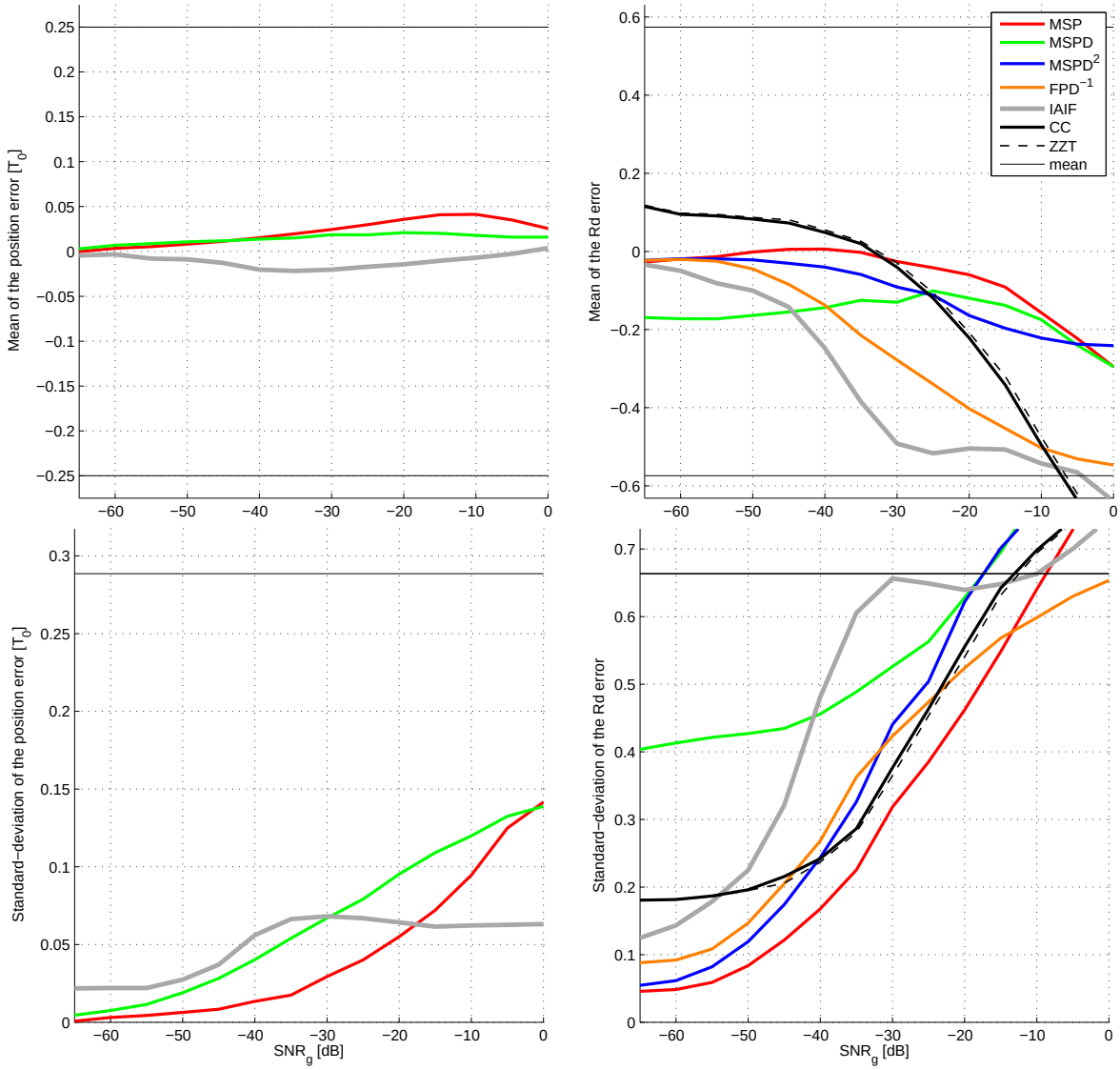


Figure 8.2: Mean and standard-deviation of the *Rd* error related to glottal noise.

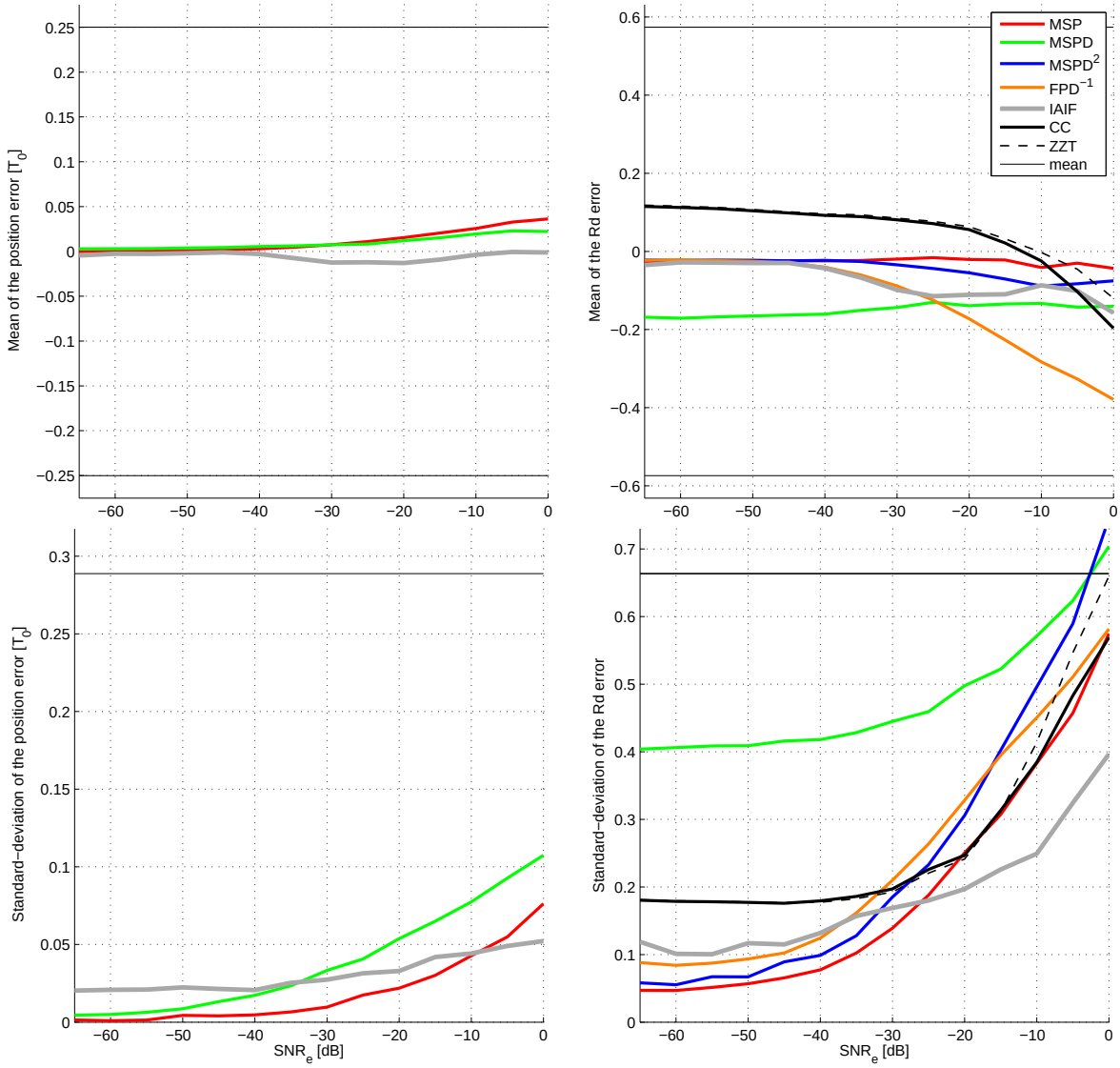


Figure 8.3: Mean and standard-deviation of the *Rd* error related to environment noise.

8.2 Comparison with electroglottographic signals

The electroglottography (EGG) is a non-invasive tool used in phoniatry to retrieve features of the motion of the vocal folds (see fig. 8.4). Among these features, one can obtain the GCIs in a given utterance [TN09]. Additionally, the open-quotient O_q can be computed by detecting the instant of opening of the glottis [HdDC04]. Assuming a high correlation between the glottal source and the motion of the vocal folds, reference sets of GCIs and O_q parameters can be created and compared to the estimation of the parameters of the proposed methods.

Four databases are used in this evaluation: APLAWD [LBN87] which is made of 5 different utterances of about 3 seconds pronounced by 5 different English female speakers and 5 different English male speakers; three CMU Arctic databases [KB03] commonly used for speech synthesis: two American male voices (*bdl* and *jmk*) and one American female voice (*slt*). Only the first 32 utterances of each Arctic database are sufficient to obtain more than 5000 comparison pairs. Note that, whereas the APLAWD database allows to evaluate the methods among various speakers (10 speakers, 5 utterances), each Arctic database has a larger phonemes variation (1 speaker, 32 utterances). All these databases are recorded with synchronized EGG and acoustic signal. The SIGMA algorithm [TN09] is used to retrieve the GCIs of the EGG recordings and the DECOM algorithm [HdDC04] is used to estimate the open quotient. In the following, the initial values used by the method based on MSP are given by the methods based on MSPD^2 and GCIGS. The fundamental frequency f_0 is estimated using the YIN method [dK02]. Moreover, the evaluation is made on voiced segments only. These segments are computed from the EGG signal. A time in the EGG signal is defined voiced if there is a reference GCI closer than half a period.

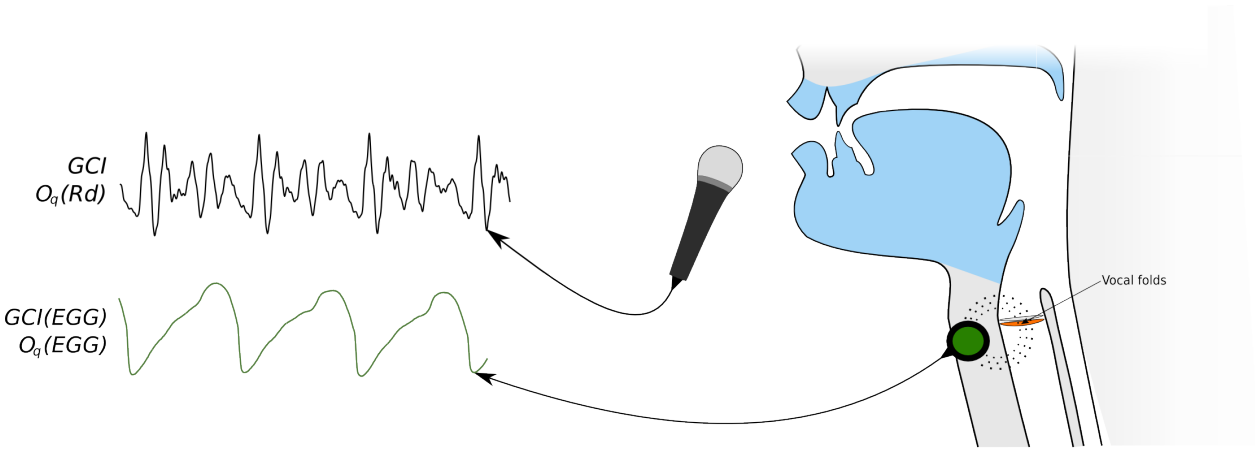


Figure 8.4: Evaluation of estimated values using electroglottography.

8.2.1 Evaluation of GCI estimates

In this test, the reference GCIs of the EGG signal are compared to the GCI of the LFRd glottal model. Due to the propagation delay between the EGG and the waveform, the reference GCIs and the detected GCIs are synchronized for each utterance by maximizing their correlation. Four methods are compared: the method based on MSP, the GCIGS method, the DYPSA method [KNB02] and another method based on Group-Delay (GD) [SY95, Fer04]. The synchronization between the reference GCIs and the detected GCIs is used for all the compared methods. Figure 8.5 shows the comparisons results.

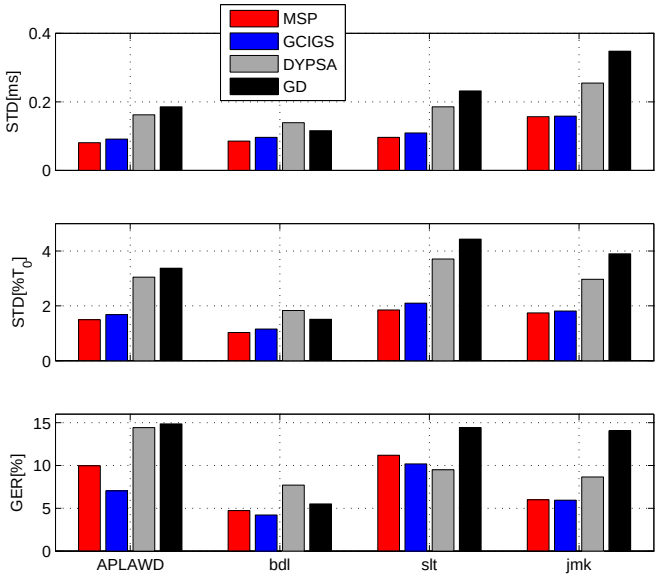


Figure 8.5: Evaluation of GCI estimation methods with 4 databases. STD is the standard-deviation computed through the interquartile range of the duration between the reference and the estimated GCIs, given in milliseconds $[ms]$ and in percent of the period $[\%T_0]$. The Gross Error Rate (GER) is the percent of that same durations$> 0.1 \cdot T_0$ among all comparison pairs.

In conclusion, as expected, the method using MSP slightly improves the precision of the GCIGS method. Indeed, by joint optimization of the shape and the position, the phase spectrum of the convolutive residual is closer to linear than without joint estimate. However, the GCIGS method assumes that a prominent peak exists in a period of the time derivative of the glottal source whereas the method based on MSP assumes that the whole phase spectrum of the glottal source corresponds to the one of the LF model. The hypothesis of the GCIGS method is thus weaker than the hypothesis of the MSP based method. With real signals, it can explain why the GCIGS method can be more robust than MSP (smaller Gross Error Rate (GER) for all databases except *jmk*). Finally, compared to the state of the art, the joint estimation of the shape and the position seems not to improve the results much more than the GCIGS method does. Removing the amplitude spectrum of a glottal model when computing the VTF has much more impact on the results (has been done in both GCIGS and MSP) than using the phase spectrum of this glottal model (has been done with MSP only).

8.2.2 Evaluation of the shape parameter estimate

In this test, the open quotient O_q measured on the EGG signals is compared to the one predicted from the estimated Rd parameter using the prediction formula of equations (2.1).

The weighting of the error functions varies among the compared methods. Indeed, the methods based on glottal source estimation (ie. IAIF, CC and ZZT) weight the mean squared error of the LF fitting in the spectral domain according to the estimated glottal source. Therefore, the glottal formant around the first three harmonics is thus reinforced compared to the spectral tilt in high frequencies. Conversely, in the proposed methods, the weighting among the harmonics is uniform. Therefore, the influence of the weighting on the efficiency of the estimators has to be evaluated. Accordingly, using APLAWD database only (the three other databases will be used as test set), figure 8.6 shows the error of the O_q estimation related to the number of harmonics taken into account in the error measure (or in the mean value for the method FPD^{-1}).

According to this figure, although it can be interesting to estimate the high frequency properties of a glottal model (e.g. the spectral tilt), the efficiency of the MSP, CC and ZZT based methods seems to substantially decreases when the considered frequency band increases. Conversely to the evaluation with synthetic signals, the MSPD^2 outperforms the MSP based method in this comparison with real signals. Although the optimization algorithm using MSP might not find the optimum, a grid search algorithm has shown the same results. In the case of real signals, the difference between the glottal model and the real glottal pulse introduces a distortion in the phase spectrum of the convolutive residual. Therefore, using MSP, ϕ and Rd can offset each other in order to minimize the error function. Conversely, the MSPD^2 can be systematically biased by this distortion but its variance can be smaller. In figure 8.6, the significant difference between MSP and MSPD^2 median values support this explanation. Comparing FPD^{-1} and MSPD^2 results, their efficiency are close to the ones evaluated with synthetic signals. Finally, the method based on MSPD^2 clearly outperforms all of the compared methods.

Using the results of figure 8.6, one can select the number of harmonics implying the smallest variance for each method: 4 for CC and ZZT, 5 for MSP, 6 for FPD^{-1}, 7 for MSPD^2 and IAIF. Figure 8.7 shows the corresponding results for the four databases. In conclusion, whereas the results of the methods based on MSP and FPD^{-1} vary significantly among the evaluated voices (*bdl*, *slt* and *jmk*), the variance of the method based on MSPD^2 is systematically smaller than the one of the methods related to the state of the art (IAIF, CC and ZZT).

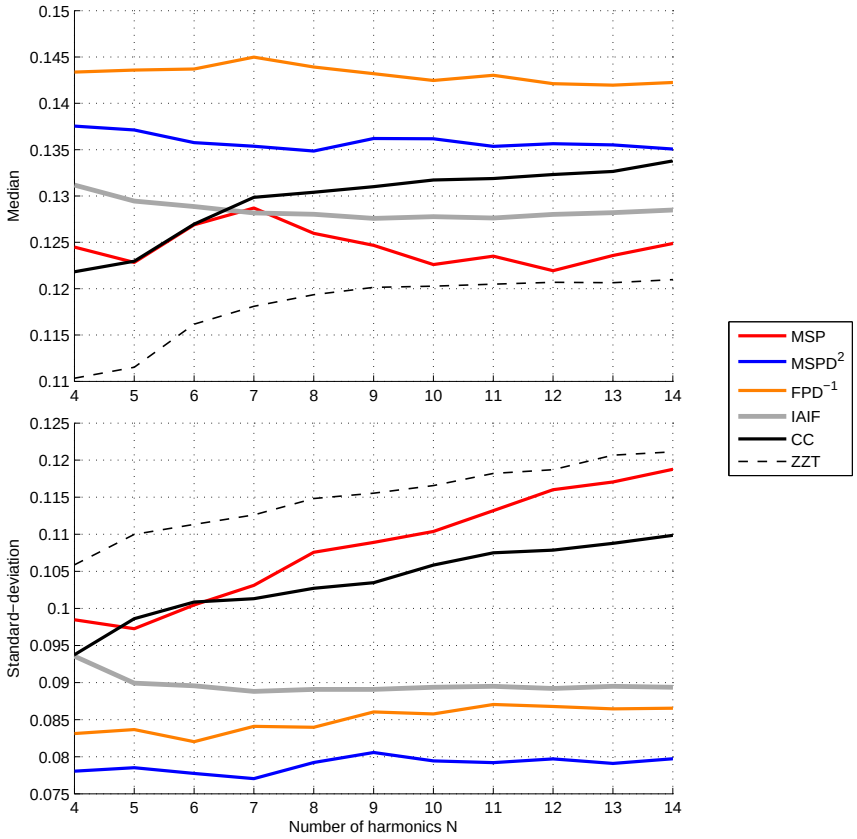


Figure 8.6: Median and standard-deviation of the O_q estimation error related to the number of harmonics taken into account in the estimators (only the APLAWD database is used).

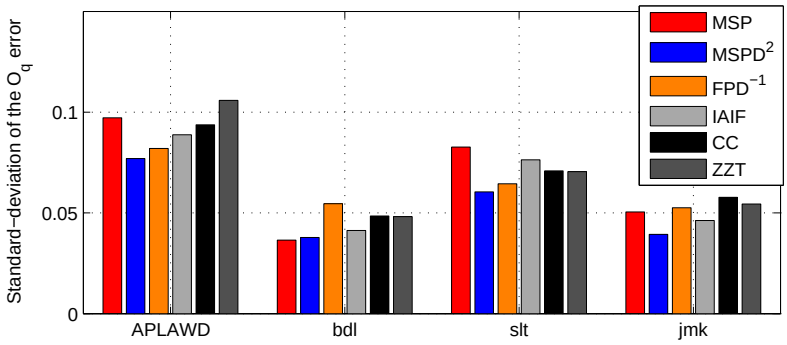


Figure 8.7: Standard-deviation of the O_q estimation error (computed through the interquartile range).

8.3 Examples of Rd estimates on real signals

In addition to the quantitative evaluations given above, it is interesting to qualitatively assess the proposed methods. Accordingly, examples of estimates of the Rd parameter are shown in this section.

Figure 8.8 shows an example of a sustained open /e/ recorded with High-Speed-Videoendoscopy (HSV) (sample `USC08.ouv_renf.50.*`). In this recording, the sound moves from a breathy phonation to a tense phonation. In addition to the Rd estimation based on MSP, MSPD^2 and FPD^{-1}, two measurements are also shown. A first Rd value is predicted from the open quotient of the EGG signal and a second Rd value is predicted from the T_0-normalized duration between the extremum of the glottal area derivative ($I_q = (t_e - t_i)/T_0$). The glottal area is estimated from the HSV images (see Appendix C). As expected, these two measurements move from values corresponding to lax to tense configurations. Moreover, this behavior is the same for all the three used methods.

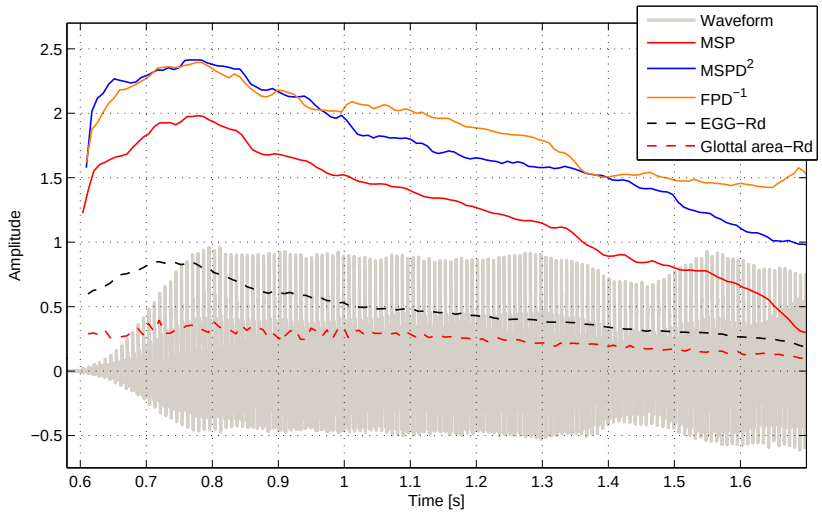


Figure 8.8: Rd estimates on a sustained open /e/ from a breathy to a tense phonation.

Figure 8.9 shows curves of Rd parameter on four real utterances. Firstly, as shown in figure 8.6, the systematic bias of the three proposed methods is different. Moreover, it is difficult to evaluate which estimate is the less biased. Indeed, only a correlation between EGG signals and estimated O_q values can be considered but their comparisons in absolute values are not possible (remember that the relation between the glottal area and the glottal flow is not straightforward. see sec. 2.3). Secondly, the curves of the Rd estimates are not perfectly smooth. Indeed, some irregularities are visible which seem inconsistent with the smooth behavior of the waveform (e.g. (b)$t = 2$ and (d)$t = 2.55$). One one hand, the methods based on MSP and MSPD^2 can be blocked at a local minimum. One the other hand, the method FPD^{-1} is a quasi closed-form expression of the observed data. Therefore, at least for this latter, these irregularities are related to the LFRd glottal model which can not fit the underlying real glottal pulse and not related to an optimization method. Consequently, either the full set of parameters (O_q, α_m, Q_a) of the LF model should be considered (with all the related optimization issues, see sec. 5.3), either another glottal model should be used. Note that the harmonic model can be also erroneous at transients (e.g. (a)$t = 0.72$ and (a)$t = 1.1$). Additionally, small variations of the Rd estimates are often visible which are not correlated to the EGG-Rd measurement (e.g. (b)$t = 1.55$ and (c)$t = 2.55$). Indeed, the glottal source is closer to the glottal flow than to the vocal folds. Thus, the articulatory configuration can have an influence on the

glottal flow whereas this influence is not measured by the EGG. In conclusion, for the study of the voice, the estimated *Rd* parameter have to be carefully interpreted. Additionally, for voice transformation and speech synthesis, it seems necessary to smooth this feature in order to obtain stable values for source-filter separation and machine learning method.

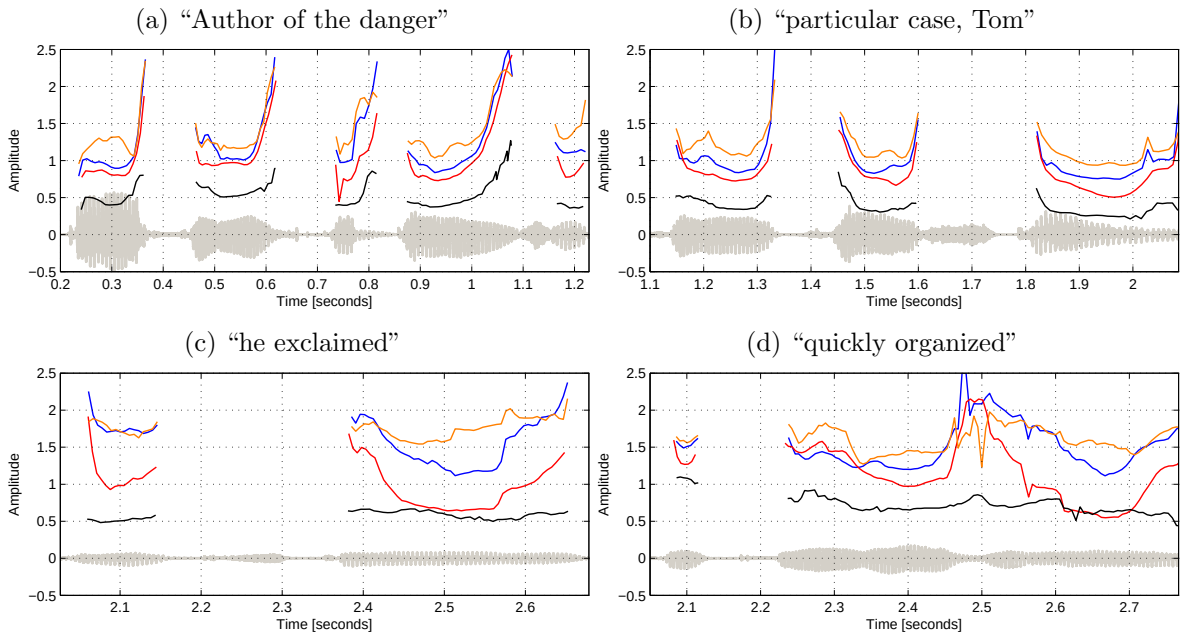


Figure 8.9: *Rd* estimates on real speech segments. First row, an English male voice (from *bdl*). Second row, an English female voice (from *slt*). The MSP in red, MSPD^2 in blue, FPD^{-1} in orange, the EGG-*Rd* parameter predicted from the EGG-O_q measurement in black and the waveform in gray.

Conclusions

- In a first test, the four proposed methods based on MSP, MSPD, MSPD^2 and FPD^{-1} have been evaluated with synthetic signals and compared to the methods IAIF, CC and ZZT.
  - As expected, the efficiency of all the methods decreases when the fundamental frequency increases.
  - For equivalent SNR, the efficiency of the estimators are less reduced by environment noise than by noise added at the source level, the glottal noise. With high environment noise level, only the IAIF method outperform all the proposed methods. Conversely, for low noise level (environment or glottal noises), the most efficient methods seem to be the ones using MSP, MSPD^2 and FPD^{-1}.
- In a second test, electroglottographic signals have been used to evaluate both position and shape estimated by the proposed methods.
  - The efficiency of the MSP based method has been evaluated using the Glottal Closure Instants estimated from EGG signals using the SIGMA method [TN09]. Four methods have been compared: the MSP based method, the GCIGS method and two other existing methods (the DYPSA method and a group-delay-based method). Globally, the GCIGS method clearly outperform the three other methods. Moreover, the GCIs estimated by this method are used as initial values for the MSP based method. Using this refinement, this evaluation shown that the number of gross errors of GCI position increases, whereas the precision is improved.
  - The DECOM method [HdDC04] has been used to estimate the open quotient on EGG signals. This quotient has been compared to the one predicted from the *Rd* estimates using: methods based on MSP, MSPD^2, FPD^{-1} and the existing methods IAIF, CC and ZZT. On one hand, the IAIF, CC and ZZT methods can estimate a glottal source which does not correspond to any glottal model. One the other hand, according to this evaluation, the MSPD^2 is more efficient than these methods to estimate the LF^{Rd} shape parameter.
  - The method based on MSPD^2 is the most efficient compared to existing methods and compared the other methods proposed in this study. Additionally, using this method, the estimation of the shape parameter is independent on the amplitude of the glottal model and independent on the position of this latter.

Part III

Voice transformation and Speech synthesis

Chapter 9

Analysis/synthesis method

The synthesis of voiced sounds has been studied since a long time [Fla72b]. Firstly by means of analog systems modeling the acoustics of the voice production [Fla57, Wei55, SKF53, Dud39]. Then by digital synthesizers of an acoustic model [Mae82a]. Synthesis by simulation of the mechanical behavior of the vocal folds has also been conducted [PHA94, ST95]. As argued in the introduction, the source-filter model is used in this study because of its simplicity and the wish to manipulate the perceived elements of the voice rather than to model all of the details of its production. Below, existing methods for voice transformation and synthesis using the source-filter model are enumerated. The method proposed in this study follows after this state of the art, and the next chapter will evaluate this new method.

9.1 Methods for voice transformation and speech synthesis

In order to transform a voice recording, there is mainly two different approaches. On one hand the speech waveform can be fully modeled and encoded into a relatively small set of parameters (e.g. STRAIGHT [KMKd99], WBVPM [Bon08], HNM [Sty96], FOF [GdR93, RPB84]). In the following, these methods will be termed *analysis/synthesis methods*. On the other hand, a part of the original signal can be reused in the transformed signal (e.g. Phase vocoder [Roe10, LD99a, FG66], PSOLA [VMT92, MC90, HMC89]). These methods will be termed *modification methods*. For example, combined with an envelope estimation method (e.g. TE, LP, etc. see sec. 4.3), the phase vocoder preserves the envelope residual in the transformed waveform. Therefore, a part of the original phase spectrum is not modeled but kept untouched by the transformations. Methods based on PSOLA assume that the signal inside a single window can be used without being modeled. Compared to analysis/synthesis methods, one can expect that modification methods are more robust because a part of the original signal is not modeled, which implies less influences of estimation errors. Conversely, one can expect that the modification methods are limited by the part which is not modeled. Accordingly, analysis/synthesis methods should be more flexible, but more sensitive to the estimation of their parameters.

For text-to-speech synthesis, mainly two different approaches exist. Firstly, synthesis by concatenation uses units extracted from a recording database (e.g. phones, diphones, non-uniform units) to reconstruct a given utterance [HB96, HMC89]. Thus, the units have to be properly selected from the database in order to obtain a sequence which is consistent in terms of certain perceived features (fundamental frequency, timbre, speech rate, etc.) [VK06]. Secondly, the parametric speech synthesis called *HMM-based synthesis* models statistically the parameters of an analysis/synthesis method using machine learning. A sequence

of parameters can be generated from the statistical model according to the wanted utterance [ZNY+07]. Whereas the first approach is not dependent on the quality of an analysis/synthesis method and provides currently better synthesis results than the HMM-based synthesis, this latter is assumed to have more flexibility in terms of applications. Indeed, methods modifying the identity, the expressivity or the voice quality can be applied to the generated parameter sequence, before the synthesis step. Since the presented research has focus on modeling voice production, only the HMM-based synthesis will be considered.

In the following, the most common methods are described which are used in voice transformation and speech synthesis.

Modification methods

The OverLap-Add (OLA) method is widely used in voice transformation: First, an utterance is segmented into frames using overlapped windows. Then these frames can be moved, duplicated, transformed, interpolated, etc. in order to create a new sequence of transformed frames which have to be properly combined to obtain the final transformed utterance. This mixing is obtained by a simple addition of the transformed frames according to their modified time position. To avoid phase cancellation issues, the inter frame correlation can be used in time domain, like proposed in the Synchronous-OLA (SOLA) method. In the frequency domain, the phase vocoder uses a spectral representation of the frames to minimize the phase cancellation issues and allow efficient spectral manipulations of the sound [LD99a, Por76, FG66]. The Shape-Invariant Phase vocoder (SHIP) has been proposed to improve the phase synchronization using only the sinusoidal component of the frames [Roe10]. These methods are known to achieve high quality time stretching. Additionally, a pitch transposition effect can be obtained by resampling a time stretched utterance. When transposing pitch upward, the harmonic content at low frequencies is pushed up toward higher frequencies which leads to a buzzy sound. However, in that case the phase can be randomized to reduce this side effect [Roe10]. When transposing pitch downward, the noise produced in high frequencies arises at low frequencies where such a noise is naturally not present. Therefore, using downward transposition, the phase vocoder is known to increase the hoarseness. Whereas the phase spectrum can be randomized to create noise from sinusoidal contents, the inverse remains a tricky problem. Note that, the phase vocoder is not dedicated to voice only and can find applications in polyphonic recordings such as music.

In order to minimize the phase cancellation issues, analysis windows of two periods duration can be used and centered on estimated GCIs, leading to the well known Pitch-Synchronous OLA method (PSOLA) [VMT92]. Note that, contrarily to the phase vocoder, PSOLA requires an estimate of the fundamental frequency and an estimation of the GCIs. Therefore, this method is restricted to monophonic recordings but especially efficient for voice processing. Because the impulse response of the VTF is forced to decay by the windows which are particularly short, the most well known drawback of this method is the lack of resonances in downward pitch transpositions. However, due to its simplicity (no interpretation of the spectral properties of the windowed signal), PSOLA is one of the most robust and artifact-free methods.

Analysis/Synthesis methods

- Sinusoidal and harmonic models [Sty96, MQ86]

  The peaks of an observed amplitude spectrum can be modeled using sinusoids. Speech can thus be parametrized in terms of amplitude, frequency and phase of a set of sinusoids [MQ86]. Moreover, if the waveform is assumed to be periodic, this set can be reduced to harmonically related frequencies

[Sty96]. In such an approach, the noise component can be modeled using an amplitude modulated Gaussian noise convolved by an AR envelope, leading to the Harmonic+Noise Model (HNM) [LSM93]. The noise envelope in the frequency domain can be also segmented in multiple frequency bands with an explicit control of the harmonic to noise ratio in each band [GL88]. Since the spectral components of the voice are never perfectly stationary, a Quasi-Harmonic Model (QHM) has been proposed to take into account the time variations [PRS08, Sty96].

- Wide-Band Voice Pulse Modeling (WBVPM) [Bon08]

  In this method, windows centered on GCIs are used like in the PSOLA method. Then, a spectral representation of each frame is modeled using a wide-band spectrum. An envelope of the amplitude spectrum is estimated by splines like in the SEEVOC method [Pau81], and the phase spectrum is modeled by a smooth envelope and a delay [Bon08]. This method models the spectral content of each frame as an impulse whereas the sinusoidal models represent exactly the same content by means of sinusoids.

- Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [KMKd99]

  Instead of estimating the spectral envelope like done in vocoder approaches, the STRAIGHT method compute the Fourier transform of a frame using a particular window function in order to retrieve a smooth interpolation of the spectral peaks in both frequency and time domain. This way of estimating the envelope of the amplitude spectrum is thus very robust since no convergence issues can appear like in iteration procedures (e.g. TE, DAP).

  Initially, the phase spectrum of voiced parts of a speech utterance are synthesized using the minimum-phase relation with the amplitude spectrum [Kaw97]. To model unvoiced parts, an all-pass filter is used to randomize the phase spectrum above a given frequency (e.g. zero in fricatives and $\approx 2\,\text{kHz}$ in voiced segments). Then, in order to deal with the mixed-phase property of the speech signal, Banno et al. proposed to extend the STRAIGHT method with a phase model using a Time-domain Smoothed Group Delay (TSGD) [BLN+98].

- ARX/ARMAX

  Many analysis/synthesis methods that take exogenous input have been proposed to manipulate the voice [AR08, VRC07, Lu02, Lju86, Hed84]. The deterministic source is synthesized using a certain glottal model and the residual of this deterministic source is modeled either using modulated and colored noise [AR08] or HNM [AR09, VRC07].

  Compared to the above methods, the implementation of the ARX/ARMAX methods is far from straightforward. Moreover, they depend on a reliable estimation of the glottal model parameters and are sensitive to inversion errors [AR08].

For the evaluation of the method proposed in this study, the SHIP, PSOLA and STRAIGHT methods will be used in the comparisons because of their availability.

9.2 Choice of approach for the proposed method

Analysis/synthesis methods can be used for both voice transformation and HMM-based speech synthesis. Conversely, modification methods do not fully model voice production and thus cannot be applied to HMM-based speech synthesis. More generally, one may expect more applications of analysis/synthesis methods. Therefore, during this research, we developed an analysis/synthesis method which will be presented in the following sections.

Additionally, as in ARX/ARMAX methods, a glottal model will be used to model the deterministic component of the glottal source. As argued in the introduction, we assume that the efficiency of an analysis/synthesis method can be improved using a glottal model, that is dedicated to voice production, compared to a method which does not take into account the particular properties of the underlying signals of voice production. Indeed, although analysis/synthesis methods are usually designed for voice modeling, most of these methods can be applied to any pseudo-periodic signal (e.g. STRAIGHT, HNM, WBVPM). For example, as shown in chapter 2, the glottal source has a spectrum which should be taken into account during the separation process conversely to most of the current analysis/synthesis methods which assume that the excitation source is made of a flat amplitude spectrum. In the proposed method, the LFRd pulse models the deterministic component of the glottal source and Gaussian noise represents the random component. Then, in the analysis step, we estimate the VTF by taking into account this source model to fit an observed speech spectrum. This method is called *Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise* (SVLN).

The next section presents the separation method: the estimation of the parameters of the source model and the parameters of the VTF from an observed spectrum. Then, presentation of the synthesis method follows.

9.3 The analysis step: estimation of the SVLN parameters

In this section, conversely to part II *Analysis* which treated the deterministic component, both the deterministic and the random frequency bands will be considered in the following. Below, equation 9.1 recall the voice production model depicted in section 3.3 using parameters of the SVLN method:

$$S(\omega) = \left[e^{j\omega\phi} \cdot H^{f_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)\right] \cdot C^{\bar{c}}(\omega) \cdot L(\omega) \tag{9.1}$$

where $G^{Rd}(\omega)$ is the LFRd glottal model, $N^{\sigma_g}(\omega)$ is a spectrum of a zero-mean Gaussian noise of standard-deviation σ_g and $\bar{c}$ are the cepstral coefficients which parametrize the VTF. In the following sections, these parameters are estimated at regular intervals of 2.5 ms along a given speech utterance. At each analysis instant, the acoustic signal is assumed to be stationary in an analysis window of 3.5 periods duration in voiced parts (with a minimum of 10 ms) and 15 ms in unvoiced parts.

9.3.1 The parameters of the deterministic source: f_0, Rd, E_e

As shown in section 2.5, the fundamental frequency f_0 can be estimated from numerous methods. For this presentation, the YIN method [dK02] is used. To estimate the LF shape parameter Rd, the method based on MSPD^2 is used because of its efficiency demonstrated in chapter 8.

Three gains co-exist in the voice production model: E_e, σ_g and the mean log amplitude of the VTF. These gains are completely dependent on each other. Indeed, if E_e and σ_g are multiplied by some

arbitrary value α, the VTF mean log amplitude may compensate α leading to the same gain of the observed spectrum (with $-log(\alpha)$). Consequently, a constraint is necessary. In this presentation, the mean log amplitude of the VTF is fixed to zero. The energy variation of the speech signal is thus only dependent on the energy of the source, the noise level σ_g and the excitation amplitude of the glottal model E_e.

9.3.2 The parameter of the random source: σ_g

According to figure 3.3, a Voiced/Unvoiced Frequency (VUF) can be estimated to split $S(\omega)$ into a deterministic source below the VUF and Gaussian noise above. Like the fundamental frequency, this frequency is assumed to be known *a priori* (see sec.3.3). In the following, the amplitude spectrum $|G^{Rd}(\omega)|$ is therefore assumed to cross the expected amplitude of the noise at the VUF (see left plot of figure 3.3). Consequently, since the amplitude spectrum $|G^{Rd}(\omega)|$ is known when the f_0 and Rd estimates are given, the noise level σ_g can be deduced from the VUF estimate:

$$\sigma_g = |G^{Rd}(\text{VUF})| \cdot \frac{\sqrt{2}}{\sqrt{\pi/2} \cdot \sqrt{\sum_t \text{win}[t]^2}}$$

where $|G^{Rd}(VUF)|$ is the expected amplitude of the LF model at the VUF which has to be converted to the Gaussian parameter σ_g: Spectral amplitudes of Gaussian noise obey a Rayleigh distribution. Therefore, $|G^{Rd}(VUF)|$ is first converted to the Rayleigh mode ($1/\sqrt{\pi/2}$), then the standard deviation of the Gaussian distribution in the time domain is retrieved from the Rayleigh mode ($\sqrt{2}$) [Yeh08]. Additionally, in the spectral domain, the noise level is proportional to the energy of the analysis window win$[t]$ used to compute $S(\omega)$. The normalization by $\sqrt{\sum_t \text{win}[t]^2}$ is therefore necessary.

9.3.3 The estimation of the vocal-tract filter

According to the difference of the underlying excitation properties, the frequency bands above and below the VUF are modeled using two different envelopes. These envelopes are aligned as follows to ensure a VTF estimate which is independent of the nature of the excitation.

In the deterministic band, where $\omega <$VUF, the contribution of the radiation $L(\omega)$ and the deterministic source $G^{Rd}(\omega)$ are removed from $S(\omega)$ by division in frequency domain. The TE cepstral envelope $\mathcal{T}(.)$ is then used to fit the top of the harmonics of the division result. Note that this envelope fits the expected amplitude of the VTF frequency response since the top of a harmonic is its expected amplitude.

In the random band, where $\omega >$VUF, $S(\omega)$ is divided by $L(\omega)$ and by the crossing value $|G^{Rd}(VUF)|$ to ensure a continuity between the two frequency bands. The division result is modeled by computing its real cepstrum $\mathcal{P}(.)$ truncated to a given order (discussed below). According to the Rayleigh distribution of the spectral amplitudes of this band, the mean log amplitude measured by $\mathcal{P}(.)$ has first to be converted to the Rayleigh mode on a linear scale (factor $e^{0.058}$ in the eq. 9.2 below) [Yeh08]. Then, the expected amplitude is retrieved from the Rayleigh mean value ($\sqrt{\pi/2}$).

Given these two envelope estimates, the estimation of the VTF can be summarized by:

$$C(\omega) = \begin{cases} \mathcal{T}\left(\frac{S(\omega)}{L(\omega)G^{Rd}(\omega)}\right) \cdot \gamma^{-1} & \text{if } \omega < VUF \\ \\ \mathcal{P}\left(\frac{S(\omega)}{L(\omega)G^{Rd}(VUF)}\right) \cdot \frac{\sqrt{\pi/2}}{\gamma \cdot e^{0.058}} & \text{if } \omega \geq VUF \end{cases} \tag{9.2}$$

where $\gamma = \sum_t \text{win}[t]/(f_s/f_0)$ stands for the number of periods in the analysis window. This normalization is necessary regarding to the synthesis step where the VTF is convolved with each period of the source. The gain of the estimated VTF has thus to be normalized according to the shape and the duration of the analysis window.

It is also necessary that the two envelopes $\mathcal{T}(.)$ and $\mathcal{P}(.)$ do not fit the harmonic structure of the observed spectrum $S(\omega)$. For the TE envelope, the optimal order $0.5 \cdot f_s/f_0$ is used. The same order is used for the cepstral envelope. Indeed, although no harmonic partial appears in the frequency band of the random source, sinusoidal peaks with distance of f_0 (but not of multiples of f_0) arise in this band because the glottal noise is amplitude modulated by the glottal area (see sec. 2.6). A last technical detail has to be taken into account. The division by $L(0) = 0$ has to be avoided in equation 9.2. $L(0)$ can be either extrapolated from $L(\omega)$ or $j\omega$ can be replaced by $1 - \mu e^{j\omega}$ with μ close to unity.

Finally, the cepstral coefficients $\bar{c}$ of the VTF are retrieved from the minimum-phase cepstrum of $C(\omega)$ to represent the VTF frequency response with a small set of parameters.

Note that this separation method is always able to model the observed amplitude spectrum $|S(\omega)|$. Indeed, the estimation of the VTF always completes the source and radiation models in order to obtain $|S(\omega)|$. Conversely, the phase of the model is imposed either by the LF model or the Gaussian noise.

Unvoiced segments

In unvoiced segments (fricatives, plosives, silence, etc.), there are no glottal pulses. Therefore, when the VUF estimate is lower than f_0, the VUF is clipped to zero. Note that in such a case, the estimation of the VTF reduces to the cepstral envelope $\mathcal{P}(.)$.

9.4 The synthesis step using SVLN parameters

This section describes the procedure which synthesizes a speech utterance from the estimated parameters of the previous section. Roughly speaking, this synthesis procedure is an overlap-add technique. Small segments of stationary signals are synthesized and theses segments are then overlap-added to construct the whole signal. The following sections first define what a segment is. Then, depicted in figure 9.2, the synthesis of the content of each segment is described.

9.4.1 Segment position and duration

In voiced parts, where the excitation amplitude E_e exceeds a given threshold, temporal marks m_k of the kth-segment are placed at intervals according to the fundamental period $1/f_0$ (see fig. 9.1). The maximum excitation instant t_e of each LFRd pulse is placed on this mark. Then the starting time t_k of the kth-segment is defined as the opening instant t_s of the LFRd model. Finally, the ending time of this segment is the starting time of the next. In unvoiced parts, a segment has a 5 ms duration, and its mark m_k is placed in the center, as illustrated in figure 9.1.

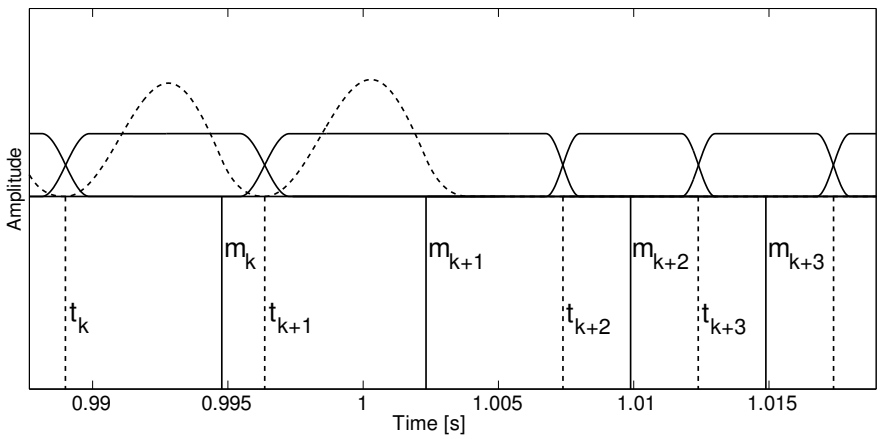


Figure 9.1: Two voiced segments followed by two unvoiced segments: Marks m_k and starting times t_k are shown with vertical lines. Synthesized LFRd models are in dashed lines, and windows win$_k[t]$ are in solid lines.

9.4.2 The noise component: filtering, modulation and windowing

For all segments, noise is generated. However, if this noise is white, the synthesized voice sounds hoarse because the lowest harmonics of the deterministic source are disturbed by this noise. Additionally, if the glottal noise is not amplitude modulated synchronously with the fundamental period, a second source is perceived separately from the deterministic source [AR09, MQ05, SK00, Her91]. Consequently, to improve the naturalness of the synthesized noise, the two following processes are used.

Firstly, the glottal noise is colored, lowest frequencies are weaker than higher frequencies [Ste71]. In this procedure, the noise is thus filtered with a high-pass filter $F_{hp}^{VUF}(\omega)$ defined by a cutoff frequency equal to the VUF and a slope of 6 dB/kHz in the transition band (since the VUF is not part of the analysis parameters, this value is obtained from the intersection between the noise level and the amplitude spectrum of the glottal model).

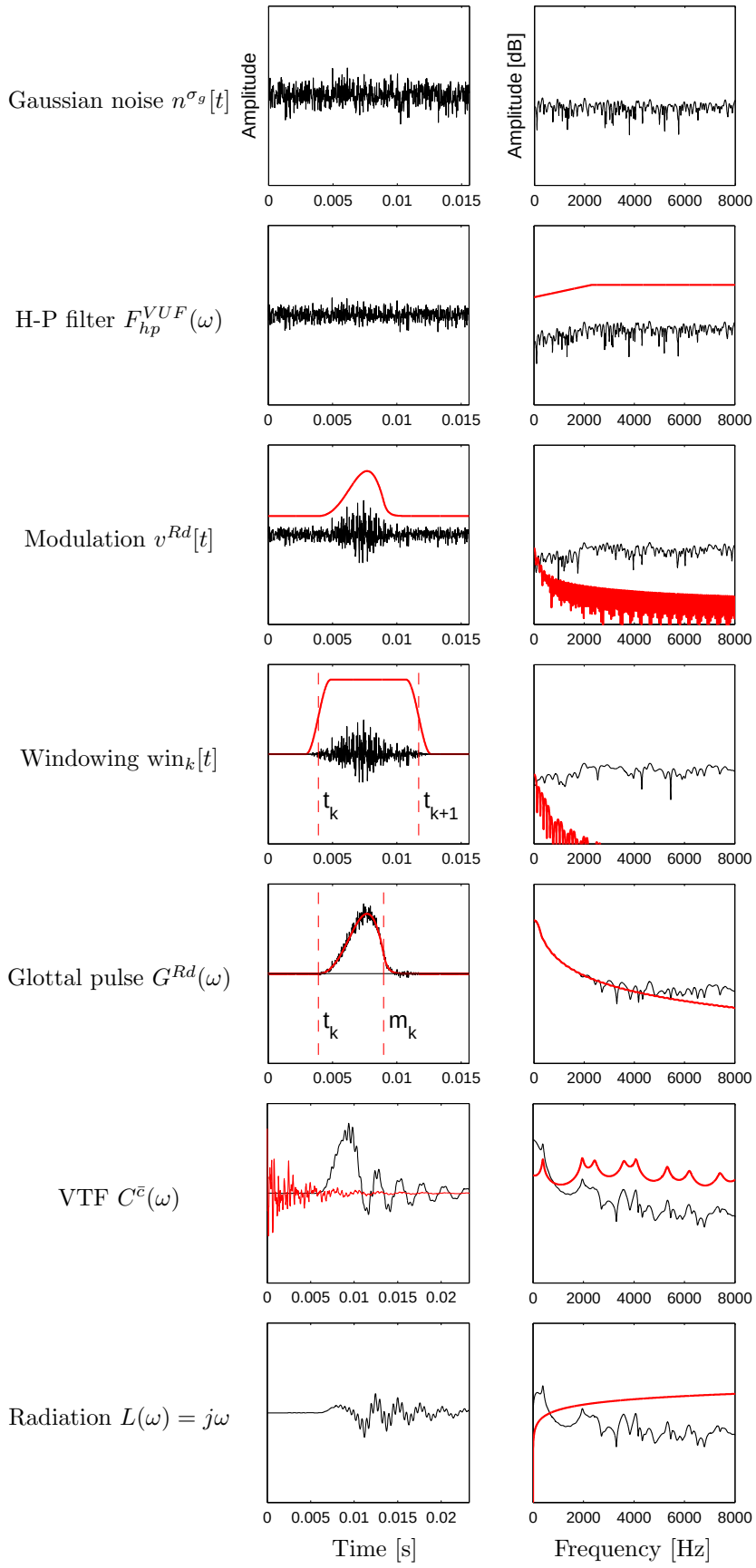


Figure 9.2: Synthesis of a segment (For the clarity of this sequence of plots, the difference of level between the deterministic and random components may not correspond to the descriptions given by the text)

Secondly, as shown in section 2.6, the amplitude of the glottal noise depends on the glottal flow and the glottal area. Accordingly, the glottal noise is amplitude modulated as proposed in [dY08] and [MQ05]. This modulation $v^{Rd}[t]$ is built from the LFRd glottal model as follows:

$$v^{Rd}[t] = \beta \cdot g^{Rd}[t] + (1 - \beta)$$

where $g^{Rd}[t]$ is the pulse of the LFRd glottal model with voicing amplitude $A_v = 1$ and β a constant. The estimation of the parameters of the noise modulation (Rd and β) has not been studied in this work. The Rd parameter here is set to the same as the one of the deterministic source. Then, from subjective listening of 10 different voices and their corresponding resynthesis, we fixed the value $\beta = 0.75$ according to the naturalness of the resynthesis. Obviously, if these two values Rd and β were properly estimated from the observed signal, the naturalness of the synthesized noise component would be improved [MQ05].

Finally and conversely to the glottal pulses, the noise does not stop at zero amplitude at the end of each segment. Therefore, a cross fade is necessary between noise segments of different color and amplitude. For each kth-segment, a window $\text{win}_k[t]$ is built with a fade-in center on t_k and a fade-out center on t_{k+1} (see figure 9.1). The fade-in/out function is a hanning half window of duration $0.25 \cdot \min(t_{k+1} - t_k, t_k - t_{k-1})$. Additionally, the fade-out of win_k is the complementary of the fade-in of win_{k+1}. Consequently, the sum of all windows is 1 at any time of the synthesized utterance. Once the synthesized segments are overlap-added, it is not necessary to normalize the result by the sum of the windows. Note that no window is necessary to cross fade the glottal pulses since they start and end at zero amplitude.

In voiced segments, according to the discussion above and the first four rows of figure 9.2, the noise spectrum of the kth segment is synthesized by:

$$N_k(\omega) = \mathcal{F}\left(\text{win}_k[t] \cdot v^{Rd_k}[t] \cdot \mathcal{F}^{-1}\left(F_{hp}^{\text{VUF}_k}(\omega) \cdot N^{\sigma_{gk}}(\omega)\right)\right) \tag{9.3}$$

where $N^{\sigma_{gk}}(\omega)$ is the spectrum of zero-mean Gaussian random signal $n^{\sigma_{gk}}[t]$. In unvoiced segments, the noise source reduces to:

$$N_k(\omega) = \mathcal{F}\big(\text{win}_k[t] \cdot n^{\sigma_{gk}}[t]\big) \tag{9.4}$$

9.4.3 Glottal pulse and filtering elements

Finally, the deterministic source $G^{Rd_k}(\omega)$ is added to the noise component (5th row of figure 9.2) and the VTF and radiation filters are applied to the source (6th and last row of figure 9.2):

$$S_k(\omega) = \left(e^{-j\omega m_k} \cdot G^{Rd_k}(\omega) + N_k(\omega)\right) \cdot C^{\bar{c}_k}(\omega) \cdot j\omega \tag{9.5}$$

where $e^{-j\omega m_k}$ is a delay placing the instant t_e of the LF model at the mark m_k and $C^{\bar{c}_k}(\omega)$ is the minimum-phase VTF corresponding to the cepstral coefficients $\bar{c}_k$.

Finally, the time domain sequence of each segment is retrieved through the inverse Fourier transform of $S_k(\omega)$. Then, the entire signal is constructed by successively overlap-adding the time segments. Note that in such a synthesis process, a window of only one period's duration is used to synthesize the noise component. The averaging effect between neighbor frames known in usual overlap-add process is thus greatly reduced. Indeed, in this method, both source and VTF are completely free to change from one period to the next without any influence of the neighboring periods. Sound examples `snd_*.resynthesis.wav` show results of analysis/synthesis using the proposed SVLN method. In the context of pitch transposition and speech synthesis, this method will be evaluated in the next chapter.

Conclusions

- This chapter presented the SVLN method (Separation of Vocal-tract and Liljencrants-Fant model plus Noise) which aims to separate a given voice recording into four parts: a deterministic source, a random source, a VTF and a radiation filter. Whereas most of the existing techniques can be applied to any pseudo-periodic signal (e.g. vocoders, PSOLA, STRAIGHT), the proposed method is dedicated to voice processing, because it separates the different components of the voice production using a glottal model (like ARX and ARMAX methods do).

- Using the LF^{Rd} model and Gaussian noise, the SVLN method models the glottal source with only four parameters (f_0, Rd, E_e, σ_g). Such a reduced number of parameters is thus very interesting in connection with machine-learning methods like used in HMM-based synthesis.

- It has been shown that the VTF can be estimated by taking into account the radiation, the amplitude spectrum of the glottal model and the different types of excitations (deterministic and random). Moreover, the estimation of the VTF complete the excitation model in order to fit the observed spectrum. Therefore, one can expect to always recover this amplitude spectrum perfectly whereas a difference exists between the observed phase spectrum and the phase spectrum of the model. Indeed, at low frequencies, the phase spectrum of the deterministic source is determined by the LF^{Rd} model while at high frequencies, the phase spectrum is almost fully random (the amplitude modulation of the noise creates some structures in the phase spectrum).

- In order to improve the naturalness of the random component of the source regarding to Gaussian noise used in this method, two known techniques are used: amplitude modulation and high-pass filtering.

Chapter 10

Evaluation of the SVLN method

In this chapter, the SVLN analysis/synthesis method described in the previous chapter is evaluated. First, some theoretical elements are discussed which should help to understand the behavior of the SVLN method related to the estimation of its parameters. Then, the results of three preference tests are presented in order to evaluate the capabilities of this method in pitch transposition. The control of the breathiness with the proposed SVLN method is evaluated with a fourth test and the results from a last test allows to evaluate the quality of an HMM-based synthesis using the proposed method.

10.1 Influence of the estimated parameters on the SVLN results

The amplitude spectrum of the observed speech signal is always reconstructed by the SVLN method (see section 9.3.3). The phase spectrum, however, can be modeled only by the LF^{Rd} model, Gaussian noise and the minimum-phase of the VTF. Therefore, in the context of encoding and decoding of the voice (without transformation), a bias of the *Rd* value implies an error of resynthesis of the phase spectrum only. In terms of stability of the modeling of the voice, this robustness related to the shape parameter is necessary according to the risk incurred by the estimation of sophisticated parameters. Even though a bias is assumed to have no consequences, the shape parameter has to be stable. According to the irregularities observed in real signals (see sec. 8.3), it is important to smooth the estimated values to ensure that the separation made by the SVLN method will be stable between adjacent analysis frames. In the context of voice transformation, biased estimates of the *Rd* parameter implies that the transformation capabilities of the SVLN method are limited by the *Rd* parameter being clipped. In addition, if the estimation of the VTF (eq. 9.2) does not properly filter out the amplitude spectrum of the real glottal pulse, one can expect that residual spectral shapes remain in the transformed voice which tends to generate some artifacts. Consequently, in order to smooth the time evolution of the glottal shape parameter and to stabilize the VTF estimate, the estimated *Rd* curves are first filtered using a median filter with a window duration of 100 ms, then a zero-phase filter is used to smooth the steps made by the median filtering. The latter filter uses a hanning window of the same duration of the median filter. In doing so, inside a single phoneme, we assume that the voice quality is constant.

The estimation of the VUF has also an impact on a synthesized voice. If the VUF is underestimated, noise is generated at low frequencies, and the synthesized voice sounds hoarse. Conversely, if the VUF is overestimated, the voice sounds buzzy because regular impulses are generated at high frequencies by the harmonic structure. Additionally, the voicing decision in the time domain is critical for a proper

reconstruction of the transients. For example, if a plosive is defined voiced by the analysis step, the excitation energy of the VTF at low frequencies will be generated by the LF model which would create a bubble-like artifact.

10.2 Preference tests for pitch transposition

A preliminary study has been carried out through three preference tests in order to evaluate to which extent the SVLN method can be used to transpose the pitch of a given utterance. The proposed method has been compared with the Shape-Invariant Phase vocoder (SHIP) [Roe10][1], the PSOLA method [Pee01, HMC89][2] and the STRAIGHT method [KMKd99][3]. The f_0 and VUF estimates were common to all compared methods. If necessary, the estimation errors of the f_0 values related to octaves errors has been corrected manually. Additionally, to avoid an influence of the voicing decision in the time domain, this parameter has been manually annotated.

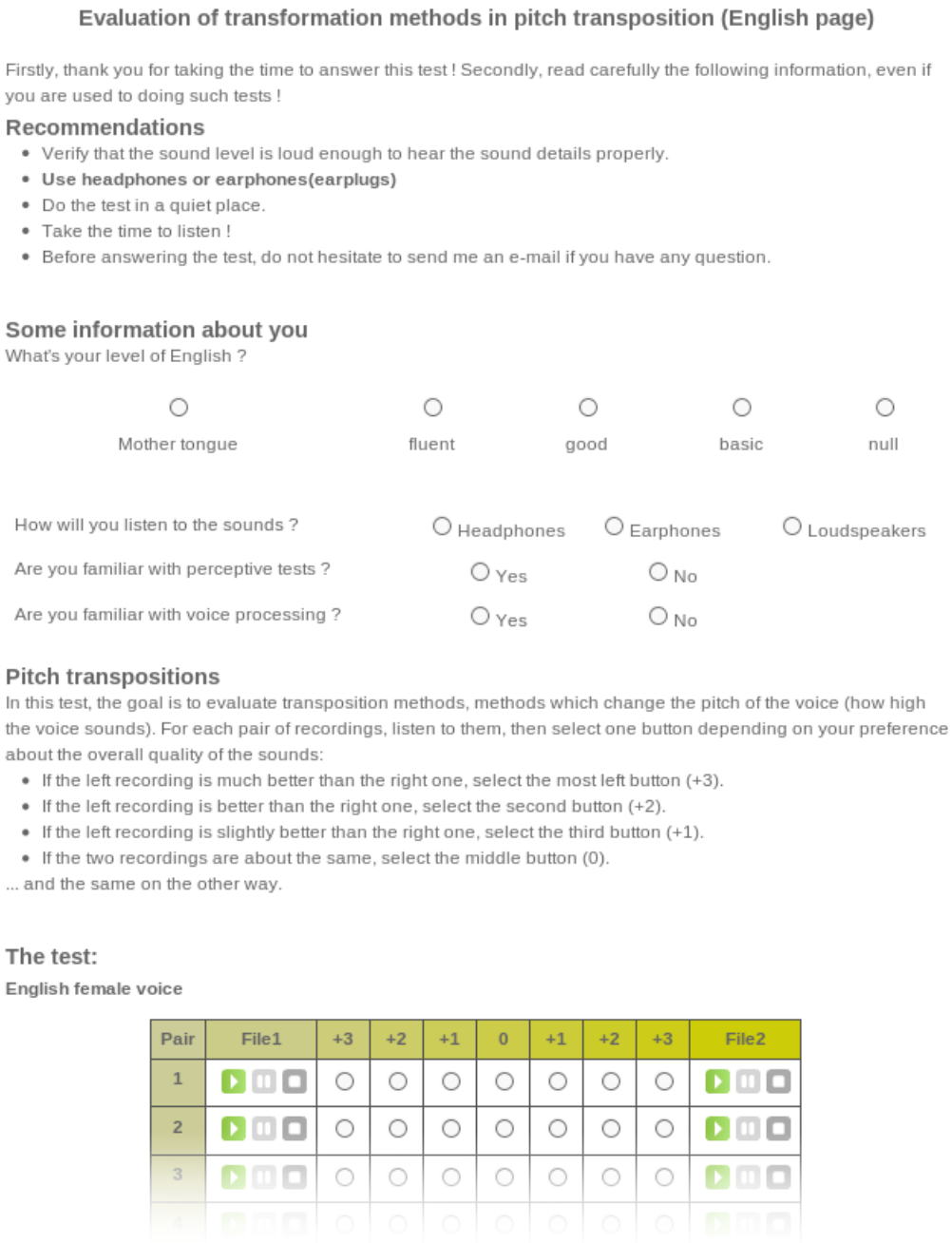
Concerning the SVLN method, the used separation method takes into account the amplitude spectrum of the source in the VTF estimation. Consequently, if the parameters controlling the source are left unmodified in pitch transposition, the glottal formant will be shifted proportionally to the transposition of the fundamental frequency. However, the voice quality is known to be correlated to f_0 [TM03, Hen01]. The higher the pitch, the more lax the source and thus the bigger the Rd value. Accordingly, in order to obtain a natural voice in pitch transposition, the glottal formant and the fundamental frequency should not be equally shifted. In these preference tests, the Rd parameter of the synthesized voice was modified according to the following formula: $Rd' = 2^{\beta \cdot T/1200} \cdot Rd$, where T is the transposition factor given in cents and β is a proportionality constant that controls the coefficient of the modified Rd parameter according to the transposition factor. For these preference tests, we chose a fixed β value from the following informal listening. The utterances used in these tests were transposed with $T = \pm 900$ cents and β values between 0.1 and 2. According to the naturalness of these transpositions we kept the transpositions with $\beta = 0.5$.

Note that in preference tests the listeners are often asked to focus on a precise characteristic of the sound quality (e.g. naturalness, artifacts, intelligibility). Furthermore, the choice of this characteristic depends on the purpose of the evaluated methods (e.g. naturalness for voice transformation methods, artifacts for encoding/decoding methods). However, if the listener is asked to evaluate only the naturalness, it is difficult to assess to which extent the compared methods can be used in real applications regarding the presence of artifacts. In the tests of this evaluation, in order to obtain a mean opinion considering all potential way of evaluation, the listeners were asked to give "their preference about the overall quality of the sounds" of each comparison pair which was produced by two different methods (see fig. 10.1). Each listener was also asked whether he/she used headphones, earphones or loudspeakers. To ensure a minimal quality of listening condition, the results of those who used loudspeakers have been set aside.

Resynthesis of unvoiced segments

First, the influence of the unvoiced segments on the overall quality of the methods was evaluated. Therefore, two tests were carried out, one with the unvoiced segments resynthesized by each method and the other one with the unvoiced segments taken from the original recording. Each test was made with English and French recordings and a dedicated page was used for each language (see English page in figure 10.1).

[1] Using the *Super Vocoder de Phase (SVP)* developed internally at Ircam.
[2] Using a MATLAB implementation developed internally at Ircam.
[3] Freely available from `http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv`, here, the used version is `V40pcode`



Evaluation of transformation methods in pitch transposition (English page)

Firstly, thank you for taking the time to answer this test ! Secondly, read carefully the following information, even if you are used to doing such tests !

Recommendations

- Verify that the sound level is loud enough to hear the sound details properly.
- **Use headphones or earphones(earplugs)**
- Do the test in a quiet place.
- Take the time to listen !
- Before answering the test, do not hesitate to send me an e-mail if you have any question.

Some information about you

What's your level of English ?

Mother tongue | fluent | good | basic | null

How will you listen to the sounds ? Headphones | Earphones | Loudspeakers

Are you familiar with perceptive tests ? Yes | No

Are you familiar with voice processing ? Yes | No

Pitch transpositions

In this test, the goal is to evaluate transposition methods, methods which change the pitch of the voice (how high the voice sounds). For each pair of recordings, listen to them, then select one button depending on your preference about the overall quality of the sounds:

- If the left recording is much better than the right one, select the most left button (+3).
- If the left recording is better than the right one, select the second button (+2).
- If the left recording is slightly better than the right one, select the third button (+1).
- If the two recordings are about the same, select the middle button (0).

... and the same on the other way.

The test:

English female voice

Pair	File1	+3	+2	+1	0	+1	+2	+3	File2
1									
2									
3									
4									

Figure 10.1: The page of a preference test submitted to the participants.

For each page, the fundamental frequency of two utterances (one male and one female) was transposed both upward and downward using $T = \pm 900$ cents using the compared methods. Thus, for each utterance, 12 comparison pairs were obtained by comparing each method to the three other methods (6 methods comparison for 2 transpositions). Each test page dedicated to one language is thus made of 24 comparison pairs to evaluate by the listener. The left/right order of the pairs was randomized in order to avoid one method from always being on the same side of the test and the order of the pairs was shuffled in order to not always evaluate the methods in the same order between the various voices and languages. Finally, for each method, the preference scores are gathered from all the evaluated pairs ($+N$ for each grade advantaging the method, and $-N$ for each grade penalizing the method). The mean preference score can thus be computed for each method which is then compared to the one of the other methods as shown in the presented results below.

In such preference tests, the number of participants has to be sufficient in order to obtain significant results regarding to the confidence intervals of the mean preferences and the variance of the answers. According to the necessary number of answers to discuss the disparities below (using the original unvoiced segments), the number of participant for the second test had to be larger than for the first one. Therefore, once significant differences were observed from the first test, this test was closed in order to ask the participants to answer the second test. Consequently, 8 participants answered the first test in English and 4 answered in French. For the second test, 25 participants answered to each English and French pages. Results are shown in figure 10.2. One can see that the resynthesized unvoiced segments penalize clearly the SVNL method. Indeed, significant artifacts can be perceived in transients. Conversely, using the original signal in the unvoiced segments, the SVLN method is comparable to the other methods. Moreover, the STRAIGHT method outperforms the other compared methods and the quality of the SVLN method is between that of STRAIGHT and those of SHIP and PSOLA. In addition, one can note that the preference for the SHIP method is lower than that of SVLN and STRAIGHT in downward transpositions.

In order to focus on the capability of the methods to transpose the pitch, the next presented results are related to the second test, the one where the original recording was used in the unvoiced segments.

Results disparity among voices

One can see that the disparity of the results can be important between the voices. For example, figure 10.3 shows the results of the male voices. Between the two voices, one can see that the preferences of the STRAIGHT and SVLN methods are opposite in upward transpositions. Moreover, the STRAIGHT method is mainly penalized by the English male voice whereas this method clearly outperforms the other methods for the upward transposition of the French voice. Finally, the SVLN method shows no clear differences with the SHIP method for the French male voice and clearly outperforms this method for the downward transposition of the English male voice. The results are thus strongly dependent on the used voices. According to this preliminary evaluation, no method outperforms all the others for all voices, although one can observe mean trends of preferences.

Results disparity among transposition factors

A third similar test has been carried out where the utterances of French and English voices (both male and female) was transposed one octave below and one octave above the original pitch (i.e. ± 1200 cents). Note that the PSOLA method was removed in this test in order to obtain less comparison pairs to evaluate. 26 and 14 participants answered the test in English and French respectively. The results of this test are shown in figure 10.4.

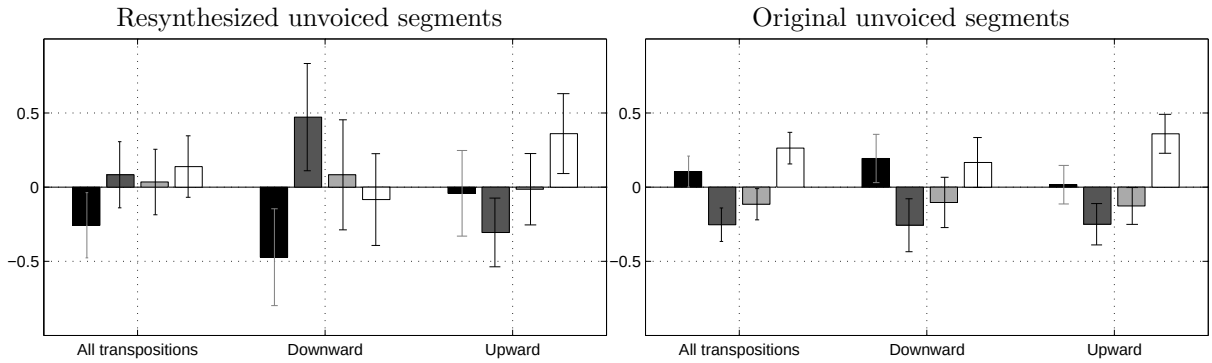


Figure 10.2: Difference in preferences using the original or the resynthesized unvoiced segments. The mean preference is shown with its 95% confidence interval.

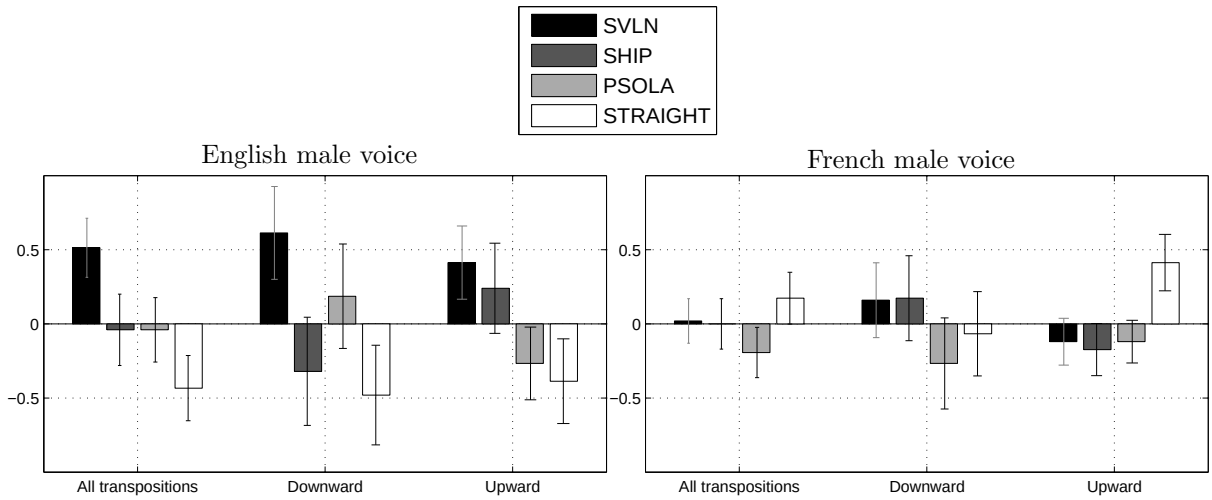


Figure 10.3: Disparity between the two male voices.

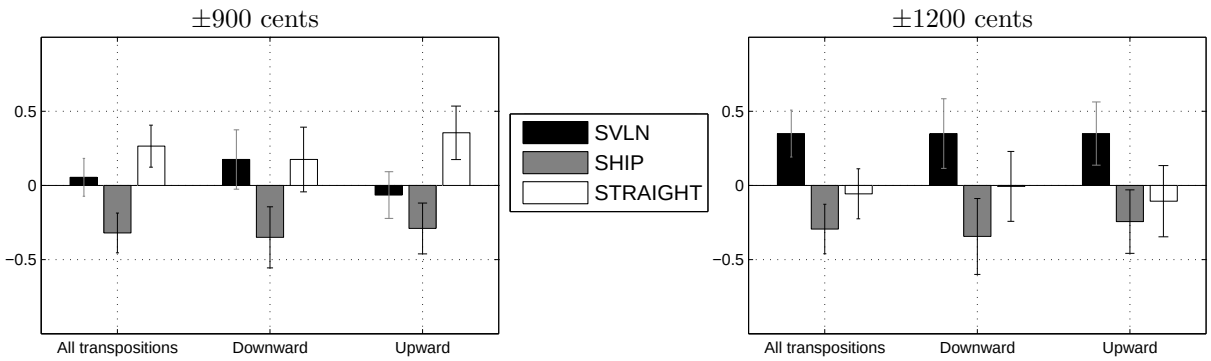


Figure 10.4: Disparity of preferences among transposition factors.

One can see that the results of the SVLN method are improved for transpositions of ±1200 cents compared to ±900 cents. According to informal listening, the quality of the SHIP and STRAIGHT methods seems to reduce when the absolute value of the transposition factor increases. Conversely, the SVLN method produces a glottal source which is consistent with the LF^{Rd} glottal model and the measured noise level σ_g. Thus, whatever the transposition factor, one can expect that the synthesized source by SVLN always respects some constraint imposed by the LF model which is important for the naturalness of the voice. This observation can explain the results disparity among the transposition factors.

10.3 Evaluation of breathiness modification

A last test was carried out to evaluate the capability of the SVLN method to modify the breathiness of a given recording. The breathiness implies a random component in the glottal source which is a priori not related to the *Rd* shape parameter. However, using the SVLN method, a modification of the *Rd* parameter changes the perception of the random source. Indeed, by increasing *Rd* the VUF is decreased since this shape parameter control also the spectral tilt of the glottal pulse. A frequency band, previously excited by harmonics, can be so made of noise. Conversely, by reducing *Rd*, noise excitation can be masked by increasing harmonics. Through the chosen voice production model, the breathiness of a voice can be therefore modified by a modification of the unique parameter *Rd*.

In the following "preference" test, similar to the previous tests, the listeners were asked to compare two utterances of different breathiness obtained by a modification of the *Rd* parameter by the SVLN method. Like in figure 10.1, a grade was selected by the listeners according to the comparison of the pair: "+3 if the left recording is much breathier than the right one; +2 if the left recording is breathier than the right one; +1 if the left recording is slightly breathier than the right one; 0 If the two recordings are about the same or if a difference exists which is, from the point of view of the listener, not related to breathiness; and the same on the other side of the comparison grid". The original recordings and four different transformations were compared. The latter were obtained by multiplying the *Rd* parameter by four different power of 2: $2^{-1} = 0.5$; $2^{-1/2} \approx 0.71$; $2^{1/2} \approx 1.41$ and 2. The test was proposed in the same two languages, English and French, using one utterance of a male voice only. "Mean breathiness scores" are computed like the mean preference score discussed above.

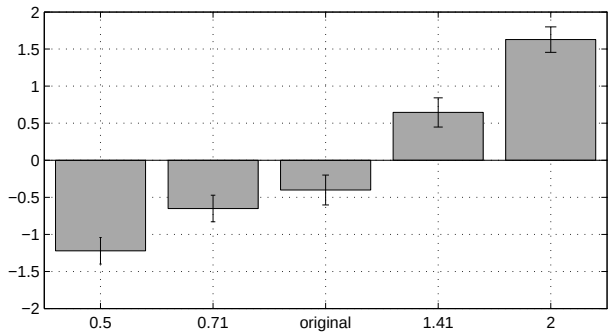


Figure 10.5: Mean breathiness scores for various scaling of *Rd* and the 95% confidence interval.

In conclusion, regarding to figure 10.5, the SVLN is clearly able to modify the breathiness of the voiced segments. It is also interesting to note that the maximum value of breathiness (factor 2) is close to be 50% times bigger than the minimum value of breathiness (factor 0.5). It seems thus easier to increase the breathiness than reduce it.

10.4 Speech synthesis based on Hidden Markov Models (HMM)

By means of machine learning, the parameters of an analysis/synthesis method estimated on a database of acoustic recordings and their corresponding textual transcriptions are used to train a set of contextual HMMs[4]. To synthesize a given text, a corresponding sequence of parameters can then be generated using the trained HMMs. Finally, the speech waveform corresponding to this parameter sequence is synthesized. In this section, we present the results of a preference test which evaluated the efficiency of three analysis/synthesis methods: SVLN, STRAIGHT and a basic method using impulses train or Gaussian noise for the voice source and mel-cepstral coefficients to model the VTF [ZNY+07] (called *impulse-source* method in the following).

Using the SVLN method, the parameters of the source f_0, E_e, Rd, σ_g and the cepstral coefficients of the VTF are first estimated according to the methods described in section 9.3. Then, in order to reduce the number of parameter in the learning procedure, the excitation amplitude E_e is merged into the first cepstral coefficient controlling the overall gain of the synthesized signal. In order to keep the relative level between the deterministic and random sources, the gain of the random source σ_g is thus also normalized by E_e. Finally, the cepstral coefficients are encoded using a mel scale and the fundamental frequency is used as voiced/unvoiced indicator as shown below. In this test, the implementation of the HMM-based synthesis was based on the HTS Toolkit [ZNY+07]. The parameters were split into independent streams. Several stream configurations were tested and the following one was considered the optimal according to informal listening:

- One single Gaussian distribution with semi-tied covariance [Gal99] for $\{Rd, \sigma_g, \bar{c}\}$;
- One multi-space distribution [TMMK02] for f_0

Both streams include first and second time derivatives of their parameters.

In order to obtain a consistent prosody and co-articulation of a synthesized utterance, it is necessary to take into account the context of each phoneme. Therefore, contextual features are used to describe the phonetic, lexical and syntactic context of the phonemes. These contextual features, detailed in table 10.1, have been automatically extracted from the speech recordings and their text transcriptions using *ircamAlign* [LMRV08], an HMM-based segmentation system relying on the HTK toolkit [You94] and the French phonetizer Lia_phon [Bec01]. For each utterance of the training set, the text was first converted into a phonetic graph with multiple pronunciation possibilities. Then, the best phonetic sequence was chosen according to the corresponding audio file and aligned temporally with it. The context features were finally extracted according to the aligned text and the extracted phonetic sequence. A 5-states left-to-right HMM was finally used to model each contextual phoneme.

Using an Expectation Maximization (EM) algorithm, the training procedure was similar to the one described in [TZB02]: monophones models were first trained and then converted to context-dependent models. Moreover, decision-tree clustering was performed according to the extracted context features in order to obtain reliable model parameters. During the synthesis step, a parameter sequence was first generated using HTS with a constrained maximum likelihood algorithm [TMY+95] from which a speech signal is synthesized according to the SVLN method described in section 9.4. The same procedure was used for the STRAIGHT method and the impulse-source method using their respective parameters.

[4]This work has been carried out with Pierre Lanchantin and has been the subject of a conference publication [LDR10].

Phonetic features:

- **Phoneme identity (SAMPA code)**, and the following phonological features: vowel(length, height, fronting, rounding) consonant(type, place, voicing) for the central phoneme and for its neighbors (2 before and 2 after).

Lexical and syntactic features

- **Phoneme and syllable structure**: position of the phoneme in its syllable; number of phoneme in the current, previous and next syllable; position of the phoneme in the word; position of the phoneme in the phrase; nucleus of the syllable.
- **Word related**: Part Of Speech (POS) of the word and its neighbors (1 before and 1 after); number of syllable in the current, previous and next word; number of content words from the start and from the end of the phrase, number of non-content words up to the previous and next content word.
- **Phrase related**: number of syllable in the phrase; number of words in the phrase; position of the phrase in the utterance.
- **Utterance related** : number of syllables, words and phrases in the utterance.
- **Punctuation related**: punctuation of the last phrase.

Table 10.1: Context features extracted by ircamAlign

Preference test

The compared synthesis systems have been trained on a database containing 1995 sentences (approximately 1h30 of speech) spoken by a French non-professional male speaker and recorded at 16 kHz in an anechoic room. The test consisted of a subjective comparison between the three systems. 5 utterances were chosen for generating the test samples for each system. Like in the comparison test for pitch transposition, the listeners were asked to give a rate between -3 and 3 for a pair of synthesized utterances using two different system. Therefore a total of 15 comparison pairs were evaluated by each listener. 14 French native listeners answered the test. The ranking of the three systems was evaluated by averaging the scores of the test for each method (see figure 10.6).

The results show that the speech synthesized by the proposed system using SVLN has a quality between the STRAIGHT and impulse-source methods. Note that the different clustering of the context features resulted in slight prosodic differences which may alter the evaluation. In addition, as shown in the preference test for pitch transposition, the transients and unvoiced segments reduces the overall quality of the SVLN method. Finally, although the breathiness of the synthesized speech can be controlled with the SVLN method (see the preference test on breathiness above), the overall quality of the SVLN method in HMM-based synthesis is highly dependent on the reliability of the parameters.

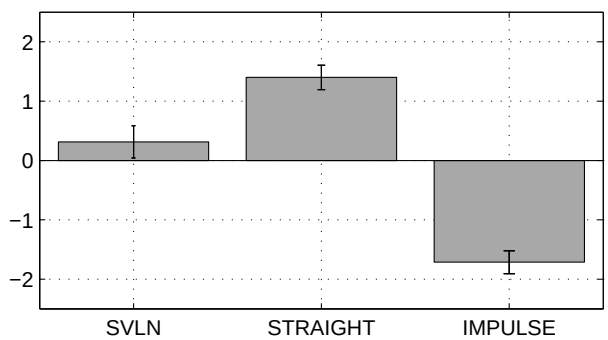


Figure 10.6: Mean preferences 95% confidence interval of the preference.

Conclusions

- In order to ensure a stable estimation of the VTF using the proposed SVLN method, the time evolution of the shape parameter estimated with the $MSPD^2$ based method is smoothed using a 100 ms median filter. In doing so, we assume that the voice quality is constant inside a single phoneme.
- Preference tests comparing SVLN, SHIP, PSOLA and STRAIGHT methods have been carried out to evaluate the overall quality of these methods in terms of pitch transposition.
  - The results of the first two tests show that the resynthesis of the unvoiced segments clearly degrades the overall quality of the transpositions made by the proposed SVLN method (see fig. 10.2).
  - Depending on the processed voice, the results show to vary for all the compared methods (see fig. 10.3). For example, whereas the STRAIGHT method outperforms on average the three other methods, this method is outperformed by the SVLN method for the male English voice used in these tests. Therefore, depending on the application, if a single method has to be chosen to process multiple voices, the STRAIGHT method should be considered. On the other hand, since there is no optimal method in the general sense, each method can have its own advantage regarding to the processing of a single voice.
  - Whereas the significance of the artifacts produced by the SHIP and STRAIGHT methods seems to be proportional to the transposition factor, the quality of the SVLN method seems to be independent of this factor. Consequently, the more important the transposition, the better the quality produced by the SVLN method compared to the other two ones (see fig. 10.4).
- Using a comparison test, we have also shown that the breathiness of two neutral male utterances can be controlled using the proposed SVLN method. Additionally, it seems easier to increase the breathiness than to reduce it.
- A last preference test showed that the quality of an HMM-based speech synthesis using the SVLN method is below the one using STRAIGHT and above the one using the basic impulse-source method.

Chapter 11

General conclusions

In this study, we assumed in the introduction that a glottal model can be used to improve existing methods of voice transformation and speech synthesis. First, the elements of the source-filter model of voice production has been described in chapters 2 and 3. According to these chapters, we assumed that the glottal pulse is a mixed-phase signal and the vocal-tract impulse response is a minimum-phase signal (see sections 2.5.3 and 3.1.2 respectively). In chapters 4 to 8, given these properties and using the LF^{Rd} glottal model, special attention has been given to methods which estimate the shape parameter of a glottal model. In addition, the efficiency of these methods has been evaluated and compared to state of the art methods using synthetic and electroglottographic signals. Finally, we proposed an analysis/synthesis method which model the glottal source using the LF^{Rd} model and Gaussian noise. The True-Envelope (TE) estimation method then complete this source to model the Vocal-Tract Filter (VTF). Preference tests have been also carried out to compare this method to state of the art methods.

The following summarizes the conclusions given all along this study at the end of each chapter.

Estimation of glottal model parameters

According to the mixed-phase and minimum-phase properties of the glottal source and the vocal-tract impulse response, it has been shown that these properties can be used to estimate the parameters of a glottal model using the Mean Squared Phase (MSP) between an observed spectrum and its model (sec. 5.1). Using this criterion, we proposed two methods which jointly estimate the position and the shape of a glottal pulse (the MSP and MSPD based methods (sec. 5.1.4 and 5.2.1)). In order to ensure that the parameters can be estimated, the conditions a glottal model and its parametrization have to satisfy have been discussed (sec. 5.1.1). Then, we proposed two other methods to estimate the shape parameter of a glottal model which are independent of the time position of this latter (the MSPD^2 method and the method FPD^{-1} based on the inversion of the Function of Phase Distortion (FPD) (sec. 6.1.1 and 6.2.1)). These two methods use a discrete approximation of the phase spectrum component which is not linear. Except for the FPD^{-1} method, all the other proposed methods estimating a shape parameter (i.e. MSP, MSPD and MSPD^2) use optimization algorithms which minimize an error function. Conversely, using the FPD^{-1} method, the shape parameter is in a quasi closed-form expression of the observed spectrum. In addition, the function of phase distortion allows to evaluate *a priori* to which extent a shape parameter of a glottal model can be estimated using the methods based on mean squared phase (6.2.2). A last method has been proposed to estimate the Glottal Closure Instants (GCI) of a speech utterance using a glottal model estimate (the GCIGS method (sec. 7.2)).

Concerning the evaluation of the proposed methods with synthetic and EGG signals: First, using the method based on MSPD^2 to estimate the shape parameter of the LFRd model, it has been shown that the GCIGS method outperforms two compared methods of the state of the art (sec. 8.2.1). By Joint estimation of the pulse position and the shape parameter, the MSP based method can be used to refine the estimated position. It has been shown this refinement increases the gross errors but improves the precision. Then, it has been shown that the weighting of the error functions involved in the estimation methods of the shape parameter influences the efficiency of these methods. In order to obtain the best efficiency for each of the compared methods, the number of harmonics taken into account in the error functions must not exceed 7 (sec. 8.2.2). Consequently, regarding to the high frequency spectral properties one can be interested in, this limit might be very low. Whereas MSP and FPD^{-1} based methods have shown promising results with synthetic signals, their efficiencies evaluated with EGG signals vary with respect to the analyzed voice. Conversely, the MSPD^2 based method outperforms all the other proposed methods and three other methods based on the estimation of the glottal source (IAIF, CC and ZZT) (sec. 8.2.2). Note that the independence of the glottal parameters estimation has been addressed with particular attention. Indeed, the estimation based on MSPD^2 is independent on the amplitude of the glottal source and independent on the glottal pulse position. In addition, to estimate one single shape parameter, the MSPD^2 based method doesn't need initial values. Indeed, a simple Brent's algorithm is sufficient which uses only the extrema of the parameter range.

Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise (SVLN)

Conversely to the STRAIGHT and WBVPM analysis/synthesis methods, the proposed SVLN method uses a glottal model to describe the deterministic component of the glottal source. This method is thus especially dedicated to voice processing (like ARX and ARMAX methods) whereas the other two methods can be applied to any pseudo-periodic signal. It has been shown that the SVLN method is always able to represent the amplitude spectrum of a given speech segment. Indeed, the TE method, which is used to estimate a smooth envelope of the VTF, completes the source model and the radiation model in order to retrieve the observed amplitude spectrum (sec. 9.3.3). Conversely, the phase spectrum of the model used by SVLN is imposed by the LFRd glottal model in low frequencies and by Gaussian noise in high frequencies.

According to stability issues encountered in the estimation of the shape parameter of the glottal model (sec. 8.3), we proposed to use a median filter and a zero-phase filter to smooth the time evolution of this parameter. The estimation of the noise level related to Gaussian noise is thus stabilized as well as the estimation of the VTF. Concerning the preference tests used to evaluate the efficiency of the SVLN method compared to three other state of the art methods (STRAIGHT, SHIP and PSOLA): The unvoiced segments resynthesized by the SVLN method clearly reduce the quality of this latter. The same issue has been encountered in HMM-based synthesis. However, it has been shown that the overall quality of important pitch transpositions ($\pm$1200 cents) of the SVLN method can be higher than the ones of the other compared methods (sec. 10.2). Finally, using a comparison test, it has been shown that the breathiness of two neutral male utterances can be clearly controlled using the SVLN method (sec. 10.3).

11.1 Future directions

Estimation of glottal parameters

In this study, the LFRd glottal model has been widely used. However, many other glottal models exist (see sec. 2.4) and many of them, like the original LF model, have multiple parameters. Whereas this study mainly focused on estimation methods using a glottal model and a parametrization which cover a fairly small shape space, it would be interesting to widen this space by studying various existing glottal models and the estimation of multiple shape parameters.

In section 5.3, it has been shown that the three shape parameters of the LF model are dependent on each other in terms of mean squared error and mean squared phase. Moreover, section 5.3.1 has shown that the two shape parameters O_q and α_m of the LF model can compensate each other in order to keep a relatively smooth evolution of a characteristic of the LF shape. A full study about the existing glottal models and the dependency of their parameters related to the existing estimation methods would thus complement the point discussed above.

Optimal glottal model

As a continuation of the previous point, a glottal model could be created which should have independent shape parameters regarding to a given error function (if these parameters can not be in a closed-form expression of observed data). Whereas previous glottal models have been mainly inspired by physiological and physical considerations, a balance should be found between these considerations and the numerical conditions which have to be satisfied in order to obtain reliable parameter estimates. According to this study, the conditions expressed in section 5.1.1 and the elements discussed in appendix B have to be considered using a method minimizing the mean squared phase.

Note that the recent progress of the High-Speed Videoendoscopy (HSV) should be also considered. For example, a study of simulations of glottal flow from glottal area estimations like proposed in figure 2.3 could be interesting when designing a glottal model. First, the diversity the glottal source can cover in various vibration modes of the vocal folds can be taken into account. Then, the discrepancy between the glottal flow, which is coupled with the vocal-tract shape, and a non-interactive glottal model like most of the current models, can be minimized.

Voice modeling using a glottal model

In this study, using the SVLN method, it has been shown that a glottal model can be used to improve important pitch transpositions compared to state of the art methods. However, the transients between voiced and unvoiced sounds are not properly managed with the proposed method. Indeed, the nature of the excitation source of the VTF changes drastically between these two types of sounds. Current methods like STRAIGHT, WBVPM, PSOLA and vocoders manage this situation with a signal model which is able to model both sounds. Conversely, using a glottal model, the proposed analysis/synthesis method has to switch between two situations where there are glottal pulses or not. Such transitions are thus sources of artifacts. In conclusion, we consider that it is still an open question how to robustly model voice production in both unvoiced and voiced segments where a glottal model is used.

Voice modeling for voice transformation and speech synthesis

The amplitude spectrum has been widely studied through the development of estimation methods of smooth envelope (e.g. LP, DAP, TE). However, we consider that the phase spectrum has not yet received the same attention. Whereas the SVLN method uses a glottal model to describe the phase spectrum at low frequencies, the high frequencies are modeled by Gaussian noise. Therefore, either the phase spectrum is described by a strictly deterministic signal or a full random signal. Obviously, one can expect that the glottal source is not only made of these two kind of phase spectra. A balance should thus be found between these two extrema. Finally, it would be interesting to study the glottal noise and its properties, known to which extent the modulation of the glottal noise is important from a point of view of the perception and how to control efficiently the properties of this noise. By giving attention to the mixture of the deterministic component at high frequencies, the noise and its modulation, the space of voices which can be handled by analysis/synthesis methods should be broadened.

Appendix A

Minimum, zero and maximum phase signals

This appendix discusses the minimum, zero and maximum phase properties of a signal related to their realizations using the real cepstrum. However, since these properties are related, by definition, to the zeros of the z-transform of the speech signal only, we first briefly discuss how the poles of the vocal-tract transfer function are managed using DFTs.

Indeed, in this study, we assume that the DFT can be used to approximate the resonances of the vocal-tract filter. Therefore, the infinite impulse response of a pole is truncated using a finite impulse response and thus represented by zeros. In doing so, both poles and zeros of the VTF are identically represented by zeros and it is thus sufficient to use the minimum, zero and maximum phase properties to describe where the zeros (and the approximated poles) are located around the unit circle and how the energy of the signal is distributed along the time scale. Note that, this approximation of the poles is possible only if the size of the DFT is sufficiently large compared to the bandwidth of the poles. The decrease of the exponential behavior of the poles has to be sufficient to ensure that the energy of the impulse response at the end of the DFT is negligible. A minimum DFT size can thus be estimated according to the minimum bandwidth of a pole B_{min} and a chosen minimum attenuation at the end of the DFT. The decay rate of a pole of bandwidth B Hz for a sampling frequency f_s is

$$d_r = B\frac{\pi}{f_s}$$

and the corresponding attenuation A_{lin} in linear amplitudes after a time t is

$$A_{lin} = e^{-d_r \cdot f_s \cdot t}$$

Finally, the minimum DFT size, for an attenuation A in dB, is thus expressed as:

$$t = -A\frac{\log(10)}{20 \cdot \pi \cdot B_{min}}$$

The approximation of a pole by the DFT is thus satisfied using a 150 ms DFT size for a typical value $B_{min} \approx 30$ Hz and an attenuation of $A = -100$ dB.

A.1 The real cepstrum and the zero and minimum phase realizations

First, the real cepstrum of a signal defined by a spectrum $X[k]$ is computed as follows [OS78]:

$$\tilde{x}[n] = \mathcal{F}^{-1}\left(\log|X[k]|\right) \tag{A.1}$$

$\tilde{x}[n]$ thus represents $|X[k]|$ which is a zero-phase spectrum. Therefore, the zero-phase realization $Y[k]$ of $X[k]$, which is the construction of $Y[k]$ from $X[k]$, is simply obtained by using the inverse computation of the real cepstrum:

$$Y[k] = \exp\left(\mathcal{F}(\tilde{x}[n])\right) \tag{A.2}$$

Note that this realization, as well as the others in the following, is multiplicative (i.e. $f(A \cdot B) = f(A) \cdot f(B)\ \forall |A|, |B| \geq 0$). Indeed,

$$\begin{aligned}
\exp(\mathcal{F}(\mathcal{F}^{-1}\left(log|A[k] \cdot B[k]|\right)) &= \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|A[k]| \cdot |B[k]|\right)) \\
&= \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|A[k]| + \log|B[k]|\right)) \\
&= \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|A[k]|\right) + \mathcal{F}^{-1}\left(\log|B[k]|\right)) \\
&= \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|A[k]|\right)) + \mathcal{F}(\mathcal{F}^{-1}\left(\log|B[k]|\right))) \\
&= \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|A[k]|\right))) \cdot \exp(\mathcal{F}(\mathcal{F}^{-1}\left(\log|B[k]|\right)))
\end{aligned}$$

Then, the causality and the minimum/maximum phase properties have strong relations in the cepstral domain. Indeed, the signal related to a causal cepstrum is minimum-phase and the signal of an anti-causal cepstrum is maximum-phase [OS78]. Accordingly, the minimum-phase realization $\tilde{x}[n]_-$ of a given real cepstrum $\tilde{x}[n]$ is computed by [OS78, p.794]:

$$\tilde{x}[n]_- = \begin{cases} \tilde{x}[n] & n = 0, N/2 \\ 2 \cdot \tilde{x}[n] & 1 \leq n < N/2 \quad \text{(the causal part)} \\ 0 & N/2 < n \leq N-1 \quad \text{(the anti-causal part)} \end{cases} \tag{A.3}$$

where N is the size of the DFT. One can see this operation as a mirroring of the amplitudes of the anti-causal part into the causal part. The energy of the amplitude spectrum is thus kept untouched whereas the minimum-phase spectrum is retrieved. The spectrum of the corresponding minimum-phase signal can be finally retrieved with:

$$X[k]_- = \exp\left(\mathcal{F}(\tilde{x}[n]_-)\right) \tag{A.4}$$

The multiplicative property of this realization is easily shown using the distributivity of the real cepstrum computation:

$$\begin{aligned}
\mathcal{F}^{-1}\left(log|A[k] \cdot B[k]|\right) &= \mathcal{F}^{-1}\left(\log|A[k]| \cdot |B[k]|\right) \\
&= \mathcal{F}^{-1}\left(\log|A[k]| + \log|B[k]|\right) \\
&= \mathcal{F}^{-1}\left(\log|A[k]|\right) + \mathcal{F}^{-1}\left(\log|B[k]|\right)
\end{aligned} \tag{A.5}$$

A weighting of the cepstral coefficients of $A[k] \cdot B[k]$ like in equation (A.3) is thus distributed to both

terms $A[k]$ and $B[k]$, and the whole realization is therefore multiplicative.

A.2 Generalized realization

It is interesting to see that the realizations of these phase properties can be generalized. First, using the the maximum-phase realization:

$$\tilde{x}[n]_+ = \begin{cases} \tilde{x}[n] & n = 0, N/2 \\ 0 & 1 \leq n < N/2 \quad \text{(the causal part)} \\ 2 \cdot \tilde{x}[n] & N/2 < n \leq N-1 \quad \text{(the anti-causal part)} \end{cases} \tag{A.6}$$

Then, by scaling the imaginary part of the log spectrum of the maximum-phase realization, the usual realizations of the three phase properties can be generalized as follows. First the log spectrum of the maximum-phase realization is:

$$\hat{X}[k] = \mathcal{F}(\tilde{x}[n]_+) \tag{A.7}$$

Then, the generalized realization of the phase properties is:

$$X^\gamma[k] = \exp\Big(\Re(\hat{X}[k]) + \gamma j \Im(\hat{X}[k])\Big) \tag{A.8}$$

The minimum, zero and maximum phase realizations can thus be obtained with γ equal to -1, 0 and 1 respectively [1]. Figure A.1 shows an example using two periods of a speech signal windowed with a hanning window $w[n]$. Its DFT is thus computed using $X[k] = \mathcal{F}(w[n] \cdot x[n])$ (where $\cdot$ is the element to element multiplication). Then, two spectra are realized using the equation (A.8) with $\gamma = -1$ and $\gamma = 1$ corresponding to the minimum and maximum phase realizations respectively (with $\gamma = 0$, the phase spectrum of the zero-phase realization is equal to zero for all frequencies and the corresponding time signal is symmetric around $t = 0$). Note that the phase spectrum of the minimum-phase signal is the opposite of the phase spectrum of the maximum-phase signal. Indeed, from equation (A.8) one can see that $X^{-\gamma}[k] = \overline{X}^{\gamma}[k]$ which also implies $\angle X^{-\gamma}[k] = -\angle X^{\gamma}[k]$ (in the case of minimum and maximum phase properties: $\angle X_-[k] = -\angle X_+[k]$). Sound examples `snd_appendixA_synth_*_phase.wav` are synthesized using an LFRd model and a /ɛ/. In terms of perception, one can notice a difference between the two extrema `snd_appendixA_synth_min_phase.wav` and `snd_appendixA_synth_max_phase.wav`.

Using the speech sample of figure A.1, figure A.2 shows that the distribution of energy of the signal is related to the γ value. The local energy of the signal is computed using a sliding hanning window $w[n]$ which is used above to compute the DFT of the signal (i.e. both signal and window have thus the same duration):

$$E_l[n] = |s[n]|^2 \circledast w[n]$$

Note that the energy distribution is symmetric for $\gamma = 0$ only since a zero-phase signal is symmetric. Moreover, the zero-phase signal is the most concentrated in the time domain. One can also see that for increasing $|\gamma|$ the spread of the signal also increases. Finally, in terms of causality, figure A.3 shows the energy of the causal part and that of the anti-causal part of the speech sample corresponding to $X^\gamma[k]$. One can see that the signal is causal if and only if $\gamma = -1$ and anti-causal if and only if $\gamma = 1$.

The minimum and maximum phase realizations of a signal can also be obtained by mirroring, around

[1] The minimum-phase realization is not used in order to keep a consistency between the delay of the center of energy of the signal and the sign of the γ value

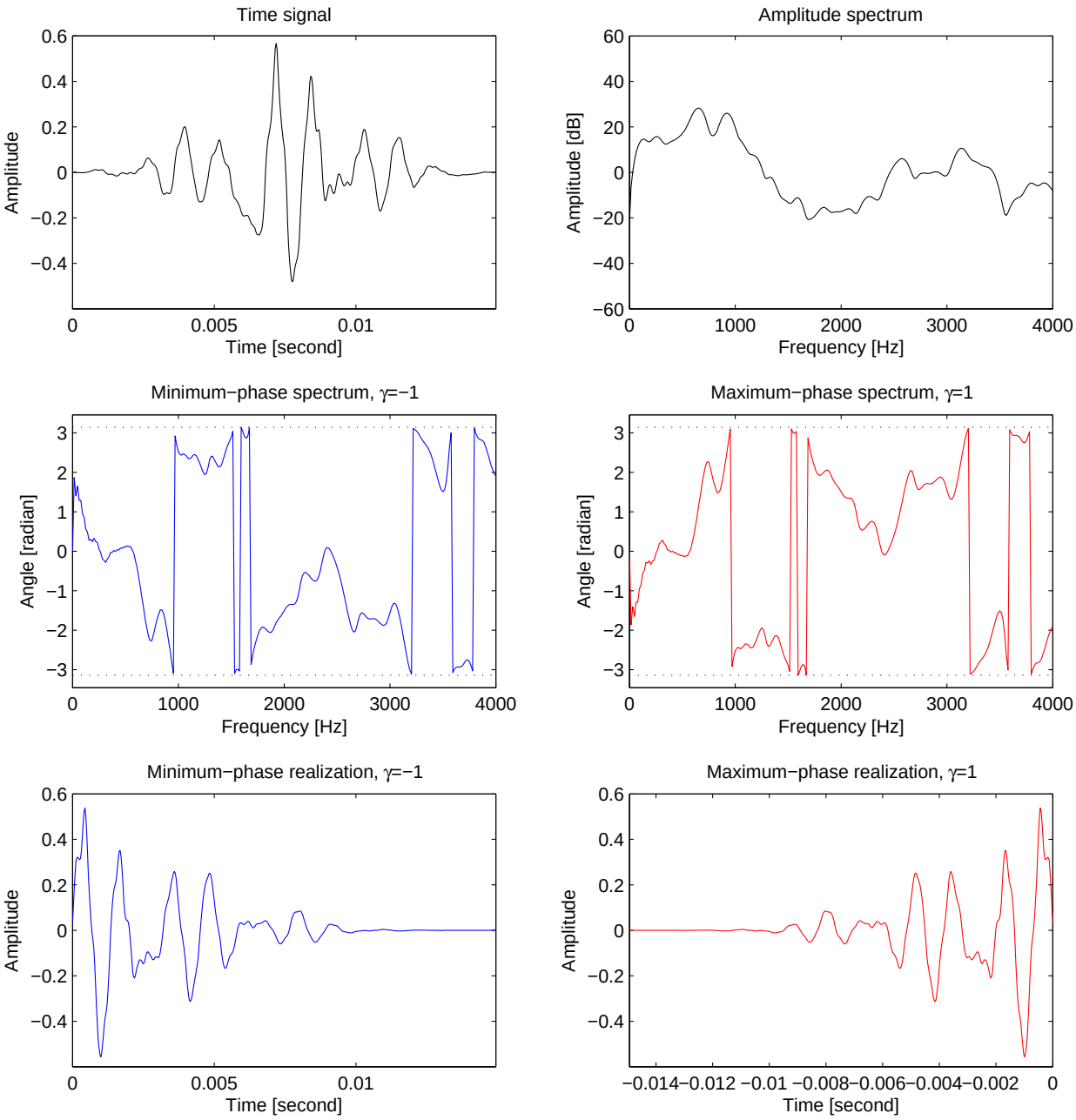


Figure A.1: Example of two periods of a speech signal $x[n]$, its amplitude spectrum $X[k]$, its corresponding minimum and maximum phase spectra and their corresponding signals.

the unit circle, the Zeros of the Z-Transform (ZZT) of this signal. However, a few differences exist between this approach and the one used in this appendix. Firstly, by scaling the imaginary part of the log spectrum (eq. A.8), the amplitude spectrum is not modified whereas a mirroring of one zero modify the gain of the spectrum regarding to the distance between this zero and the unit circle. Additionally,

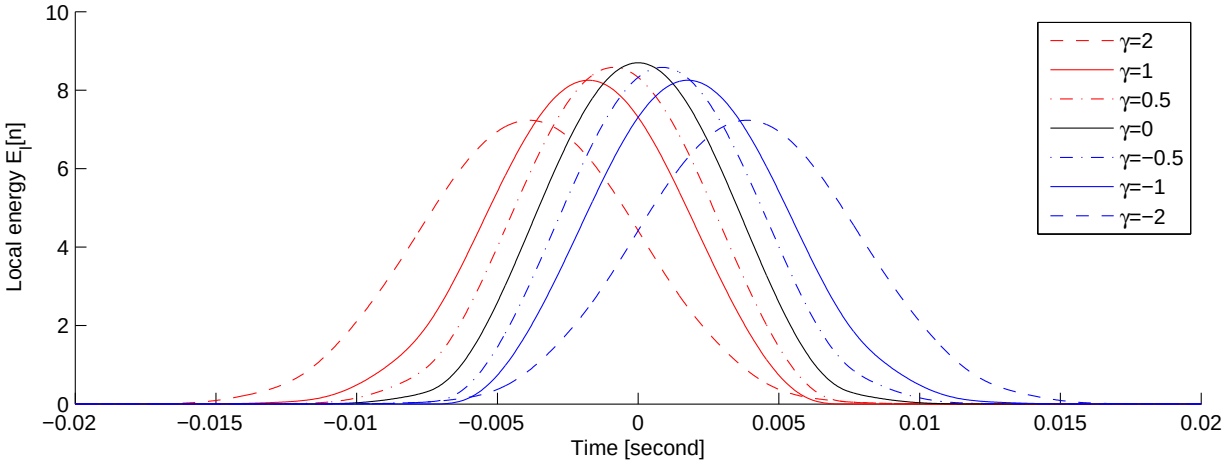


Figure A.2: Local energy distribution for various γ values.

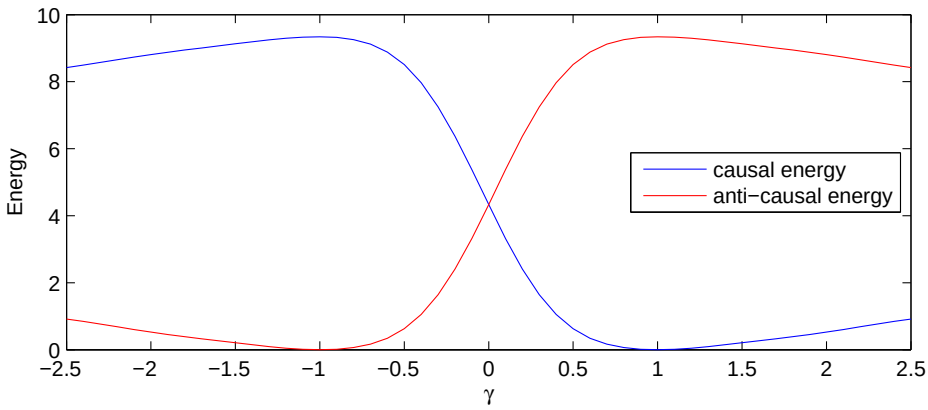


Figure A.3: Causal and anti-causal energy for varying γ values.

no linear-phase component exists in $X^{\gamma}[k]$. Indeed, whereas each mirrored zero outside the unit circle add an extra delay of one sample, $X^{\gamma}[k]$ always satisfies the maximum flat phase criteria [Bon04] (the phase spectrum has to start at zero radian at DC frequency and has to finish at zero radian at Nyquist frequency). Finally, the signals spaces covered by mirroring zeros or using $X^{\gamma}[k]$ are very different. On one hand, the phase spectrum of $X^{\gamma}[k]$ is expressed as a linear combination of only one base vector (here, the phase spectrum of the maximum-phase realization). On the other hand, by mirroring the zeros of the z-transforms a given frequency band can be drastically modified keeping the other parts of the spectrum almost untouched.

Appendix B

Error minimization and glottal parameters estimation

The reliability of estimation of glottal model parameters is a known issue in speech analysis. Indeed, many glottal models have been proposed but stability issues have been reported concerning the estimation of their parameters [Mil86]. Various heuristics have been proposed to deal with this issue: use multiple periods [PB09, Mil86], use a rough glottal model and refine the solution with a more sophisticated one [VRC05a, Lu02], or use smoothing filters like the ones in section 10.1. In this appendix, we first discuss a few elements which significantly influence the results of the estimation methods of glottal parameters. To the best of our knowledge, these elements are often considered as technical details and seldom discussed in the literature to the benefit of the presentation of higher level and novel ideas. Two procedures are then presented which are used to estimate the LF^{Rd} glottal model based on methods estimating the glottal source (IAIF, CC and ZZT methods).

B.1 Error minimization

Since the estimation of glottal parameters are usually not expressed in a closed-form of the observed data, an optimization algorithm (e.g. Brent's method, Preconditioned Conjugate Gradient, Simplex, etc.) is used to minimize a given error function. In order to find a relevant solution using these algorithms, some of the conditions discussed below have to be absolutely satisfied while others properties can improve the speed as well as the efficiency of the methods.

Continuity

It is necessary that the error function is a continuous function with respect to a variation of the parameters to optimize. Firstly, the shape of a glottal model has to be a continuous function of its parameters. Note that the LF model partly respects this condition because its synthesis parameters (see α and ϵ in sec. 2.4) are not expressed in a closed form of the shape parameters. An approximation method is thus necessary (e.g. Newton's method). Since such a method stops for a given threshold, the difference of glottal shapes generated by two neighbor parameter sets are related to that threshold. In the following, this threshold is assumed to be arbitrarily small to ensure a pseudo-continuity of the glottal shape with respect to its shape parameters. Secondly, the VTF estimated by $\mathcal{E}(X)$ has to be continuous with respect

to X. Therefore, the envelope estimation methods, which are iterative, are not advised. Indeed, for the TE method, the DFT bins which are fitted by the envelope can change drastically for two very close arguments X. Using the TE method, a discontinuous relation between the argument and the estimated envelope can be expected. Note also that although the support points used by the DAP method are fixed conversely to the TE method, this method stops for a given threshold which has to be small enough in order to ensure a pseudo-continuity of the envelope with respect to X (like discussed above for the synthesis parameters of the LF model).

In the proposed methods in this study (the ones using MSP, MSPD and MSPD^2), the condition of continuity is satisfied because the minimum-phase reconstruction through the power-cepstrum is a closed-form expression of a spectral division of the observed spectrum by the glottal and radiation models.

Uniqueness of the minimum

To ensure that a solution can be found by most of the algorithms minimizing an error function, this latter has to have only one single minimum, i.e. the error function has to be monotone and increasing towards the left of the optimal parameter and monotone and increasing towards the right. Moreover, if this condition is satisfied, the minimum found by those algorithms is not sensitive to an initial value (see MSP and MSPD error functions, figures 5.1 and 5.3). Finally, if the error function is convex, it implies that it has a unique minimum (all points of the error function are "visible" from the optimal point).

B.2 Parameter estimation based on glottal source estimate

In the evaluation tests of chapter 8, three external methods were compared with the proposed ones. Since these methods estimate the glottal source and not the glottal parameters directly, it is necessary to fit the glottal model on the source estimates. Therefore, the following two sections describe the procedures used in this study to do this fitting.

First, for all of the compared methods, the analyzed signal is resampled to 16 kHz in order to avoid the influence of high frequencies which can be irrelevant compared to the estimated parameters. Moreover, for synthetic signals, the error measure is limited to a Voiced/Unvoiced Frequency fixed to 2 kHz. This value is kept constant in order to have all of the compared methods equally affected by this limit. Finally, the analyzed real signals are high-pass filtered at 40 Hz in order to avoid any influence of irrelevant low frequency components (note that this filter greatly influence the reliability of the CC and ZZT methods).

B.2.1 Procedure for the Iterative Adaptive Inverse Filtering (IAIF)

In order to obtain parameter estimates of a glottal model from the estimated glottal source of the IAIF method [ATN99, Alk92], the LFRd model is fitted to the estimated source with a Preconditioned Conjugate Gradient (PCG) algorithm by minimizing the mean squared error in the time domain (see figure B.1). In the original implementation of the IAIF method available in the *Aparat* toolkit [Air08, APBA05], the 3 LF shape parameters are estimated as well as the excitation amplitude E_e. To obtain a valuable comparison with the proposed methods, that implementation has been replaced in order to estimate the Rd shape parameter only (note that E_e has to be estimated jointly with Rd because the mean squared error is sensitive to the glottal model amplitude). Additionally, the original fitting procedure has been corrected according to the condition of continuity discussed above. Consequently, the position of the glottal pulse is optimized using a linear-phase on the spectral representation of the glottal model and not

with an integer shift of its time domain representation. Secondly, the dependency between the optimized parameters has to be as small as possible. Whereas the original implementation optimizes the time domain parameters (t_p, t_e) which are both directly dependent on the glottal model position, the linear-phase of the glottal model and the Rd parameter, which does not influence this linear-phase. Finally, in order to obtain a smooth influence of the time position of the glottal model on the error function, the mean squared error is weighted in time domain by a two-period hanning window centered on t_e. Taking into account these considerations, the results of the original IAIF implementation have been greatly improved.

B.2.2 Procedure for the Complex Cepstrum (CC) and the Zeros of the Z-Transform (ZZT)

For these two methods, the LF fitting procedure discussed above for the IAIF method is used on the glottal source estimated with the CC and ZZT. In addition, in figure B.1, one can see that the glottal pulse is damped to the left by the analysis window used in these decomposition algorithms. Therefore, during the fitting of the LF model, the same window is applied to the glottal model in order to reduce a possible bias between the observed pulse and its model. In this study, the implementations of the CC and ZZT decomposition methods are the ones used in [DBD09].

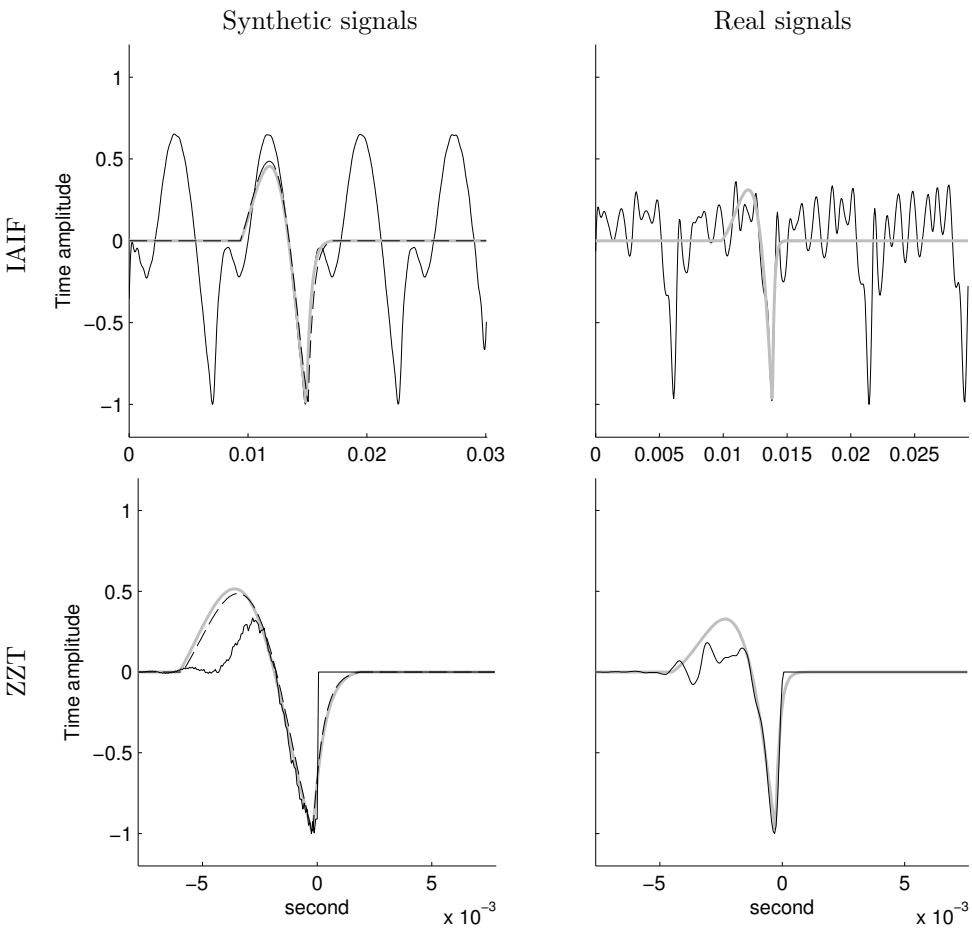


Figure B.1: Examples of fitting of the LF^{Rd} glottal model on glottal source estimates of real and synthetic signals. The estimated glottal source in thin solid black line, using IAIF or ZZT. The synthetic glottal pulse is shown in dashed line and the estimated LF^{Rd} pulse denoted by thick gray line.

Appendix C

The glottal area estimation method

Although stroboscopy is currently the most used technology for videendoscopy, the High-Speed Videoendoscopy (HSV) received a lot of attention these last years mainly because this tool starts to be accessible even though it has been developed since decades. Moreover it has been shown that HSV overcome the limitation of the stroboscopy [Del06]. Therefore, one can expect that HSV replace stroboscopy in a close future. The extraction of features related to the vocal-fold motion recorded by HSV is thus an interesting research subject. Both stroboscopy and HSV are mainly designed in a clinical context for pathology assessment [MHT00, NMEH01] but other applications can be found like in voice analysis where the area of the glottis can be extracted from the captured images and compared to other features extracted from the acoustic waveform [DBR08]. In this appendix, such a method of glottal area estimation is presented. Most of the current methods of glottal area estimation uses edge detection algorithms to locate the vocal folds limits [DP03, Pul05, NMEH01]. In addition, it also possible to obtain an estimation of the glottal area without estimating the geometry of the glottis [Del06].

C.1 The proposed method for glottal area estimation

Basically, we propose to use a threshold technique of the HSV images. We assume that the glottis has a darker luminance than that of the vocal folds. Therefore, the luminance is first thresholded according to the method described below (see also diagram of figure C.3 and figures C.1 C.2). Secondly, a thresholding of the variance of the HSV images is also used in order to exclude static parts of the HSV images from the estimate glottal area. We propose also to estimate the necessary thresholds automatically.

Automatic threshold estimation

Below is described the estimation of the luminance and variance thresholds. Firstly, the luminance images are high-pass filtered in the time domain with a cutoff frequency of 40 Hz in order to remove the slow varying motions:

$$H_{ij}^t = L_{ij}^t \circledast f^t \qquad \forall i, j$$

where L_{ij}^t at time t is the two dimensional luminance of the HSV images and f^t is the impulse response of a zero-phase high-pass FIR filter. Secondly, the variance of the HSV images is computed for each time

t over four fundamental periods T_0 using H_{ij}^t:

$$V_{ij}^t = \mathrm{var}(H_{ij}^{t-4T0/2 \ \dots \ t+4T0/2}) \quad \forall i, j$$

The main vibrating parts of the image is thus emphasized in V_{ij}^t. Thirdly, we assume that the edge of the vocal-folds are good candidates to find the threshold values. These edges are thus extracted as follows. Since motions, which are not part of the glottis, can appear outside of the glottis (e.g. sparkling effects), we propose to first compute a minimum luminance over four fundamental periods for each time t:

$$M_{ij}^t = \min(L_{ij}^{t-4T0/2 \ \dots \ t+4T0/2}) \quad \forall i, j$$

a representation of the most open glottis in the four periods is thus retrieved. Then, the horizontal derivative of each minimum image M_{ij}^t is computed using a Sobel filter in order to emphasis the edge of the vocal folds:

$$D_{ij}^t = dM_{ij}^t/dj$$

Finally, for each time t, both luminance and variance thresholds $l[t]$ and $v[t]$ are computed through a weighted mean of the luminance L_{ij}^t and the variance V_{ij}^t respectively. In order to emphasize the folds edge and the motion in the weights, this latter is computed as follows:

$$W_{ij}^t = (D_{ij}^t)^2 \cdot V_{ij}^t$$

where the squared value removes the difference between positive and negative values of the horizontal derivative and '$\cdot$' is the pixel-to-pixel multiplication. The thresholds are thus:

$$l[t] = \sum_i \sum_j W_{ij}^t \cdot L_{ij}^t / w$$

$$v[t] = \sum_i \sum_j W_{ij}^t \cdot V_{ij}^t / w$$

where $w = \sum_i \sum_j W_{ij}^t$.

Masks and glottal area estimation

Finally, the two dimensional glottal area is retrieved by pixel-to-pixel multiplication of the two masks obtained through the thresholding of the luminance and the variance (see fig. C.2). The one dimension glottal area is the sum of the pixels of the two dimension glottal area. Examples of 1D glottal area estimates can be seen in figures 2.2 and 2.3. Note that for each HSV recording used in this study, the validity of the estimated 2D glottal area has been verified visually by superimposing the area perimeter on the colored image. The validity of these estimates are thus similar to the one of a manual annotation.

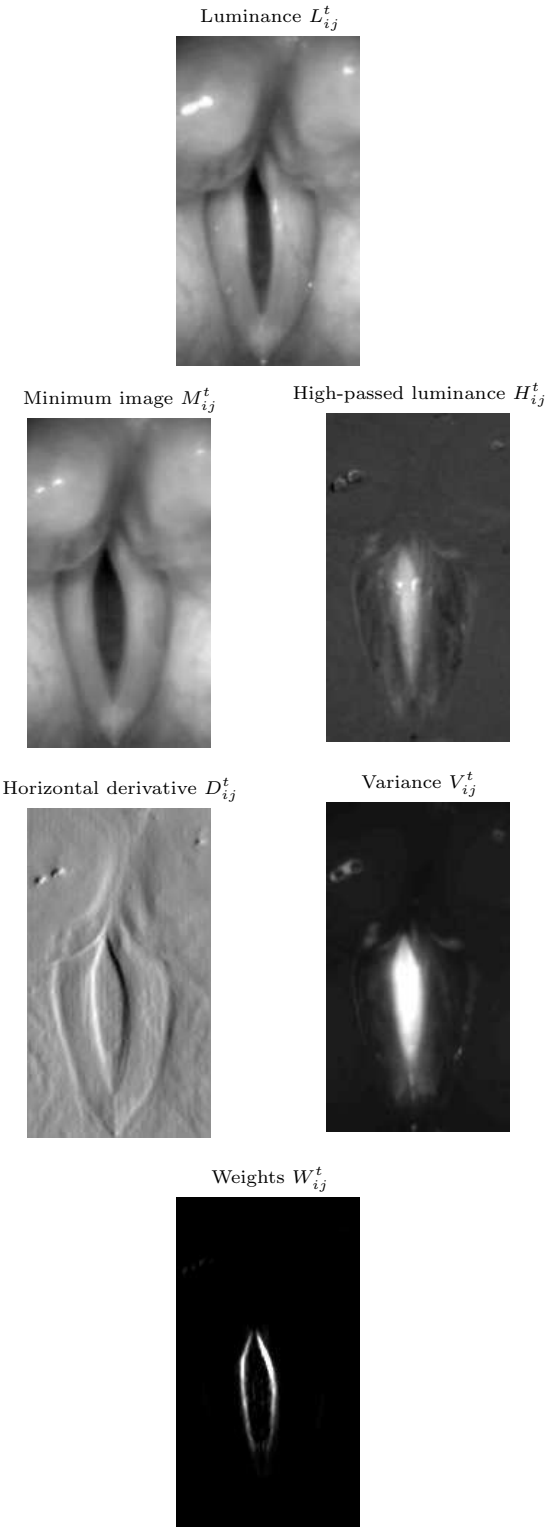


Figure C.1: Images related to the automatic threshold estimation.

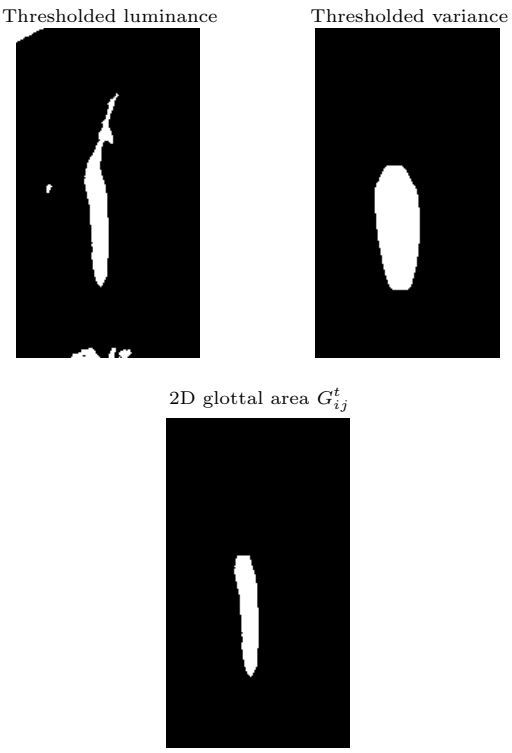


Figure C.2: Images related to the masks estimation and the 2D glottal area.

©2010 All HSV images are properties of Erkki Bianco & Gilles Degottex - Ircam.

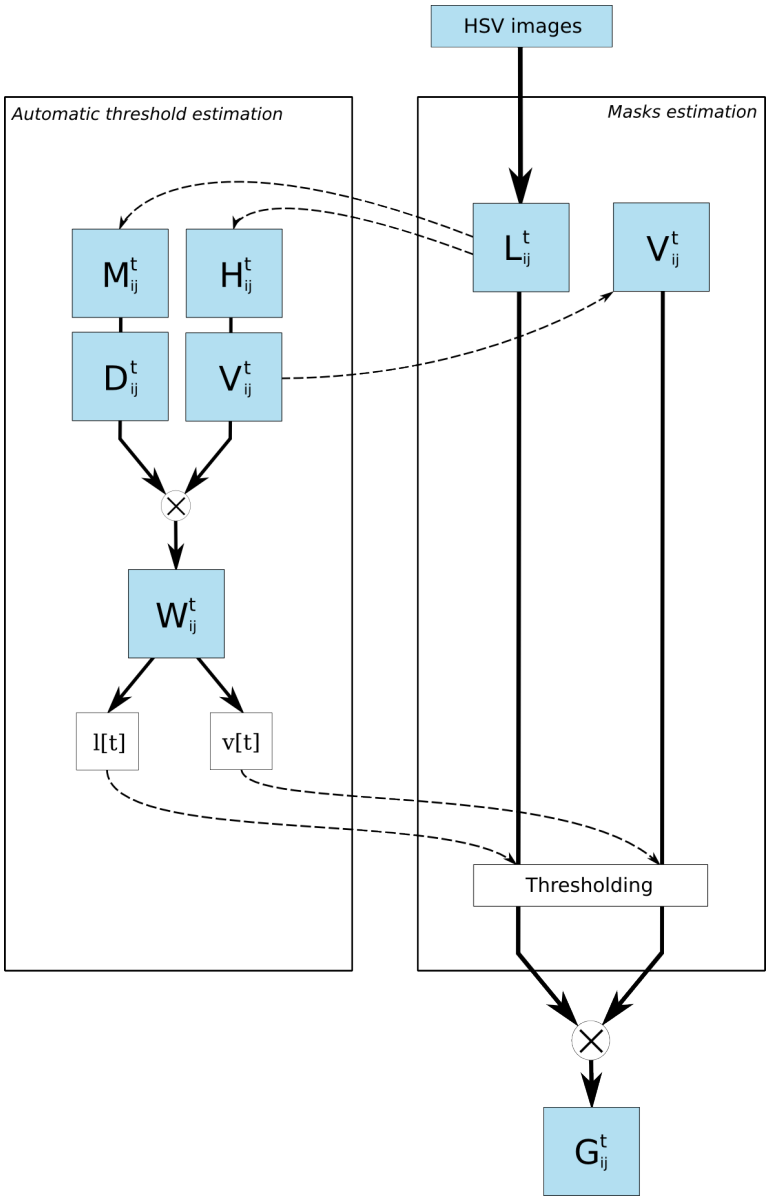


Figure C.3: Diagram of the estimation of the 2D glottal area.

C.2 Spectral aliasing

Using HSV, the motion of the vocal folds are sampled at a given frequency (4000 Hz with the used system for this study). This sampling obviously raises an issue of spectral aliasing. For example, given a fundamental frequency of 100 Hz and a Nyquist frequency of 2000 Hz, only the first 20 harmonics have a frequency below the Nyquist frequency. However, one may expect higher harmonics which will be thus aliased in the frequency domain. Figure C.4 shows the DFT of 4 periods of a 1D glottal area estimate. One can see that aliased partials exist at 1242 Hz and 1573 Hz which are pointed out by the vertical lines. Therefore, these aliased partials create some distortion on the partials below the Nyquist frequency. In conclusion, the glottal area has to be carefully interpreted if its spectral slope do not fall fast enough compared to the sampling of the high-speed camera.

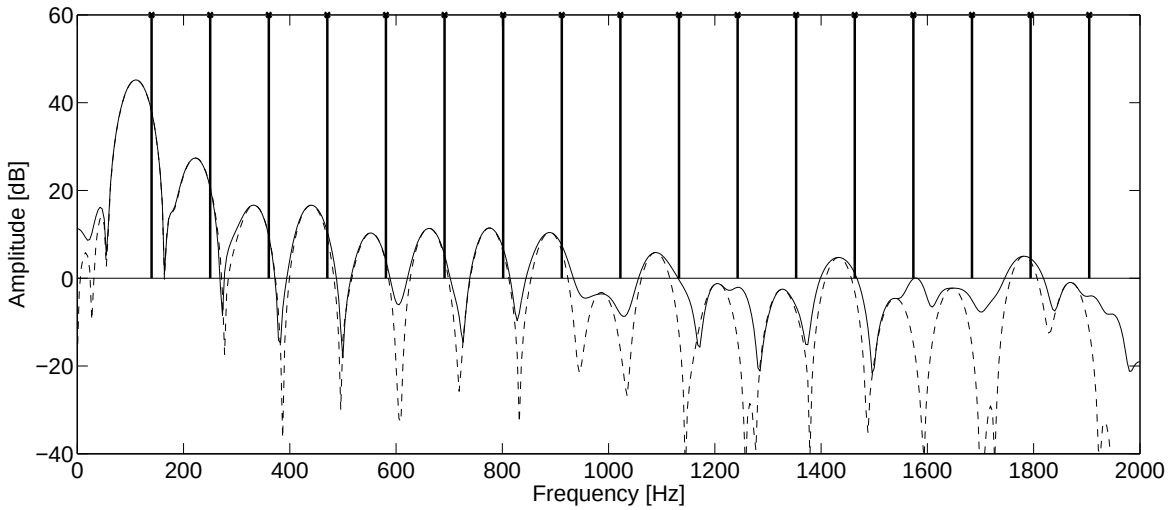


Figure C.4: The spectrum of the estimated glottal area in full line, a harmonic model of the glottal area in dashed line and the frequencies of aliased partials pointed out by vertical lines.

C.3 Equipment

During the measurements used in this study, the HSV system *Richard Wolf - ENDOCAM 5562* has been used. The vocal folds are filmed through a rigid endoscope which passes through the mouth, connected to a high-speed camera which provides 4000 colored images per second in 256x256 pixels (other resolutions and speed setups are available, but in this study, the speed has been preferred).

Appendix D

Maeda's voice synthesizer

Using equation of motion, that of continuity and that of wall vibration, S. Maeda proposed a transmission-line representation of the voice production (fig. 1.2) and a corresponding time-domain simulator of the speech production [Mae82a]. The shape of the vocal-tract is modeled by sections and the area of these latter are parametrized by the Linear Articulatory Model (LAM) [Mae79] (see table D.1 for the parameters used in this study and the resulting VTFs in figure D.4).

In this study, the lengths of the pharynx, the buccal cavity and the nasal cavity are fixed to 9 cm, 8 cm and 10 cm respectively. Whereas the glottal area can be controlled by various analytical models in the original implementation, we used the glottal area estimated from HSV recordings using the method presented in appendix C. Based on the code source available in [Gho99], many options are available. Here, we chose a configuration according to the base article presenting this synthesizer [Mae82a].

From our experiments and according to [Mae82b], the high vowels /ɛ/ and /œ/ can be nasalized by increasing the nasal coupling area resulting in natural sounds. However, nasalization of /a/ and /ɔ/ are more difficult to obtain. In [Mae82b], Maeda proposed to add a sinus cavity representing the maxillary sinuses to improve the naturalness of these nasalized sounds. However, this option is not available in the used implementation when the frequency response of the VTF is estimated. Consequently, although the naturalness of these two nasalized sounds are not optimal, we used these latter in this study.

D.1 Space and time sampling

In such a time domain simulation, due to space and time sampling, problems of frequency warping exists [WF78]. By limiting the model to a lossless uniform tube with rigid walls, it is possible to express the continuous to discrete frequency mapping [Mae82a, eq.35] (figures D.1 and D.2). From [Mae82a], a maximum section length of 1 cm and a sampling frequency of 20 kHz should be sufficient to assure proper formants frequencies up to 4 kHz. However, since we mostly worked with 44.1 kHz recordings in this study, we used a maximum space sampling of 0.5 cm and a time sampling of $3 \cdot 44.1$ kHz.

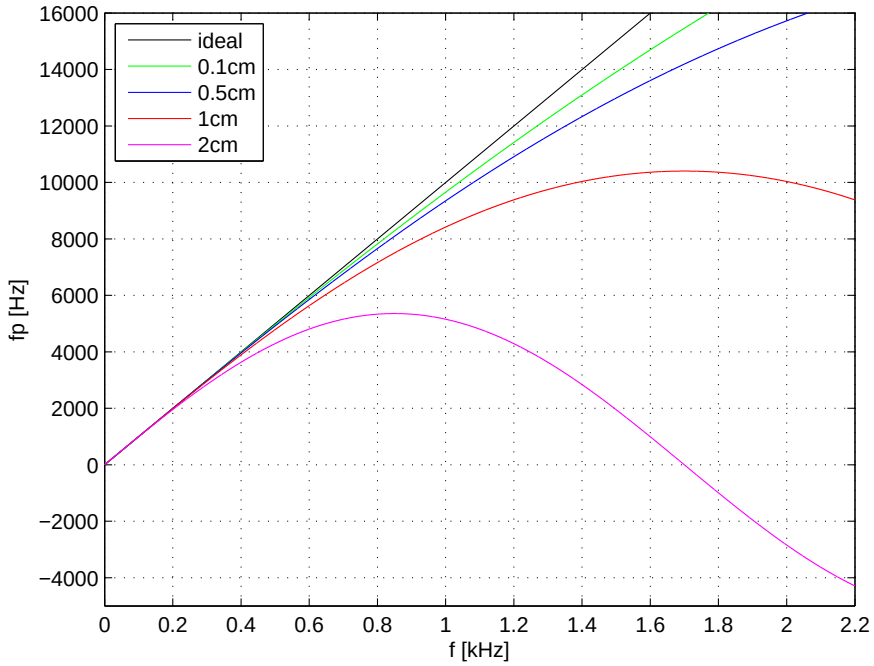


Figure D.1: Frequency mapping between continuous frequency f and discrete frequency fp. $f_s = 3 \cdot 44.1\,\text{kHz}$. The curves show different space sampling.

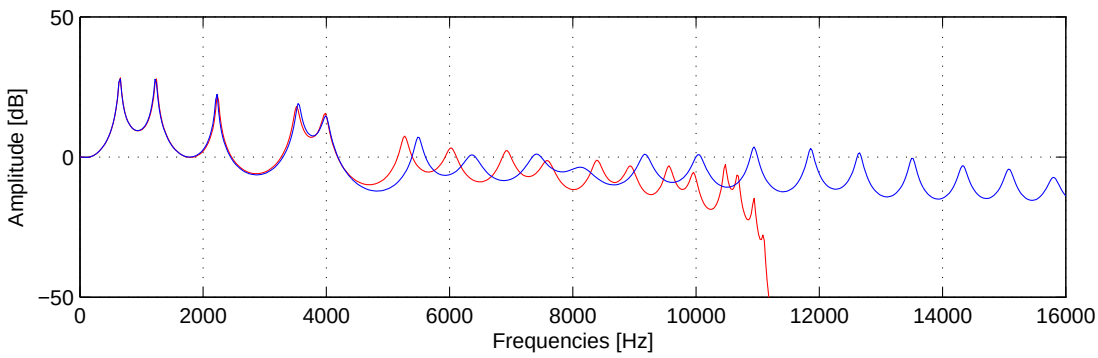


Figure D.2: Synthesized spectra (colors are the same as in figure D.1. Note that the time sampling is not relevant for the estimation of the VTF).

IPA	X-SAMPA	Example[1]	jaw	tongue	shape	apex	lip_ht	lip_pr	larynx	nasal[2]
i	i	ou**i**	0.50	-2.00	1.00	-2.00	1.00	-1.00	0.00	0.00
e	e	s**e**s	0.00	-1.00	1.00	-2.00	1.00	-1.00	0.00	0.00
ɛ	E	m**ê**me	-1.00	0.00	1.00	-2.00	1.00	-0.50	0.00	0.00
a	a	d**a**me	-1.50	2.00	0.00	-0.50	0.50	-0.50	0.00	0.00
ɔ	O	gr**o**s	-0.70	3.00	1.50	0.00	-0.60	0.00	0.00	0.00
u	u	l**ou**p	0.50	2.00	1.50	-2.00	-1.00	1.50	0.00	0.00
ø	2	d**eu**x	0.50	-1.00	1.00	-2.00	-0.50	1.00	0.00	0.00
œ	9	n**eu**f	-1.00	-0.50	0.50	-2.00	0.20	-0.50	0.00	0.00
œ̃	9∽	l**un**di	-1.50	2.00	0.00	-0.50	0.50	3.00	0.00	1.00
ɔ̃	O∽	b**on**bon	-1.50	3.00	0.50	-3.00	0.00	0.00	0.00	0.30
ɛ̃	E∽	c**in**q	-1.50	1.00	0.00	-0.50	3.00	-3.00	0.00	0.20
ɑ̃	A∽	b**an**que	-1.50	3.00	0.50	-3.00	1.00	-0.50	0.00	0.30
	"hsv"		-3.00	-1.00	-3.00	-1.50	3.00	0.00	-1.50	0.00

[1] in french [2] nasal coupling area $[cm^2]$

Table D.1: Phonemes and their corresponding LAM parameters. The corresponding sound examples `snd_appendixD_vowels_*_sf.wav` are obtained by convolving a train of LF models and the VTF impulse response computed by the acoustic simulator.

D.2 Minimum-phase property of the vocal-tract

Here, the minimum-phase property of the synthesized VTF is discussed. In figure D.3, the phase of the acoustic model is shown in black lines for the four phonemes which can be nasalized. The minimum phase of each VTF can be computed from its amplitude spectrum using the real cepstrum (appendix A) (blues line in fig. D.3). One can see the following two elements on the difference between these two phases (red line in fig. D.3):

1) The difference has a principal distortion which is smooth. However, since this effect exist for non-nasalized vowels as well as for nasalized vowels, we assume that this distortion is made by the space sampling which creates a frequency aliasing (see the previous section).

2) Around the pole-zero pair made by the nasalization (∽ 500 Hz) the difference is almost linear-phase. However, if the zero of this pair was outside the unit circle, its phase contribution would be opposite to that of the minimum-phase version (remember that the peak of the group-delay of a zero inside the unit circle is negative while this peak is positive for a zero outside of the unit circle). Therefore, if the zero is outside the unit circle, a noticeable local distortion would exit around the frequency of this zero. Obviously, it's not the case and this zero is thus inside the unit circle.

In conclusion, it seems reasonable to assume that the VTFs synthesized by this simulator are minimum-phase.

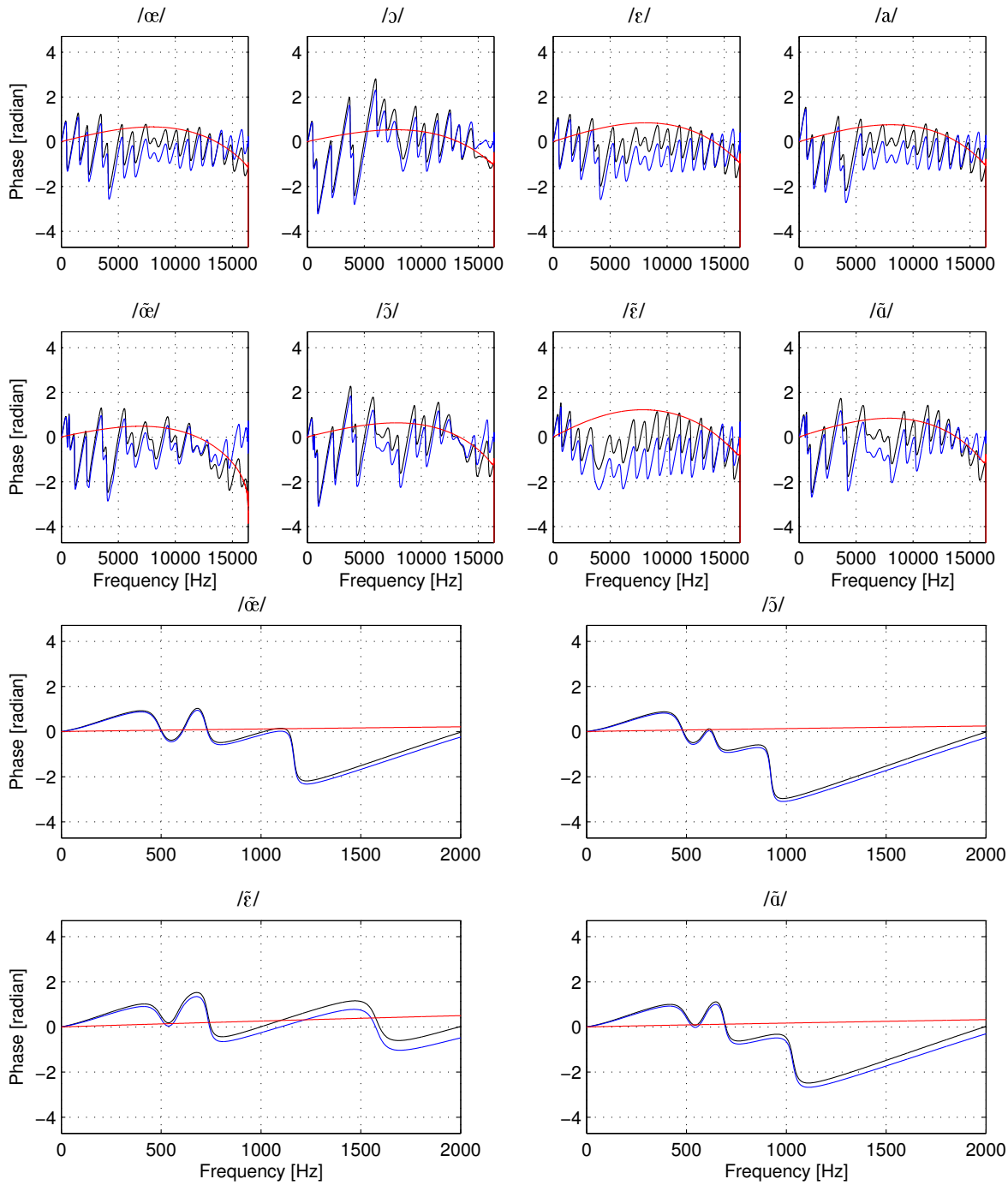


Figure D.3: Phase of the synthesized VTFs in black lines. The phase of the corresponding minimum-phase version in blue lines. The four plots below represent the first 2 kHz of the nasalized versions.

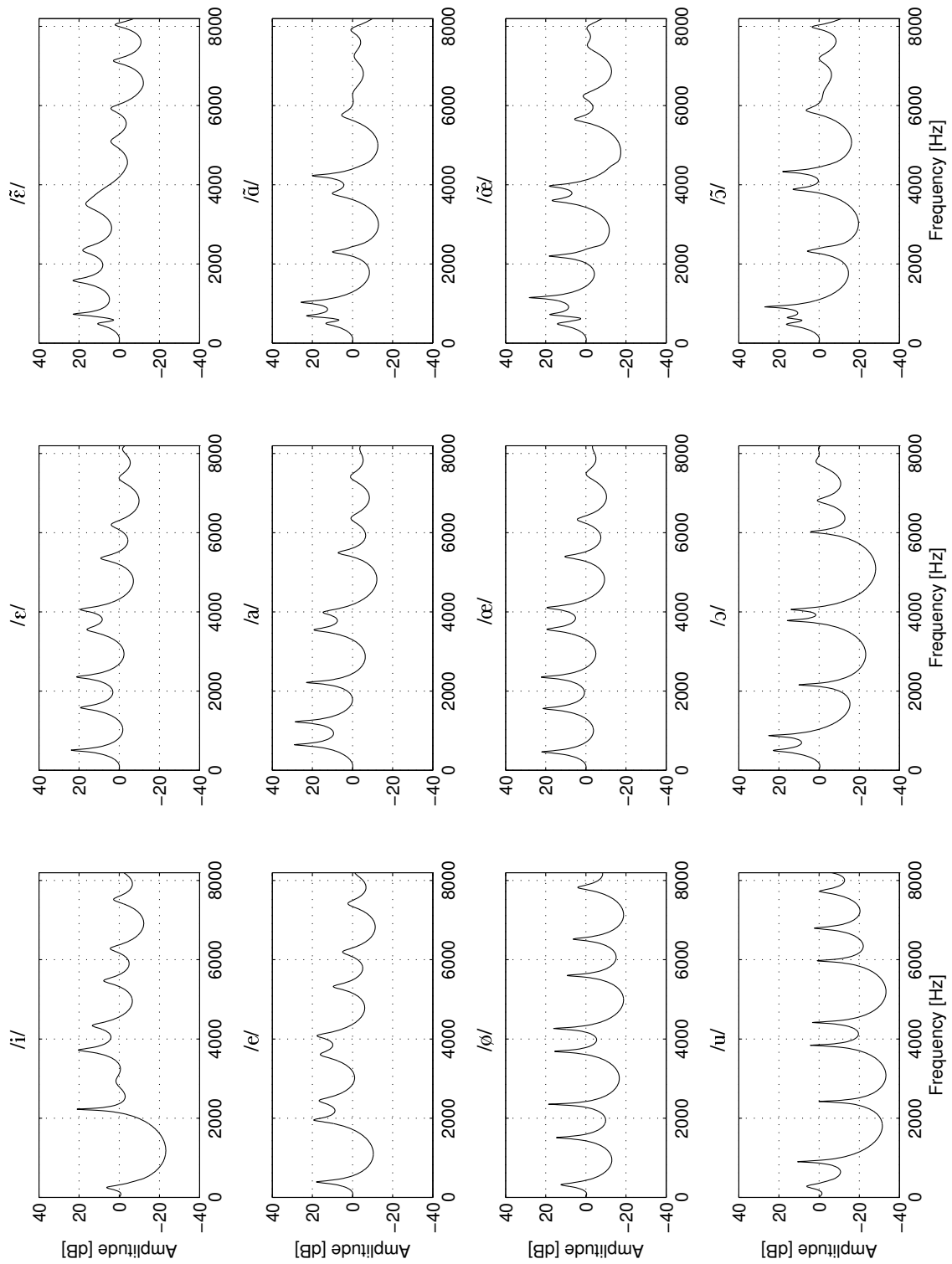


Figure D.4: Amplitude spectrum of the VTFs synthesized with the Maeda's simulator.

Bibliography

[AABP05] P. Alku, M. Airas, T. Backstrom, and H. Pulakka. Group delay function as a means to assess quality of glottal inverse filtering. In *Proc. Interspeech*, pages 1053–1056, 2005.

[AH71] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

[Air08] Matti Airas. *Methods and studies of laryngeal voice quality analysis in speech production.* PhD thesis, Helsinki University of Technology, Finland, 2008.

[Alk92] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[AM05] O. O. Akande and P. J. Murphy. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, 46(1):15–36, 2005.

[Ana84] T. V. Ananthapadmanabha. Acoustic analysis of voice source dynamics. *STL-QPSR*, 25(2-3):1–24, 1984.

[APBA05] M. Airas, H. Pulakka, T. Backstrom, and P. Alku. Toolkit for voice inverse filtering and parametrisation. In *Proc. Interspeech*, pages 2145–2148, 2005.

[AR82] B. S. Atal and J. R. Remde. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 614–617, 1982.

[AR08] Y. Agiomyrgiannakis and O. Rosec. Towards flexible speech coding for speech synthesis: an LF + modulated noise vocoder. In *Proc. Interspeech*, pages 1849–1852, 2008.

[AR09] Y. Agiomyrgiannakis and O. Rosec. ARX-LF-based source-filter methods for voice modification and transformation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3589–3592, 2009.

[ATN99] P. Alku, H. Tiitinen, and R. Naatanen. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology*, 110(8):1329–1333, 1999.

[AV94] P. Alku and E. Vilkman. Estimation of the glottal pulseform based on discrete all-pole modeling. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 1619–1622, 1994.

[AY79] T. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):309–319, 1979.

[BALA05] T. Backstrom, M. Airas, L. Lehto, and P. Alku. Objective quality measures for glottal inverse filtering of speech pressure signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 897–900, 2005.

[BDdD05] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12(4):344–347, 2005.

[BdH07] A. Barney, A. de Stefano, and N. Henrich. The effect of glottal opening on the acoustic response of the vocal tract. *Acta Acustica united with Acustica*, 93(6):1046–1056, 2007.

[Bec01] F. Bechet. Liaphon: un système complet de phonetisation de textes. *Traitement Automatique des Langues*, 42(1):47–67, 2001.

[BFH+61] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House. Reduction of speech spectra by anaylis-by-synthesis techniques. *Journal of the Acoustical Society of America*, 33(12):1725–1736, 1961.

[BLN+98] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara. Efficient representation of short-time phase based on group delay. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 861–864, 1998.

[Bon04] J. Bonada. High quality voice transformations based on modeling radiated voice pulses in frequency domain. In *Proc. Digital Audio Effects (DAFx)*, 2004.

[Bon08] Jordi Bonada. *Voice Processing and Synthesis by Performance Sampling and Spectral Models.* PhD thesis, Universitat Pompeu Fabra, Spain, 2008.

[Boz05] Baris Bozkurt. *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals.* PhD thesis, Faculté Polytechnique de Mons, Belgium, 2005.

[Bre73] R. P. Brent. *Algorithms for Minimization without derivatives.* Prentice-Hall, 1973.

[BRW10] M. A. Berezina, D. Rudoy, and P. J. Wolfe. Autoregressive modeling of voiced speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5042–5045, 2010.

[Cam07] Arturo Camacho. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music.* PhD thesis, University of Florida, USA, 2007.

[CC95] K. E. Cummings and M. A. Clements. Glottal models for digital speech processing: A historical survey and new results. *Digital Signal Processing*, 5(1):21–42, 1995.

[CL93] T. F. Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Control and Optimization*, 6(2):418–445, 1993.

[CL94] T. F. Coleman and Yuying Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Mathematical Programming*, 67(1-3):189–224, 1994.

[CL96] T.F. Coleman and Yuying Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996.

[CO89] Y. M. Cheng and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1805–1815, 1989.

[CW94] D.G. Childers and Chun-Fan Wong. Measuring and modeling vocal source-tract interaction. *IEEE Transactions on Biomedical Engineering*, 41(7):663–671, 1994.

[DBD09] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. Interspeech*, pages 116–119, 2009.

[DBR08] G. Degottex, E. Bianco, and X. Rodet. Usual to particular phonatory situations studied with high-speed videoendoscopy. In *Proc. International Conference on Voice Physiology and Biomechanics (ICVPB)*, pages 19–26, 2008.

[Dd97] B. Doval and C. d'Alessandro. Spectral correlates of glottal waveform models: an analytic study. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1295–1298, 1997.

[Dd99] B. Doval and C. d'Alessandro. The spectrum of glottal flow models. Technical Report NDL N 99-07, LIMSI, 1999.

[DdD97] B. Doval, C. d'Alessandro, and B. Diard. Spectral methods for voice source parameters estimation. In *Proc. Eurospeech*, pages 533–536, 1997.

[DdH03] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *Proc. ISCA Voice Quality: Functions, Analysis and Synthesis (VOQUAL)*, pages 16–20, 2003.

[DdH06] B. Doval, C. d'Alessandro, and N. Henrich. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006.

[DDM+08] T. Drugman, T. Dubuisson, A. Moinet, N. d'Alessandro, and T. Dutoit. Glottal source estimation robustness. In *Proc. International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, 2008.

[Del06] D. Deliyski. High-speed videoendoscopy: Recent progress and clinical prospects. In *Proc. Advances in Quantitative Laryngology (AQL)*, pages 1–16 vol. 7, 2006.

[Den05] Hui Qun Deng. *Estimations of glottal waves and vocalt-tract area functions from speech signals*. PhD thesis, University of British Columbia, Canada, 2005.

[dK02] A. de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[DP03] D. Deliyski and P. Petrushev. Methods for objective assessment of high-speed videoendoscopy. In *Proc. Advances in Quantitative Laryngology (AQL)*, pages 1–16, 2003.

[DR93] B. Doval and X. Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmms. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 221–224, 1993.

[DRR09a] G. Degottex, A. Roebel, and X. Rodet. Glottal closure instant detection from a glottal shape estimate. In *Proc. Conference on Speech and Computer (SPECOM)*, pages 226–231, 2009.

[DRR09b] G. Degottex, A. Roebel, and X. Rodet. Shape parameter estimate for a glottal model without time position. In *Proc. Conference on Speech and Computer (SPECOM)*, pages 345–349, 2009.

[Dud39] H. Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11(2):169–177, 1939.

[dY08] A. del Pozo and S. Young. The linear transformation of lf glottal waveforms for voice conversion. In *Proc. Interspeech*, pages 1457–1460, 2008.

[EJM91] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423, 1991.

[Eli09] Benjamin Elie. Characterisation of vocal tract acoustics in the case of oro-nasal coupling. Master's thesis, UPMC/UNSW, France/Australia, 2009.

[EMCB08] J. H. Esling, S. R. Moisik, and L. Crevier-Buchman. A biomechanical model of aryepiglottic trilling. In *Proc. International Conference on Voice Physiology and Biomechanics (ICVPB)*, pages 90–112, 2008.

[EMPSB96] S. El-Masri, X. Pelorson, P. Saguet, and P. Badin. Vocal tract acoustics using the transmission line matrix (TLM) method. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 953–956 vol. 2, 1996.

[Fan79] G. Fant. Vocal source analysis - a progress report. *STL-QPSR*, 20(3-4):31–53, 1979.

[Fan95] G. Fant. The LF-model revisited. transformations and frequency domain analysis. *STL-QPSR*, 36(2-3):119–156, 1995.

[Fer04] Raul Fernandez. *A Computational Model for the Automatic Recognition of Affect in Speech.* PhD thesis, Massachusetts Institute of Technology, USA, 2004.

[FG66] J. L. Flanagan and R. M. Golden. Phase vocoder. Technical report, The Bell System Technical Journal, 1966.

[FKLB94] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Bavegdrd. Voice source parameters in continuous speech, transformation of LF-parameters. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 1451–1454, 1994.

[FL71] O. Fujimura and J. Lindqvist. Sweep-tone measurements of vocal-tract characteristics. *Journal of the Acoustical Society of America*, 49(2B):541–558, 1971.

[FL86] H. Fujisaki and M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 1605–1608, 1986.

[FL88] G. Fant and Q. Lin. Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR*, 29(2-3):1–21, 1988.

[Fla57] J. L. Flanagan. Note on the design of "terminal-analog" speech synthesizers. *Journal of the Acoustical Society of America*, 29(2):306–310, 1957.

[Fla72a] J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer Verlag, 1972.

[Fla72b] J. L. Flanagan. Voices of men and machines. *Journal of the Acoustical Society of America*, 51(5A):1375–1387, 1972.

[FLL85] G. Fant, J. Liljencrants, and Q.-G. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985.

[FM06] Qiang Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):492–501, 2006.

[FMS01] M. Frohlich, D. Michaelis, and H. W. Strube. SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *Journal of the Acoustical Society of America*, 110(1):479–488, 2001.

[FMT99] K. Funaki, Y. Miyanaga, and K. Tochinai. Recursive ARMAX speech analysis based on a glottal source model with phase compensation. *Elsevier Signal Processing*, 74:279–295, 1999.

[FRR09] S. Farner, A. Roebel, and X. Rodet. Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications. In *Proc. Conference of the Audio Engineering Society (AES)*, 2009.

[Gal99] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3):272–281, 1999.

[GdR93] S. Grau, C. d'Alessandro, and G. Richard. A speech formant synthesizer based on harmonic + random formant-waveforms representations. In *Proc. Eurospeech*, pages 1697–1700, 1993.

[Gho99] S. S. Ghosh. VTCALCS for matlab v1.0. Technical report, Boston University, USA, 1999. http://speechlab.bu.edu/VTCalcs.php.

[GL88] D. W. Griffin and J. S. Lim. Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(8):1223–1235, 1988.

[GM74] Jr. Gray, A. and J. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 22(3):207–217, 1974.

[Gol62] A. S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369–375, 1962.

[GR90] T. Galas and X. Rodet. An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals. In *Proc. Computer Music (ICMC)*, 1990.

[Han95] Helen M. Hanson. *Glottal characteristics of female speakers.* PhD thesis, Harvard University, USA, 1995.

[Han97] H. M. Hanson. Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101(1):466–481, 1997.

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

[HDd99] N. Henrich, B. Doval, and C. d'Alessandro. Glottal open quotient estimation using linear prediction. In *Proc. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 12–17, 1999.

[HdD01] N. Henrich, C. d'Alessandro, and B. Doval. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In *Proc. Eurospeech*, pages 47–50, 2001.

[HdDC04] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America*, 115(3):1321–1332, 2004.

[Hed84] P Hedelin. A glottal LPC-vocoder. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 21–24, 1984.

[Hen01] Nathalie Henrich. *Etude de la source glottique en voix parlée et chantée.* PhD thesis, UPMC, in french, France, 2001.

[Her88] D. J. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

[Her91] D. J. Hermes. Synthesis of breathy vowels: Some research methods. *Speech Communication*, 10:497–502, 1991.

[HM95] J. W. Hawks and J. D. Miller. A formant bandwidth estimation procedure for vowel synthesis. *Journal of the Acoustical Society of America*, 97(2):1343–1344, 1995.

[HMC89] C. Hamon, E. Mouline, and F. Charpentier. A diphone synthesis system based on time-domain prosodic modifications of speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 238–241, 1989.

[IA79] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Electronics and Communication*, 62-A(4):10–17, 1979. in japanese.

[Jac89] L. B. Jackson. Noncausal arma modeling of voiced speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(10):1606–1608, 1989.

[JS05] P. Jinachitra and J. O. III Smith. Joint estimation of glottal source and vocal tract for vocal synthesis using kalman smoothing and em algorithm. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 327–330, 2005.

[Kaw97] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1303–1306, 1997.

[KB03] J. Kominek and A. W. Black. The CMU ARCTIC speech databases. In *Proc. ISCA Speech Synthesis Workshop*, pages 223–224, 2003. http://www.festvox.org/cmu_arctic.

[KdB+05] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi, and T. Irino. Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Proc. Interspeech*, pages 537–540, 2005.

[KH07] S.-J. Kim and M. Hahn. Two-band excitation for HMM-based speech synthesis. *IEICE - Transactions on Information and Systems*, E90-D(1):378–381, 2007.

[KK90] Dennis H. Klatt and Laura C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[KM79] I. Konvalinka and M. Matausek. Simultaneous estimation of poles and zeros in speech analysis and ITIF-iterative inverse filtering algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(5):485–492, 1979.

[KMKd99] H. Kawahara, I Masuda-Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207, 1999.

[KMT+08] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3933–3936, 2008.

[KNB02] A. Kounoudes, P. A. Naylor, and M. Brookes. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[KNZ97] H. S. Kim, H. Nolmes, and W. Zhang. Investigation on the spectral envelope estimator (SEEVOC) and refined pitch estimation based on the sinusoidal speech model. In *Proc. IEEE Speech and Image Technologies for Computing and Telecommunications*, volume 2, pages 575–578, 1997.

[KOT77] G. Kopec, A. Oppenheim, and J. Tribolet. Speech analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(1):40–49, 1977.

[LAB+07] L. Lehto, M. Airas, E. Bjorkner, J. Sundberg, and P. Alku. Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *Journal of Voice*, 21(2):138–50, 2007.

[LBN87] G. Lindsey, A. Breen, and S. Nevard. SPAR's archivable actual-word databases. Technical report, University College London, U.K., 1987.

[LD99a] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

[LD99b] J. Laroche and M. Dolson. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *IEEE Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 91–94, 1999.

[LDR10] P. Lanchantin, G. Degottex, and X. Rodet. A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4630–4633, Dallas, USA, 2010.

[LGS72] J. Lindqvist-Gauffin and J. Sundberg. Acoustic properties of the nasal tract. *STL-QPSR*, 13(1):13–17, 1972.

[Lil85] Johan Liljencrants. *Speech synthesis with a reflection-type line analog.* PhD thesis, KTH - Royal Institute of Technology, Sweden, 1985.

[Lju86] Mats Gunner Ljungvist. *Speech Analysis-Synthesis based on modeling of voice source and vocal-tract characteristics.* PhD thesis, University of Tokyo, Japan, 1986.

[LL93] Il-Taek Lim and Byeong Gi Lee. Lossless pole-zero modeling of speech signals. *IEEE Transactions on Speech and Audio Processing*, 1(3):269–276, 1993.

[LMRV08] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux. Automatic phoneme segmentation with relaxed textual constraints. In *Proc. Language Resources and Evaluation Conference*, pages 2403–2407, 2008.

[Lob01] A. P. Lobo. Glottal flow derivative modeling with the wavelet smoothed excitation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 861–864, 2001.

[LSM93] J. Laroche, Y. Stylianou, and E. Moulines. Hns: Speech modification based on a harmonic+noise model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 550–553, 1993.

[Lu02] Hui-Ling Lu. *Toward a High-quality Singing Synthesizer with Vocal Texture Control.* PhD thesis, Stanford University, USA, 2002.

[Mae79] Shinji Maeda. An articulatory model of the tongue based on a statistical analysis. *Journal of the Acoustical Society of America*, 65(S1):S22–S22, 1979.

[Mae82a] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3-4):199–229, 1982.

[Mae82b] S. Maeda. The role of the sinus cavities in the production of nasal vowels. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 7, pages 911–914, May 1982.

[Mak75] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[MC90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, 1990. Neuropeech '89.

[MG76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Springer Verlag, 1976.

[MHT00] Patrick Mergell, Hanspeter Herzel, and Ingo R. Titze. Irregular vocal-fold vibration—high-speed observation and modeling. *Journal of the Acoustical Society of America*, 108(6):2996–3002, 2000.

[Mil59] R. L. Miller. Nature of the vocal cord wave. *Journal of the Acoustical Society of America*, 31(6):667–677, 1959.

[Mil86] P. Milenkovic. Glottal inverse filtering by joint estimation of an ar system with a linear input model. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(1):28–42, 1986.

[MKW94] Changxue Ma, Y. Kamp, and L.F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994.

[MMEED61] M. V. Mathews, Joan E. Miller, and Jr. E. E. David. Pitch synchronous analysis of voiced sounds. *Journal of the Acoustical Society of America*, 33(2):179–186, 1961.

[MMN86] Y. Miyanaga, N. Miki, and N. Nagai. Adaptive identification of a time-varying ARMA speech model. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(3):423–433, 1986.

[MQ86] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

[MQ05] D. Mehta and T. F. Quatieri. Synthesis, analysis, and pitch modification of the breathy vowel. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 199–202, 2005.

[MT08] E. Moore and J. Torres. A performance assessment of objective measures for evaluating the quality of glottal waveform estimates. *Speech Communication*, 50(1):56–66, 2008.

[NKGB07] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):34–43, 2007.

[NKv05] Xiaochuan Niu, A. Kain, and J. P. H. van Santen. Estimation of the acoustic properties of the nasal tract during the production of nasalized vowels. In *Proc. Interspeech*, pages 1045–1048, 2005.

[NMEH01] J. Neubauer, P. Mergell, U. Eysholdt, and H. Herzel. Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes. *Journal of the Acoustical Society of America*, 110(6):3179–3192, 2001.

[Oli93] L. C. Oliveira. Estimation of source parameters by frequency analysis. In *Proc. Eurospeech*, pages 99–102, 1993.

[OS78] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing.* Prentice-Hall, 2nd edition, 1978. note that this edition contains a chapter about complex cepstrum which has been removed in the 3rd edition (and seems to come back in the 4th).

[OS89] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing.* Prentice-Hall, 3rd Edition, 1989.

[OSS68] A. Oppenheim, R. Schafer, and T. Stockham. Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56(8):1264–1291, 1968.

[PAD10] C.F. Pedersen, O. Andersen, and P. Dalsgaard. Separation of mixed phase signals by zeros of the z-transform - A reformulation of complex cepstrum based separation by causality. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5050–5053, 2010.

[Pau81] D. B. Paul. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(4):786–794, 1981.

[PB05] J. Perez and A. Bonafonte. Automatic voice-source parameterization of natural speech. In *Proc. Interspeech*, pages 1065–1068, 2005. Department of Signal Theory and Communication TALP Research Center Technical University of Catalonia (UPC), Barcelona, Spain javierp,antonio@gps.tsc.upc.edu.

[PB09] J. Perez and A. Bonafonte. Towards robust glottal source modeling. In *Proc. Interspeech*, pages 68–71, 2009.

[Pee01] G. Peeters. *Modeles et modification du signal sonore adaptees a ses caracteristiques locales.* PhD thesis, UPMC, in french, France, 2001.

[PHA94] X. Pelorson, A. Hirschberg, and Y. Auregn. Modelling of voiced sounds production using a modified two-mass model. *Journal de Physique*, 4(C5):453–456, 1994.

[Pie81] A. D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications.* McGraw-Hill, 1981.

[Por76] M. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):243–248, 1976.

[PQR99] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999.

[PRS08] Y. Pantazis, O. Rosec, and Y. Stylianou. On the properties of a time-varying quasi-harmonic model of speech. In *Proc. Interspeech*, pages 1044–1047, 2008.

[Pul05] H. Pulakka. Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. Master's thesis, Helsinki University of Technology, Finland, 2005.

[RHC09] B. Roubeau, N. Henrich, and M. Castellengo. Laryngeal vibratory mechanisms: The notion of vocal register revisited. *Journal of Voice*, 23(4):425–438, 2009.

[Roe10] Axel Roebel. A shape-invariant phase vocoder for speech transformation. In *Proc. Digital Audio Effects (DAFx)*, 2010.

[Ros71] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2B):583–590, 1971.

[Rot73] M. Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, 53(6):1632–1645, 1973.

[RPB84] X. Rodet, Y. Potard, and J.-B. BarriÃ¨re. The CHANT project: from synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31, 1984.

[RR05] A. Roebel and X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proc. Digital Audio Effects (DAFx)*, pages 30–35, 2005.

[RS78] Lawrence R. Rabiner and Ronald W. Schafer. *Digital processing of speech signals*. Prentice-Hall, 1978.

[RSP+08] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. HMM-based finnish text-to-speech system utilizing glottal inverse filtering. In *Proc. Interspeech*, pages 1881–1884, 2008.

[RVR07] A. Roebel, F. Villavicencio, and X. Rodet. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11):1343–1350, 2007.

[SA85] M. Schroeder and B. Atal. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 937–940, 1985.

[SA10] Yen-Liang Shue and A. Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5134–5137, 2010.

[SdD06] N. Sturmel, C. d'Alessandro, and B. Doval. A spectral method for estimation of the voice speed quotient and evaluation using electroglottography. In *Proc. Advances in Quantitative Laryngology (AQL)*, 2006.

[SdR09] N. Sturmel, C. d'Alessandro, and F. Rigaud. Glottal closure instant detection using lines of maximum amplitudes (loma) of thewavelet transform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4517–4520, 2009.

[SK00] J. Skoglund and W. B. Kleijn. On time-frequency masking in voiced speech. *IEEE Transactions on Speech and Audio Processing*, 8(4):361–369, 2000.

[SKA09] Y.-L Shue, J. Kreiman, and A. Alwan. A novel codebook search technique for estimating the open quotient. In *Proc. Interspeech*, pages 2895–2898, 2009.

[SKF53] K. N. Stevens, S. Kasowski, and C. Gunnar M. Fant. An electrical analog of the vocal tract. *Journal of the Acoustical Society of America*, 25(4):734–742, 1953.

[SL07] K. Schnell and A. Lacroix. Estimation of speech features of glottal excitation by nonlinear prediction. In *Proc. Non-Linear Speech Processing (NOLISP)*, pages 116–119, 2007. Institute of Applied Physics, Goethe-University Frankfurt Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany schnell@iap.uni-frankfurt.de.

[SL08] K. Schnell and A. Lacroix. Time-varying linear prediction for speech analysis and synthesis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3941–3944, 2008.

[SL09] K. Schnell and A. Lacroix. Iterative inverse filtering by lattice filters for time-varying analysis and synthesis of speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4017–4020, 2009.

[Son79] M. Sondhi. Estimation of vocal-tract areas: The need for acoustical measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(3):268–273, 1979.

[ST95] Brad H. Story and Ingo R. Titze. Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.

[STAV02] J. Sundberg, M. Thalen, P. Alku, and Erkki Vilkman. Estimating perceived phonatory pressedness in singing from flow glottograms. *TMH-QPSR*, 43(1):89–96, 2002.

[Ste60] Kenneth N. Stevens. Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 32(1):47–55, 1960.

[Ste71] Kenneth N. Stevens. Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50(4):1180–1192, 1971.

[Ste77] K. Steiglitz. On the simultaneous estimation of poles and zeros in speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(3):229–234, 1977.

[Str74] H. W. Strube. Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America*, 56(5):625–1629, 1974.

[Str98] Helmer Strik. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103(5):2659–2669, 1998. Inverse filter comes from Miller (1959); optimize parameters using dUg is better; do not consider the inverse problem, only the fitting.

[Sty96] Y. Stylianou. *Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification.* PhD thesis, TelecomParis, France, 1996.

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

[SY95] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, 1995.

[TAK+06] Hironori Takemoto, Seiji Adachi, Tatsuya Kitamura, Parham Mokhtari, and Kiyoshi Honda. Acoustic roles of the laryngeal cavity in vocal tract resonance. *Journal of the Acoustical Society of America*, 120(4):2228–2238, 2006.

[Tit08] Ingo R. Titze. Nonlinear source–filter coupling in phonation: Theory. *Journal of the Acoustical Society of America*, 123(5):2733–2749, 2008.

[TM03] M. Tooher and J. G. McKenna. Variation of the glottal LF parameters across f0, vowels, and phonetic environment. In *Proc. ISCA Voice Quality: Functions, Analysis and Synthesis (VOQUAL)*, pages 41–46, 2003.

[TMMK02] K. Tokuda, T. Masuko, N. Myizaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems*, E85-D:455–464, 2002.

[TMY+95] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Proc. Eurospeech*, pages 757–760, 1995.

[TN09] M.R.P. Thomas and P.A. Naylor. The sigma algorithm: A glottal activity detector for electroglottographic signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(8):1557–1566, nov. 2009.

[TZB02] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to english. In *Proc. IEEE Workshop on Speech synthesis*, 2002.

[van03] Ralph van Dinther. *Perceptual aspects of voice-source parameters.* PhD thesis, Technical University of Eindhoven, The Netherlands, 2003.

[VB85] D. Veeneman and S. BeMent. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):369–377, 1985.

[Vel98] R. Veldhuis. A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103(1):566–571, 1998.

[Vin07] Damien Vincent. *Analyse et controle du signal glottique en synthese de la parole.* PhD thesis, France Telecom, ENST, in french, France, 2007.

[VK06] J. Vepa and S. King. Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1763–1771, 2006.

[VMT92] H. Valbret, E. Moulines, and J.P. Tubach. Voice transformation using PSOLA technique. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 145–148, 1992.

[VRC05a] D. Vincent, O. Rosec, and T. Chonavel. Estimation du signal glottique basée sur un modèle ARX. In *Proc. Groupe d'Etude sur le Traitement du Signal et des Images (GRETSI), in french*, 2005.

[VRC05b] D. Vincent, O. Rosec, and T. Chonavel. Estimation of LF glottal source parameters based on an ARX model. *Proc. Interspeech*, pages 333–336, 2005.

[VRC06] Damien Vincent, Olivier Rosec, and Thierry Chonavel. Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 381–384, 2006.

[VRC07] D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 525–528, 2007.

[VRR06] F. Villavicencio, A. Robel, and X. Rodet. Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 869–872, 2006.

[Wei55] E. S. Weibel. Vowel synthesis by means of resonant circuits. *Journal of the Acoustical Society of America*, 27(5):858–865, 1955.

[WF78] H. Wakita and G. Fant. Toward a better vocal tract model. *STL-QPSR*, 19(1):9–29, 1978.

[WMG79] D. Wong, J. D. Markel, and A. H. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):350–355, 1979.

[Yeh08] Chunghsin Yeh. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, UPMC, France, 2008.

[Yos02] Takayoshi Yoshimura. *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems*. PhD thesis, Nagoya Institute of Technology, Japan, 2002.

[You94] S.J. Young. The HTK hidden markov model toolkit: Design and philosophy. Technical report, University of Cambridge, 1994.

[YR04] C. Yeh and A. Roebel. A new score function for joint evaluation of multiple f0 hypothesis. In *Proc. Digital Audio Effects (DAFx)*, pages 234–239, 2004.

[YTMK01] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura. Mixed-excitation for HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2259–2262, 2001.

[Yul27] G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical transactions of The Royal Society*, 1927.

[ZNY+07] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. ISCA Workshop on Speech Synthesis (SSW), http://hts.sp.nitech.ac.jp*, 2007.

[ZRR08] M. Zivanovic, A. Roebel, and X. Rodet. Adaptive threshold determination for spectral peak classification. *Computer Music Journal*, 32(2):57–67, 2008.

Publications during the study

Journal paper

G. Degottex, A. Roebel, and X. Rodet. *Phase minimization for glottal model estimation.* IEEE Transactions on Acoustics, Speech and Language Processing, accepted on August, 2010.

International conferences

G. Degottex, E. Bianco, and X. Rodet. Usual to particular phonatory situations studied with high-speed videoendoscopy. In *Proc. International Conference on Voice Physiology and Biomechanics (ICVPB)*, pages 19–26, Tempere, Finland, 2008.

G. Degottex, A. Roebel, and X. Rodet. Glottal closure instant detection from a glottal shape estimate. In Proc. *Proc. Conference on Speech and Computer (SPECOM)*, pages 226–231, St-Petersbourg, Russia, 2009.

G. Degottex, A. Roebel, and X. Rodet. Shape parameter estimate for a glottal model without time position. In Proc. *Proc. Conference on Speech and Computer (SPECOM)*, pages 345–349, St-Petersbourg, Russia, 2009.

G. Degottex, A. Roebel, and X. Rodet. Joint estimate of shape and time-synchronization of a glottal source model by phase flatness. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5058–5061, Dallas, USA, 2010.

P. Lanchantin, G. Degottex, and X. Rodet. A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4630–4633, Dallas, USA, 2010.

Seminars, workshops, talks

G. Degottex, A. Roebel, and X. Rodet. Transformation de la voix à l'aide d'un modèle de source glottique. At *Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS)*, Ircam, Paris, France, 2010.

G. Degottex and X. Rodet. Evolution de paramètres de modèle glottique et comparaisons avec signaux physiologiques. At *Summer school: Sciences et voix approche pluri-disciplinaire de la voix chantée*, Giens, France, 2009.

G. Degottex, A. Roebel, and X. Rodet. Estimation du filtre du conduit-vocal adaptée à un modèle d'excitation mixte pour la transformation et la synthèse de la voix. At *Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS)*, LMA, Marseille, France, 2009.

G. Degottex, E. Bianco, and X. Rodet. Estimation of glottal area with high-speed videoendoscopy. At *Speech Production Workshop: Instrumentation-based approach*, ILPGA, Paris, France, 2008.

G. Degottex, E. Bianco, and X. Rodet. Mesure de la source glottique par vidéoendoscopie à haute vitesse. At *Séminaire Recherche-Technologie*, Ircam, Paris, France, 2008.

E. Bianco, G. Degottex, and X. Rodet. Mécanismes vibratoires ou registres ? At *Congrès de la société française de phoniatrie*, Paris, France, 2008.

Contributions to a journal paper or a book chapter

G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin, X. Rodet IrcamCorpusTools : Plateforme Pour Les Corpus de Parole. *Traitement Automatique des Langues*, vol. 49, no. 3, pages 77–103, 2009.

X. Rodet with G. Beller, N. Bogaards, G. Degottex, S. Farner, P. Lanchantin, N. Obin, A. Röbel, C. Veaux and F. Villavicencio. Parole et musique, chapitre Transformation et synthèse de la voix parlée et de la voix chantée. *Odile Jacob*, 2009.