



ÉCOLE
POLYTECHNIQUE
DE BRUXELLES



UNIVERSITÉ LIBRE DE BRUXELLES

Analysis of vocal tremor in normophonic and dysphonic speakers

Thèse présentée en vue de l'obtention du grade
de docteur en Sciences de l'Ingénieur et Technologie

Christophe Mertens

Promoteur
Professeur Jean Schoentgen

Co-Promoteur
Professeur Francis Grenez

Service
Laboratory of Images, Signals and Acoustics (LISA)

Septembre 2015

Abstract

The study concerns the analysis of vocal cycle length perturbations in normophonic and dysphonic speakers.

A method for tracking cycle lengths in voiced speech is proposed. The speech cycles are detected via the saliences of the speech signal samples, defined as the length of the temporal interval over which a sample is a maximum. The tracking of the cycle lengths is based on a dynamic programming algorithm that does not request that the signal is locally periodic and the average period length known a priori.

The method is validated on a corpus of synthetic stimuli. The results show a good agreement between the extracted and the synthetic reference length time series. The method is able to track accurately low-frequency modulations and fast cycle-to-cycle perturbations of up to 10% and 4% respectively over the whole range of vocal frequencies. Robustness with regard to the background noise has also been tested. The results indicate that the tracking is reliable for signal-to-noise ratios higher than 15dB.

A method for analyzing the size of the cycle length perturbations as well as their frequency is proposed. The cycle length time series is decomposed into a sum of oscillating components by empirical mode decomposition the instantaneous envelopes and frequencies of which are obtained via AM-FM decomposition. Based on their average instantaneous frequencies, the empirical modes are then assigned to four categories (declination, physiological tremor, neurological tremor as well as cycle length jitter) and added within each. The within-category size of the cycle length perturbations is estimated via the standard deviation of the empirical mode sum divided by the average cycle length. The neurological tremor modulation frequency and bandwidth are obtained via the instantaneous frequencies and amplitudes of empirical modes in the neurological tremor category and summarized via a weighted instantaneous frequency probability density, compensating for the effects of mode mixing.

The method is applied to two corpora of vowels comprising 123 and 74 control and 456 and 205 Parkinson speaker recordings respectively. The results indicate that the neurological tremor modulation depth is statistically significantly higher for female Parkinson speakers than for female control speakers. Neurological tremor frequency differs statistically significantly between male and female speakers and increases statistically significantly for the pooled Parkinson speakers compared to the pooled control speakers. Finally, the average vocal frequency increases for male Parkinson speakers and decreases for female Parkinson speakers, compared to the control speakers.

This document has been written in L^AT_EX. The book template has been inspired by the *Legrand Orange Book* template, downloadable from www.LaTeXTemplates.com and under licence CC BY-NC-SA 3.0.

Chapter illustrations have been retrieved from different websites that are listed hereafter :

Chapter 1 (www.omnilexica.com), Chapter 2 (<http://www.iis.fraunhofer.de>), Chapter 3 (<http://agoras.typepad.fr>), Chapter 4 (<http://biomedical.egr.vcu.edu>), Chapter 5 (<https://wolfpaulus.com>), Chapter 6 (<http://worldcmcs.com/>), Chapter 7 (<http://www.forwallpaper.com>), Chapter 8 (<http://alecwren.com>), Chapter 9 (<http://tudogal.com>), Chapter 10 (<http://i.telegraph.co.uk>), Chapter 11 (<http://www.cloud-experience.fr>), Bibliography (<https://elgarblog.files.wordpress.com>) and other parts (<https://i1.wp.com/www.freeppt.net>).

To my valiant knight, Ethan, and
my lovely princess, Léa, . . .

This thesis was created at the heart of the Images, Signals and Acoustic department under the supervision of professors *Francis Grenez* and *Jean Schoentgen*. In addition to having exceptional personalities from a technical, social and cultural perspective, these two professors were able, throughout the years, to ally their strengths and form a complimentary duo which commands admiration and utmost respect. Many thanks to you for your support, insightful comments, and time spent towards the preparation of this thesis.

Thank you *Sabine Skodda*, *Alain Ghio*, and *François Viallet* for providing the Parkinson and control sample recordings. Thank you also *Pierre Mathys* and *Olivier Debeir*, as well as *Jorge Lucero* and *Antoine Nonclercq*, for accepting to be part of my dissertation committee and jury of this thesis.

A mention to all of the department members, current and past, thanks to whom my years at the university were stimulating and enriching. Thank you *Geoffrey Vanbienne* for logistical and amicable support throughout these years, as well as our good laughs. Thank you *Rudy Ercek* for precious advice in computing and networks, for giving me access at the ideal time to the computing power required for the simulations, and for the famous yearly harvest of plums. Thank you *Abdellah Kacha*, *Ali Alpan* and *Samia Fraj* for sharing the results of your research, and for our many conference adventures. I also wish the three of you success and satisfaction in your new roles. Thank you *Gatien Hocepiéd* for having placed your trust in me, and for allowing me to discover the many facets of our Alma Mater. Thank you *Nicolas Julémont* for the daily flyover news updates and many coffee break debates. Thank you *Laurent Lonys*, *Ammar Rouibah*, and *Agnès Sitchi*, with whom I shared many labs and exercises. Thank you *Laurent Lejeune*, *Xiaoya Fan*, *Adrien Debelle*, and *Federico Luccheti*, the newcomers, for various activities.

My gratitude also goes to my family and friends for their support and interest in my work. I'd especially like to thank my wife, *Luciné*, and my children, *Ethan* and *Léa*, for their unconditional support, their timeless patience, and their understanding regarding my absences these last weeks.

Finally, thanks to the students for their welcome. The questions you asked, and the projects on which we worked, allowed me to cover the subjects taught - and their applications - at angles from which would not have otherwise come to mind.

Contents

I	Background and literature review	
1	Introduction	3
1.1	Voice production	5
1.1.1	Anatomy	5
1.1.2	The nervous system	7
1.1.3	Principles of voice production	8
1.2	Clinical and functional assessment of pathological voices	10
1.3	Recording of a voice sample	12
1.3.1	Tasks	12
1.3.2	Segmentation	13
1.4	Analysis of vocal cycle length perturbations	13
1.4.1	Vocal jitter	14
1.4.2	Vocal tremor	14
1.5	Objectives and motivations	20
2	Cycle length tracking methods	23
2.1	Introduction	25
2.2	Short-term analysis	26
2.2.1	Temporal domain	26
2.2.2	Frequency domain	28
2.2.3	Model-based methods	31
2.3	Cycle-synchronous event analysis	33
2.3.1	Narrow-band temporal pattern analysis	33
2.3.2	Broad-band temporal pattern analysis	33

3	Sample salience analysis	39
3.1	Introduction	41
3.2	Definition	42
3.3	Salience allocation methods	43
3.3.1	Basic algorithm	43
3.3.2	Sliding analysis window	45
3.4	Reduction of computational load	50
3.5	Validation	52
3.5.1	Preliminary remarks	52
3.5.2	Theoretical developments	52
3.6	Example	55
3.6.1	Application to an arbitrary array	55
3.6.2	Application to a voiced speech sound	58
3.7	Conclusions	58
4	Tracking of cycle lengths via dynamic programming	61
4.1	Introduction	63
4.2	Overview	63
4.3	Problem formulation	63
4.3.1	Topology	63
4.3.2	Stages	64
4.3.3	States	65
4.3.4	Initialization	66
4.3.5	Optimal path search	67
4.3.6	Backtracking	67
4.4	Application to an example array	68
4.4.1	Initialization of the optimization network	68
4.4.2	Optimal path search	70
4.5	Conclusions	74
5	Application of the SCLT method to sustained speech sounds	77
5.1	Introduction	79
5.2	Preprocessing	80
5.3	Speech sample salience analysis	82
5.4	Cycle length tracking	84
5.5	The vocal cycle length time series	86
5.6	Conclusions	88
6	The wonderful story of the rolling wheel	91
6.1	Introduction	93
6.1.1	Geometry	93
6.1.2	Wheel in sustained motion	93

6.1.3	Wheel mechanism in sustained motion	94
6.2	Fourier analysis	95
6.2.1	The Fourier series and Fourier transform	95
6.2.2	Finite length time series and frequency resolution	96
6.2.3	Global analysis	98
6.3	Time-frequency analysis	100
6.4	Instantaneous frequency and amplitude	102
6.4.1	Instantaneous phase, amplitude and envelope	102
6.4.2	Instantaneous frequency	105
6.4.3	Relevance of the instantaneous values	106
6.4.4	Analysis of multi-component time series	107
7	Time-frequency analysis via Empirical mode decomposition . . .	109
7.1	Introduction	111
7.2	Empirical Mode Decomposition	111
7.2.1	Definition	111
7.2.2	Extraction of empirical modes	111
7.2.3	Sifting	113
7.3	Extraction of the instantaneous frequencies and envelopes	116
7.3.1	AM-FM decomposition	117
7.3.2	Computation of the instantaneous frequency of the empirical modes	119
7.4	Discussion	121
7.4.1	Preprocessing	121
7.4.2	Mode mixing : the empirical compromise	121
7.4.3	Conclusion	122
8	Analysis of vocal cycle length perturbations	125
8.1	Introduction	127
8.2	Time-frequency analysis of the vocal cycle length time series	127
8.3	Categorization	131
8.4	Average vocal frequency and cycle length perturbation size	135
8.4.1	Average vocal frequency	135
8.4.2	Perturbation sizes	135
8.5	Neurological tremor frequency	136
8.5.1	Overview	136
8.5.2	Complex neurological tremor time series	136
8.5.3	Neurological tremor frequency	137
8.5.4	Neurological tremor frequency estimate	139
8.5.5	Neurological tremor frequency content analysis	141
8.6	On the choice of the weights	149
8.7	Conclusions	154

9	Validation	159
9.1	Introduction	161
9.2	The synthesizer of disordered voices	161
9.2.1	Glottal source phase function perturbation model	161
9.2.2	Glottal airflow model	163
9.2.3	Propagation through the vocal tract	164
9.3	Tracking of cycle lengths	164
9.3.1	Overview	164
9.3.2	Corpus	164
9.3.3	Influence of the perturbation sizes	165
9.3.4	Influence of background noise	168
9.3.5	Comparison with Praat Software	172
9.4	Perturbation analysis	177
9.4.1	Corpus	177
9.4.2	Method	177
9.4.3	Results	178
9.5	Conclusions	183
10	Parkinson and control speakers	185
10.1	Introduction	187
10.2	Corpora	188
10.2.1	Corpus from Bochum University Clinic	188
10.2.2	Corpus from Pays d'Aix Hospital	190
10.3	Vocal cues	191
10.4	Analysis of the corpus from Bochum University Clinic	192
10.4.1	Quartiles and histograms of vocal cues	192
10.4.2	Correlation	192
10.4.3	Comparison between control and Parkinson speakers	194
10.5	Analysis of the corpus from Pays d'Aix Hospital	196
10.5.1	Quartiles and histograms of vocal cues	196
10.5.2	Correlation	196
10.5.3	Comparison between control and Parkinson speakers	198
10.5.4	Effects of the use of medication	200
10.6	Comparison between corpora	202
10.6.1	Quartiles and histograms of vocal cues	202
10.6.2	Correlations	204
10.6.3	Three-way variance analysis	204
10.6.4	Perturbation size	205
10.6.5	Neurological tremor frequency and bandwidth	207
10.6.6	Vocal frequency F_0	208
10.7	Discussion and conclusion	209
10.7.1	Discrepancies between corpora	209
10.7.2	Patient attributes	209
10.7.3	Statistically significant effects of corpus, gender or pathology	209

11	Conclusions & perspectives	213
11.1	Objectives and motivations	215
11.2	Key results	215
11.3	Improvements & perspectives	216
	Bibliography	219
	Index	225

List of Figures

1.1	Anatomy	6
1.2	Vocal fold structure	7
1.3	Nervous system	7
1.4	Nervous system loops	8
1.5	Laryngoscopic view and frontal section through the midportion of the glottis . . .	9
1.6	Subglottal and supraglottal part of the vocal apparatus	9
1.7	Modelling of the vocal tract by means of concatenated acoustic tubes	10
1.8	Analysis scheme	21
2.1	Cycle length tracking methods	25
2.2	Sustained vowel, autocorrelation function (ACF) and AMDF	27
2.3	Cepstrum of a voiced sound	28
2.4	Compressed spectra and harmonic product spectrum	29
2.5	Continuous wavelet transforms	30
2.6	Linear prediction and inverse filtering	31
2.7	Residual signal after linear prediction	32
2.8	Linear prediction of the decimated signal	32
2.9	Cycle tracking by threshold crossing	33
3.1	Signal observation	41
3.2	Topographic characterization of a mountain summit	41
3.3	Topographic salience	42
3.4	Salience definition	42
3.5	Basic algorithm for salience computation	44
3.6	Partial salience allocation	48
3.7	Global salience allocation (flow chart)	49
3.8	Salience allocation method (flow chart)	51
3.9	Running left and right salience values obtained for sample k	53
3.10	Running salience allocation for successive positions of the sliding window	56
3.11	Salience allocation : application to an example signal	57
3.12	Salience allocation : application to a voiced speech sound	58

4.1	Dynamic programming approach : Optimization network	64
4.2	Example : Two triplets and ending at peak i	64
4.3	Set of preceding triplets	65
4.4	Triplet sequence with increasing inter-peak distances	66
4.5	Initialization of triplet states	66
4.6	Example : Multistage interconnection network	70
4.7	Example : First experiment	71
4.8	Example : Second experiment	72
4.9	Example : Third experiment	73
5.1	Fragment of vowel (a) produced by 3 speakers	79
5.2	FIR filter characteristics : frequency response and impulse response	80
5.3	Band-pass filtered speech signals	81
5.4	Length N of the sliding analysis window used for salience allocation	82
5.5	Example : Saliences assigned to speech signal samples or peaks	83
5.6	Example : Results of the vocal cycle length tracking	85
5.7	Constant-step resampling	86
5.8	Example : Vocal cycle length time series	87
6.1	Geometry of wheel	93
6.2	Wheel in motion and temporal evolution of a point fixed to its circumference	94
6.3	Chain of two wheels in motion	95
6.4	Periodic time series and its Fourier coefficients	96
6.5	Observation window in temporal and frequency domains	97
6.6	Influence of the window duration on the Fourier spectrum	98
6.7	Fourier analysis applied to data obtained via (non-)stationary processes	99
6.8	Time-frequency analysis of a time series via STFT and CWT	101
6.9	Instantaneous frequency and amplitude : a practical illustration	102
6.10	Wheel in harmonic or non-harmonic motion	103
6.11	Instantaneous amplitude and phase time series	104
6.12	Determination of the instantaneous frequency function	105
6.13	Relevance of the instantaneous values	106
7.1	Empirical mode decomposition of a time series	112
7.2	Application of a sifting iteration to a signal	114
7.3	Boundary effect management by mirror symmetry	115
7.4	AM-FM decomposition of an empirical mode into a carrier and an envelope	118
7.5	Determination of the instantaneous phase function	119
7.6	Determination of the instantaneous frequency time series	120
7.7	Illustration of the mode mixing : Time series	121
7.8	Illustration of the mode mixing : Empirical mode decomposition	122
7.9	Illustration of the mode mixing : Instantaneous values	123
8.1	Control speaker : EMD and instantaneous values	128
8.2	Parkinson speaker : EMD and instantaneous values	129
8.3	Speaker with essential tremor : EMD and instantaneous values	130
8.4	Control speaker : Categorization	132
8.5	Parkinson speaker : Categorization	133
8.6	Speaker with essential tremor : Categorization	134
8.7	Complex empirical mode sum	137
8.8	Neurological tremor frequency : Weight computation	139
8.9	Estimation of the neurological tremor frequency	140
8.10	Neurological tremor frequency density estimation	142
8.11	Characterization of the neurological tremor frequency density	144
8.12	Scalar quantization of the neurological tremor frequency density	146
8.13	Time-varying representation of the neurological tremor frequency content	148

8.14	Weight formulation : time series	149
8.15	EMD & AM-FM decomposition applied to a sinusoidal time series	150
8.16	Frequency density estimates of a sinusoidal time series	151
8.17	EMD & AM-FM decomp. applied to a sinusoidal time series (with perturbation)	153
8.18	Frequency density estimates of a sinusoidal time series (with perturbation)	154
9.1	Synthesizer of disordered voices : Phase function perturbation	162
9.2	Synthesizer of disordered voices : Frequency response of tremor filter	162
9.3	Synthesizer of disordered voices : Jitter model	163
9.4	Synthesizer of disordered voices : Glottis model	164
9.5	Influence of perturbation size : Results of cycle length tracking	166
9.6	Influence of the low-frequency perturbation size	167
9.7	Influence of the jitter size	167
9.8	Illustration of noise addition	168
9.9	Cycle length tracking with additive noise	169
9.10	Influence of background noise (jitter only)	170
9.11	Influence of background noise (low-frequency perturbations only)	171
9.12	Comparison between <i>SCLT</i> and unconstrained Praat methods	173
9.13	Praat user interface (unconstrained method)	174
9.14	Praat user interface (constrained method)	175
9.15	Comparison between <i>SCLT</i> and constrained Praat methods	176
9.16	Linear regression model coefficients of the perturbation size cues	179
9.17	Linear regression model coef. for the perturbation size cues (with interaction)	180
9.18	Linear regression model coefficients for the neurological tremor frequency cues	180
9.19	Comparison of the neurological tremor frequency cue candidates	181
9.20	Linear regression model coefficients for the neurological tremor bandwidth cues	182
9.21	Tremor bandwidth corresponding to the resonator parameters of the synthesizer	183
10.1	Distribution of control speaker and patient attributes (Bochum)	189
10.2	Distribution of the control speaker and patient chronological ages (Aix)	190
10.3	Distributions of the vocal cues (Bochum)	193
10.4	Variations of vocal cue averages of control and Parkinson speakers (Bochum)	195
10.5	Distributions of the vocal cues (Aix)	197
10.6	Variations of vocal cue averages of control and Parkinson speakers (Aix)	199
10.7	Vocal cue averages of DOPA-ON and DOPA-OFF Parkinson speakers (Aix)	201
10.8	Distributions of the vocal cues (Pool)	203
10.9	Variations of perturbation size averages (corpora)	206
10.10	Variations of neur. tremor frequency and bandwidth averages (corpora)	207
10.11	Variations of vocal frequency averages (corpora)	208
10.12	Distribution of male, female, control and Parkinson speaker recordings	209

List of Tables

1.1	Conventional techniques used to assess the voice quality of a patient	12
1.2	Differences between types of recording devices	13
1.3	Chronological literature review	19
4.1	Example : Inter-peak distances	68
4.2	Example : Triplets, candidate predecessors and initial state values	69
4.3	Example : State values associated to the final triplets	70
9.1	Experimental conditions for the tracking validation	165
9.2	Experimental conditions for the perturbation analysis validation	177
9.3	Summary of vocal cues considered for perturbation analysis validation	178
10.1	Quartiles of the patient attributes (Bochum)	189
10.2	Quartiles of the control speaker and patient chronological ages (Aix)	190
10.3	Summary of retained vocal cues	191
10.4	Quartiles of the vocal cues (Bochum)	192
10.5	Correlation coefficients between the cues and patient attributes (Bochum)	194
10.6	Correlation coefficients between the cues and control speaker ages (Bochum)	194
10.7	Comparison between control and Parkinson speakers (Bochum)	194
10.8	Quartiles of the vocal cues (Aix)	196
10.9	Correlation coefficients between the cues and patient attributes (Aix)	198
10.10	Correlation coefficients between the cues and control speaker ages (Aix)	198
10.11	Comparison between control and Parkinson speakers (Aix)	198
10.12	Comparison between DOPA-ON and DOPA-OFF Parkinson speakers (Aix)	200
10.13	Quartiles of the vocal cues (Pool)	202
10.14	Correlation coefficients between vocal cues for Parkinson speakers (pool)	204
10.15	Correlation coefficients between vocal cues for control speakers (pool)	204
10.16	Comparison between the corpora, control and Parkinson speakers	205

List of Symbols

Sample salience analysis

$s(k), s_l(k), s_r(k)$	Total, left and right saliences (in samples, or s)	43
$s(k, i), s_l(k, i), s_r(k, i)$	Local total, left and right saliences (in samples, or s)	46
$s^*(k, i), s_l^*(k, i), s_r^*(k, i)$	Running total, left and right saliences (in samples, or s)	46
$s_f(k)$	Final total salience (in samples, or s)	46
$w_N(i)$	Analysis window (of length N) used for salience allocation	45

Salience-based cycle length tracking

α	Maximal expected local length perturbation	64
γ_1	Weight of the second-order differences of inter-peak durations	65
γ_2	Weight of the peak saliences	65
γ_3	Admissible preceding triplet selection	66
γ_4	Triplet sequence length weight	67
$\mu_{(g,h,i)}$	Average inter-peak distance of the triplet sequence (in samples, or s)	65
$c_{(f,g,h,i)}$	Absolute length perturbation increment of the triplet sequence	65
$C_{(g,h,i)}$	Overall length perturbation of the triplet sequence	65
$d_{(g,h)}$	Inter-peak distance (in samples, or s)	64
d_{min}, d_{max}	Minimal or maximal expected cycle length (in samples, or s)	64
$F_{0,min}, F_{0,max}$	Minimal or maximal expected vocal frequency (in Hz)	
$L_{(g,h,i)}$	Number of cycles in the triplet sequence	65
$T_{(g,h,i)}$	Set of admissible preceding triplets	65

Empirical mode decomposition & AM-FM decomposition

$\phi_i(t)$	Instantaneous mode phase (in rad)	116
θ_1, θ_2	Parameters involved in the sifting stopping test	116
$a_i(t)$	Instantaneous mode envelope	116
$c_i(t)$	Empirical mode, or intrinsic mode function	111
$f_i(t)$	Instantaneous mode frequency (in Hz)	116
$r(t)$	Residue of the empirical mode decomposition	111

Analysis of vocal cycle length perturbations

$\hat{F}(f;h)$	Neurological tremor frequency density estimate (in Hz)	141
$\hat{f}_{neur}(t)$	Estimated neurological tremor frequency time series (in Hz)	139
$ z_{neur}(t) $	Instantaneous neurological tremor envelope (in s)	137
$\phi_{neur}(t)$	Instantaneous neurological tremor phase (in rad)	137
$f_{neur}(t)$	Neurological tremor frequency time series (in Hz)	137
h	Gaussian kernel bandwidth	141
$w_i(t)$	Weight of the instantaneous mode frequency (in s)	138
$x_{int}(t)$	Intonation time series (in s)	131
$x_{jit}(t)$	Vocal jitter time series (in s)	131
$x_{neur}(t)$	Neurological tremor time series (in s)	131
$x_{phys}(t)$	Physiological tremor time series (in s)	131
$z_i(t)$	Complex empirical mode (in s)	137
$z_{neur}(t)$	Complex neurological tremor (in s)	137

Vocal cues

$\delta \hat{f}_{neur}$	Neurological tremor frequency variability (in Hz)	139
$\hat{b}_{dens,neur}$	Neurological tremor bandwidth (frequency density) (in Hz)	143
$\hat{b}_{quant,neur}$	Neurological tremor bandwidth (scalar quantization) (in Hz)	146
$\hat{f}_{\mu,neur}$	Neurological tremor frequency (temporal average) (in Hz)	139
$\hat{f}_{dens,neur}$	Neurological tremor frequency (frequency density) (in Hz)	143
$\hat{f}_{quant,neur}$	Neurological tremor frequency (scalar quantization) (in Hz)	146
σ_{jit}	Vocal jitter size	135
σ_{neur}	Neurological tremor depth	135
σ_{pert}	Total perturbation size	135
σ_{phys}	Physiological tremor depth	135
σ_{tre}	Total tremor depth	135
F_0	Average vocal frequency (in Hz)	135
T_0	Average cycle length (in s)	135

Synthesizer of disordered voices

Φ	Propagation delay between glottis entrance and exit (in rad)	163
$\phi_0(t)$	Glottal source phase function (in rad)	161
ξ_{entr}, ξ_{exit}	Vibration/abduction amplitudes of the vocal folds (in m)	163
A_g	Glottis area (in m^2)	164
d_{entr}, d_{exit}	Virtual glottal hemi-widths (in m)	163
$f_{0,trend}(t)$	Vocal frequency trend (in Hz)	163
L_g	Glottal length (in m)	164
$p(t)$	Total phase perturbation time series (in rad)	163
w_{entr}, w_{exit}	Glottal widths (in m)	163

Other symbols

F_s	Sampling frequency (in Hz)
T_s	Sampling period (in s)

Abbreviations

<i>ACF</i>	Autocorrelation function
<i>ALS</i>	Amyotrophic lateral sclerosis
<i>AMDF</i>	Average magnitude difference function
<i>AR</i>	Autoregressive
<i>CNS</i>	Central nervous system
<i>CV</i>	Coefficient of variation
<i>CWT</i>	Continuous wavelet transform
<i>DUR</i>	Duration since diagnosis
<i>EGG</i>	Electroglottography
<i>EMD</i>	Empirical mode decomposition
<i>EMG</i>	Electromyography
<i>ET</i>	Essential tremor
<i>HPS</i>	Harmonic product spectrum
<i>HY</i>	Hoehn and Yahr scale
<i>PD</i>	Parkinson's disease
<i>PNS</i>	Peripheral nervous system
<i>SCLT</i>	Salience-based tracking method
<i>STFT</i>	Short-term Fourier transform
<i>UPDRS</i>	Unified Parkinson's disease rating scale
<i>snr</i>	Signal-to-noise ratio (in dB)



Background and literature review

1	Introduction	3
1.1	Voice production	
1.2	Clinical and functional assessment of pathological voices	
1.3	Recording of a voice sample	
1.4	Analysis of vocal cycle length perturbations	
1.5	Objectives and motivations	
2	Cycle length tracking methods	23
2.1	Introduction	
2.2	Short-term analysis	
2.3	Cycle-synchronous event analysis	



1. Introduction

Objectives of this chapter

- Describe the mechanisms involved in voice production
- Give an overview of advanced voice function assessment techniques
- Define the objectives and motivations of the study

Contents

1.1	Voice production	5
1.1.1	Anatomy	5
1.1.2	The nervous system	7
1.1.3	Principles of voice production	8
1.2	Clinical and functional assessment of pathological voices	10
1.3	Recording of a voice sample	12
1.3.1	Tasks	12
1.3.2	Segmentation	13
1.4	Analysis of vocal cycle length perturbations	13
1.4.1	Vocal jitter	14
1.4.2	Vocal tremor	14
1.5	Objectives and motivations	20

1.1 Voice production

1.1.1 Anatomy

Voice production involves a complex and precise control by the central nervous system of a series of events in the peripheral phonatory organs. Some of the activity in the central nervous system is finally reflected in muscular activity of voice organ [Hir81]. Figure 1.1a illustrates a sagittal view of the human trachea, larynx, pharynx, mouth and nose cavities. Lungs included, all these play a role in the production of speech. Conventionally, they can be classified as follows :

- The sub-glottal part comprises the lungs and the trachea. During expiration, the pulmonary system supplies the aerodynamic power required for speech production.
- The glottal part is composed of the larynx and the vocal folds (Figures 1.1b and 1.1c). It constitutes the vibratory system required for voice production. Combining aerodynamic constraints and laryngeal muscular activity, the vocal folds vibrate to produce a pulsatile airflow propagating through the larynx. As a result, part of the aerodynamic power is converted into acoustical power.
- The supra-glottal part comprises the pharynx, mouth and nose cavities. It plays the role of resonator. By modifying the shape of the supraglottal apparatus, human speakers are able to modify vocal timbre and fix the timbre of the speech sounds.

As mentioned previously, the glottal part is crucial for voice production. The two vocal folds are located at the upper part of the trachea within the thyroid cartilage. They are constituted of soft tissue that is organized in different layers. These layers are conventionally considered as the *mucosa*, the *ligament* and the *muscle* [Hir81] [Tit93](Figure 1.2).

- The *mucosa* is composed of a flexible mucus membrane that forms the outer covering of the vocal fold (called *epithelium*), and the soft gelatinous superficial layer of the so-called *lamina propria*.
- The *ligament* comprises the intermediate (elastic fibers) and deep (collagenous fibers) layers of the *lamina propria*.
- The *muscle* consists of the *thyroarytenoid muscle* (or *vocalis muscle*)

Other intrinsic laryngeal muscles are involved in the vibratory system : the *cricothyroid*, *posterior cricoarytenoid*, *lateral cricoarytenoid*, *interarytenoid* muscles. When activated, these muscles modify the position, shape as well as mechanical property (elasticity, viscosity) of the vocal folds.

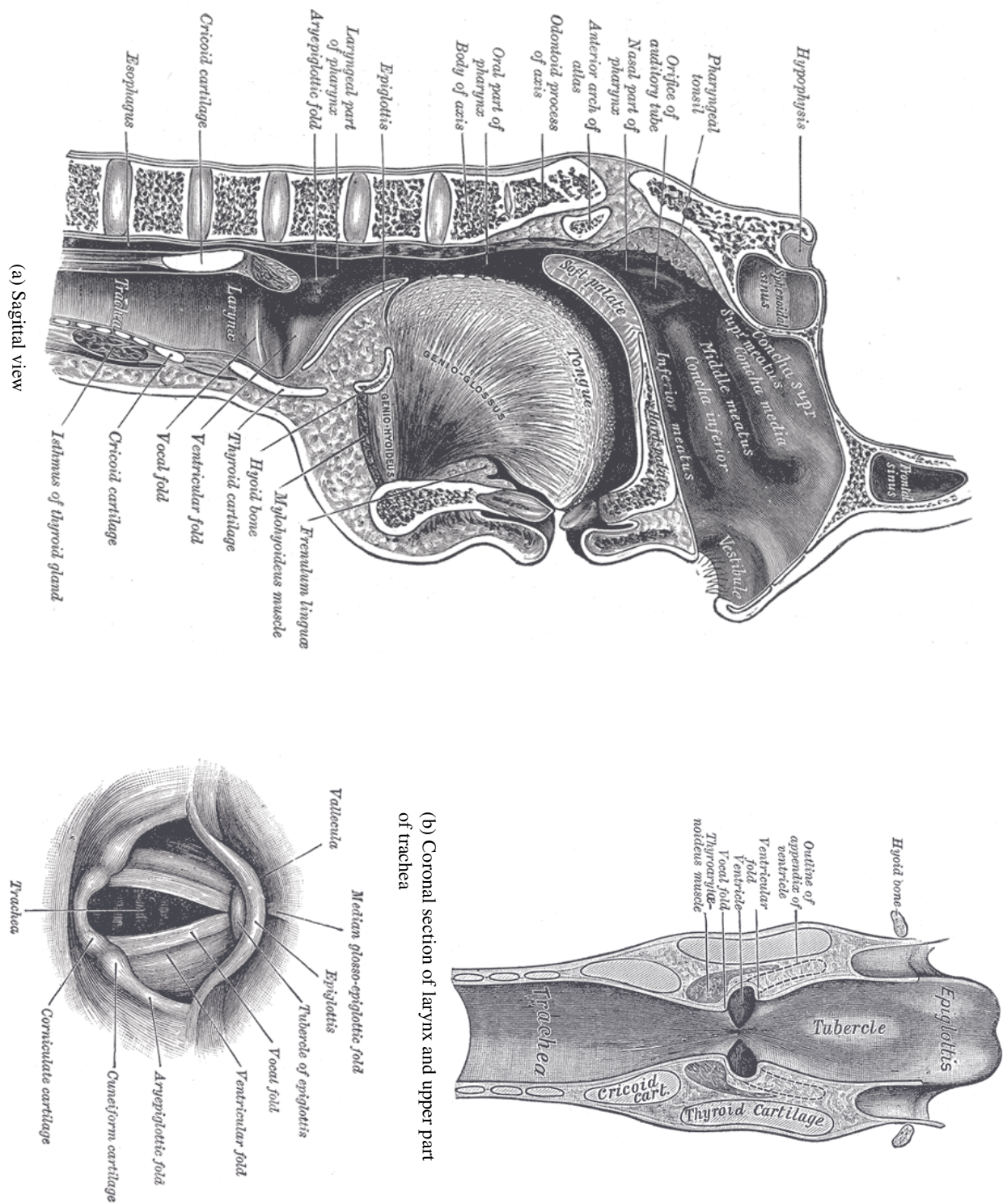


Figure 1.1 – Anatomy [HM18]

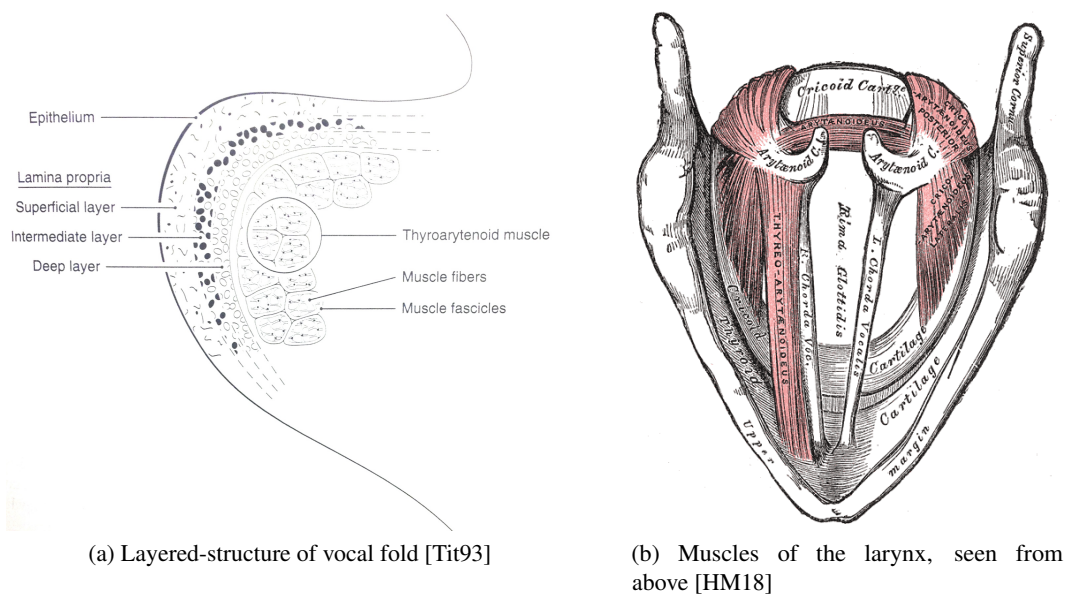


Figure 1.2 – Vocal fold structure

1.1.2 The nervous system

The nervous system coordinates voluntary and involuntary actions and transmits signals between different body parts. The nervous system comprises two main parts : the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS is composed of the brain, where information is processed, interpreted and stored, and the spinal cord which ensures the link between the brain and the peripheral nervous system. The PNS connects the central nervous system (CNS) to the body limbs and/or organs (Figure 1.3).

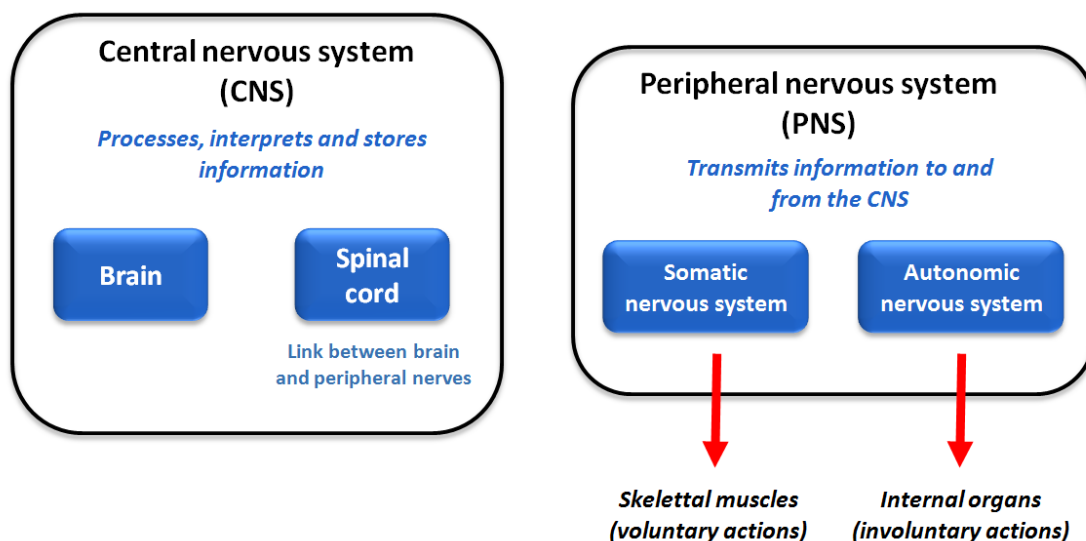


Figure 1.3 – Nervous system

The positioning or motion of a body part involves the activation of muscle fibers. For that, the central nervous system sends electrical impulses to the peripheral nervous system via individual or groups of specific nerve cells (called efferent motor neurons). Additionally, other kinds of nerve

cells (called afferent neurons) transmit information from the body sensors to the brain.

All these actors communicate with each other to produce an action, via several regulation loops. Three possible feedback loops may exist : the *short latency spinal reflex* arcs from afferent stretch receptors, the *long loop transcortical reflex* pathways from afferent stretch receptors or the *central feedback* from the motor neuron [MM00] (Figure 1.4). Because the goal of this report is not to explain in depth how the brain and the nervous system work, we won't delve into the details; however let us remind that neurological disorders may be considered as the consequence of a peripheral phenomenon limiting or impairing motor performance and/or dysfunction of the central nervous system.

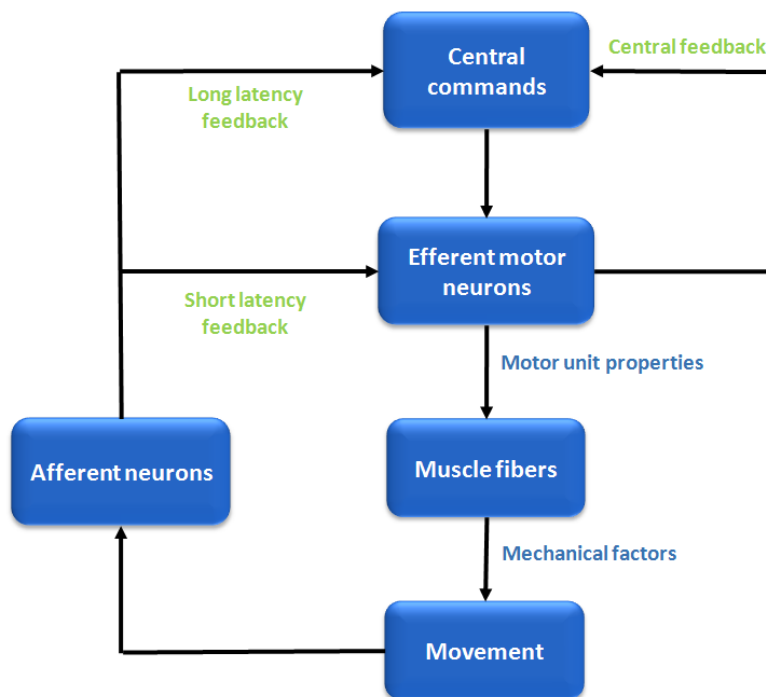


Figure 1.4 – Nervous system loops [MM00]

1.1.3 Principles of voice production

1.1.3.1 Glottal excitation

During breathing, the vocal folds are separated so that the glottal area is maximal. To produce a voiced speech sound, the vocal folds have to vibrate. For that, the laryngeal muscles (adduction of *arythenoid cartilages*, elongation of vocal cords via the *thyroid cartilage*) are activated to bring the vocal folds in contact. During expiration, the pulmonary system supplies the aerodynamic power required for voice production. As a result, an air flows through the trachea and glottis. Increasing lung pressure causes the pre-phonatory vocal fold position to become unstable when a pressure threshold is exceeded. The unstable pre-phonatory solution is then replaced by a self-sustained oscillation. Figure 1.5 illustrates the successive positions of the vocal folds during a vibration cycle.

To sum up, during phonation, the vocal folds convert the steady airflow coming from the lungs into a pulsatile airflow by cyclically opening and closing the glottis. This glottal flow provides the sound source for the excitation of the upper part of the vocal apparatus, also called the *vocal tract*. The bottom part of Figure 1.6 illustrates the temporal evolution of the glottal area and the resulting

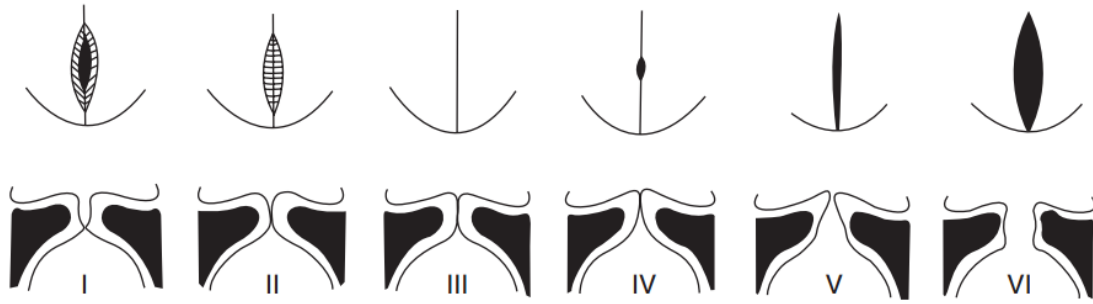


Figure 1.5 – Laryngoscopic view and frontal section through the midportion of the glottis. Point III corresponds to the maximal closure of the glottis [Dej10]

pulsated glottal airflow during phonation. These time series have been obtained via simulations by means of a synthesizer of disordered voices [Fra10].

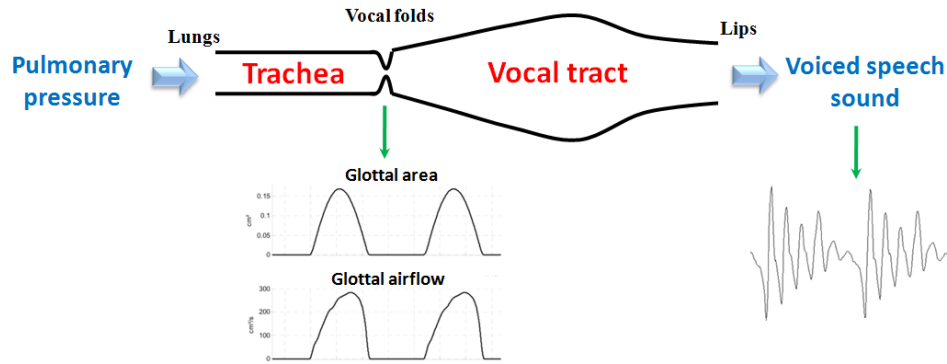


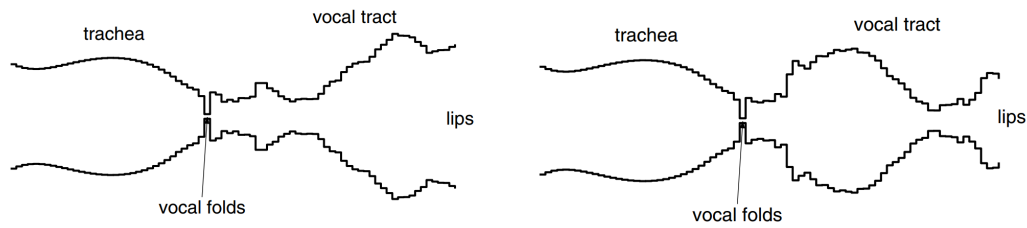
Figure 1.6 – Subglottal and supraglottal part of the vocal apparatus

1.1.3.2 Acoustic propagation through the vocal tract

The vocal tract is composed by the epi-larynx, pharynx, mouth and nose cavities. The production of a speech sound requires the coordination of the articulators via the central and peripheral nervous system. The perceived vocal timbre is directly related to the shape of the vocal tract which plays the role of resonator. The propagation of the acoustic waveform through the vocal tract modifies the temporal and spectral characteristics of sounds by amplifying/attenuating some frequency components. At the lips, the sound is radiated in the environment.

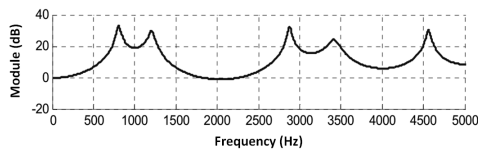
A conventional way, proposed in [KL62], to model the vocal tract consists in considering a concatenation of acoustic tubes with different sections and same length. As an example, Figure 1.7 illustrates the tubular modelling as well as the frequency response of the vocal tract for the production of sustained vowels [a] and [i]. One observes the effect of the shape of the vocal tract on the frequency response. One observes also that some frequency components are amplified. These components are called *formants* and are a distinguishing characteristics of the speech sounds. Notice that, for the sake of simplicity, the nasal cavity is often not considered.

The study of the acoustic propagation involves the resolution of equations that are obtained from Navier-Stokes equations (conservation of mass, energy, ...). Ad hoc modelling techniques

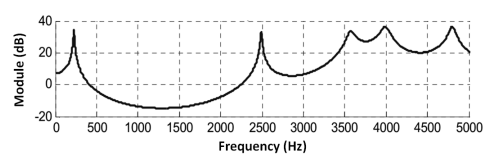


(a) Tubular representation of the subglottal and supraglottal tracts. In this case, the supraglottal tract is configured for a /a/ vowel [Sto02]

(b) Tubular representation of the subglottal and supraglottal tracts. In this case, the supraglottal tract is configured for a /i/ vowel [Sto02]



(c) Frequency response of the vocal tract for a vowel /a/ [Fra10]



(d) Frequency response of the vocal tract for a vowel /i/ [Fra10]

Figure 1.7 – Modelling of the vocal tract by means of concatenated acoustic tubes as well as the frequency response of the vocal tract for the production of sustained vowels /a/ and /i/

consider additional factors to describe the sound propagation (time-varying geometry, radiation at the lips, losses due to heat conduction and friction in the vocal tract walls, nasal coupling,...

1.2 Clinical and functional assessment of pathological voices

In clinical application of speech analysis, several techniques are used to assess and monitor the voice quality of patient during treatment. The assessment of voice and laryngeal functions is based on auditory-perceptual ratings, self-evaluation questionnaires, aerodynamic measurements, vocal folds imaging as well as acoustic analyses of speech sounds. Table 1.1 gives an overview.

Acoustic feature-based assessment methods are popular because they are non-invasive and enable clinicians to monitor the voice of patients numerically.

Techniques & Description	Objectives
--------------------------	------------

Aerodynamics : Measurement of glottal airflow and pressure as well as their relationship during phonation to assess the efficiency of phonation.

Pneumotachograph : A device including a fine mesh wire screen that creates a resistance to airflow is placed, during phonation, in the region of the patient's mouth.	Measurement of the mean airflow rate
Flowglottograph : A mask, including a high-frequency pressure transducer, produces, by removing the resonances of the vocal tract, a time-varying estimate of the glottal airflow.	Measurement of the leakage airflow (DC component appearing in case of partial glottal closure) and the pulsated airflow.
Spirometer : Device that measures over time the maximal forced expiratory and inspiratory airflow rates.	The resulted flow-volume graph enables the investigation of voice problems that are related to laryngeal obstructions.

Vocal folds imaging : These invasive techniques enable the investigation and understanding of the vocal fold vibratory dynamics.

Videolaryngostroboscopy : An endoscope as well as a strobe light are inserted into the vocal tract to observe the vibrations of vocal folds in "slow motion".	Investigation of the glottal closure, the glottal cycle regularity, the symmetry of vocal folds as well as the mucosal wave.
Digital high speed pictures : A camera, inserted into the vocal tract, captures the vibrations of vocal folds and stores images at a rate of 2000 images/s or more.	Enables the analysis and understanding of the vibration dynamics with high temporal resolution.
Videokymography : A modified video camera, inserted into the vocal tract, selects a single horizontal line from the whole image of the glottis and monitors it at a rate of $\approx 8kHz$	The concatenation of these line images provides relevant information for comparing the vibration amplitude/mucosal wave of both folds over time.

Other techniques

Electroglottography (EGG) : Two electrodes are placed on the speaker's neck in the vicinity of the thyroid cartilage measuring over time the electrical impedance during phonation.	Enables the monitoring of vocal fold contact (low impedance value), the rate of vibration and the perturbation of regularity during voice production.
Electromyography (EMG) : An electrode is inserted in a laryngeal muscle (often the <i>thyroarytenoid</i> and <i>cricothyroid muscles</i>) recording the muscular activity during phonation.	Enables the electrophysiologic investigation of the neuromuscular function by evaluating the integrity of the motor system and recording action potentials generated in the muscle fibres.

Acoustic analysis

Supra-glottal/glottal coordination cues	Example : voice onset time
Prosodic or supra-segmental features	They characterize the intonation, fluidity and rhythm of speech.
Morphological features	They report the shape of the glottal excitation signal, e.g. open quotient, speed quotient.
Vocal dysperiodicity cues	These vocal cues describe the irregularities of the glottal cycles.
Continued on next page	

Continued from previous page	
Techniques & Description	Objectives
Perceptual assessment : The voice quality of patient is assessed via a standardized auditory-perceptual procedure. The rating is made on conversational speech, on a phonetically balanced read text or sustained vowels.	
GRBAS scale : A popular scale that comprises 5 factors (Grade (G), Roughness (R), Breathiness (B), As-thenicity (A), Strain (S)) assessed via 4 degrees of severity (0,1,2,3)	That scale rates the overall voice quality (G), the irregularity of vocal fold vibrations (R), the audible air leakage through an insufficient glottal closure (B), the weakness of the voice (A) as well as hyperfunction (S).
Subjective evaluation by the patient	
Standardized questionnaires like the <i>Voice Handicap Index (VHI)</i> or the <i>Unified Parkinson Disease Rating Scale (UPDRS)</i>	The purpose of the subjective self-evaluation is to determine the deviance of voice quality and the severity of disability or handicap in daily professional and social life and the possible emotional repercussions of the dysphonia.

Table 1.1 – Conventional techniques used to assess the voice quality of patient [Ann+10]

1.3 Recording of a voice sample

A speech sound fragment is acquired and stored via a recording device. Table 1.2 provides a brief comparison of three frequently used devices. (conventional aerial microphone, throat microphone and accelerometer). The acoustic microphone records directly the signal which is radiated at the lips and the other two devices record an acoustic or mechanical signal that has been propagated through soft and hard body parts. As a consequence, the characteristics of these signals as well as their robustness with regard to background noise differ widely.

During acquisition, attention must be paid to recording conditions, the quality of which has a direct impact on the further analysis :

- the recording room has to be selected to reduce the background noise (babble noise, fan noise, ...)
- When an aerial microphone is used, an unidirectional device (sensitive to sounds from only one direction) placed at a small and constant distance from the mouth and not aligned with the mouth axis is recommended to maintain a high signal-to-noise ratio and reduce aerodynamic noise generated by the airflow.
- the recording device gain has to be selected high enough to reduce quantization error but to avoid signal clipping.
- a sampling frequency higher than $20kHz$ is recommended to guarantee a good temporal resolution.

1.3.1 Tasks

The acoustic feature-based assessment of voice quality requires the analysis of speech sounds produced by the patient. Conventionally, two major tasks are considered :

- *Sustained voiced speech sounds*, produced by the patient during a short or the longest possible duration. Here, the analysis is focused on glottal excitation. Indeed, the shape of the vocal




	Acoustic microphone	Throat microphone	Accelerometer
			
Transducer			
• type	condenser	moving coil	piezoelectric
• supply	✓ (phantom 48V)	✓ (5V)	✗
• pre-amplifier	✗	✗	✓
• frequency range	below 20kHz	below 5kHz	below 5kHz
Speech signal acquisition			
• Propagation through the neck wall	✗	✓	✓ (stuck to skin)
• Propagation through the nose wall	✗	✗	✓ (stuck to skin)
• Radiation at the lips	✓	✗	✗
Signal characteristics			
• shaped by the vocal tract	✓	✗	(depends on use)
• strong modulation by heart beat and breathing	✗	✓	✗
• tainted by environmental noise	✓	✗	✗

Table 1.2 – Differences between types of recording devices : acoustic microphone (*Sennheiser, HS-2*), throat microphone (*Clearer Communications, Stryker PC*) and accelerometer (*K&K, Twin Spot Classic*)

tract is approximatively kept unchanged and the glottal source dynamics may be investigated via morphological features or vocal dysperiodicity cues.

- *Connected speech*, recorded during spontaneous speech production or reading of a phonetically balanced text [Com20], enables the investigation of articulation, glottal, supra-glottal coordination, prosody as well as vocal dysperiodicities.

In this study, only sustained voiced speech sounds are considered.

1.3.2 Segmentation

Segmentation consists in selecting a fragment of recorded speech to be analyzed. Fragments that are deteriorated by recording conditions or not related to the speech production tasks are discarded. This may be done manually by visual inspection and/or auditory perception, or automatically by means of speech/voice activity detection algorithms. In this study, a user-friendly interface has been developed to select manually fragments of interest.

1.4 Analysis of vocal cycle length perturbations

Here, the investigation of vocal dysperiodicities is limited to voice analysis rather general speech analysis. The voice of normal speakers is characterized by small perturbations of the amplitude or duration of the vocal cycles. These perturbations contribute to perceived timbre and are related to muscle activity or neurological activity as well as aerodynamic noise and distribution of mucus.

Disorders of phonation are often a consequence of the inability of vocal folds to vibrate regularly. Larger than normal disturbances of the periodicity of the glottal source signal are therefore observed frequently as a consequence of organic or functional disorders of the larynx. These

disorders, interfering with the pseudo-regularity of vocal cord vibrations, include modifications of the shape and/or composition of vocal folds (nodules, polyps,...), variations in the function or control of the muscles involved in the vibrations. As a consequence, the perturbations of the voice may increase because of an alteration of the response of the glottal vibrator to outside perturbations that are in the normal range or an increase of the size of these perturbations.

1.4.1 Vocal jitter

Vocal jitter is a measurement of irregularities in the vibration of the vocal folds and designates small, fast and involuntary cycle-to-cycle perturbations of vocal cycle lengths. The size of vocal jitter in modal voice is expected to be $< 1\%$ of the typical cycle length. However, various conditions may affect the vocal cords and their vibration dynamics (for instance nodules, polyps and weakness of the laryngeal muscles).

1.4.2 Vocal tremor

1.4.2.1 Definition

Tremor designates a movement disorder that is characterized by small involuntary oscillatory movements of a body part [DBB98]. It occurs in different body parts and is mainly observed at the extremities. Based on their specific clinical features, tremors are grouped into syndromes that can be separated on the basis of clinical observations alone [DBB98]. Several types of tremor are therefore identified in the literature and are often classified according to the range of involved muscles and its association with specific movement conditions. Particularly, *rest tremor* is defined by tremor that occurs in a body part that is not voluntarily activated. On the other hand, *action tremor* is any tremor that is observed during voluntary muscle contraction while performing daily activities.

Vocal tremor designates involuntary low-frequency modulations of the vocal cycle lengths and is also a movement disorder that is characterized by involuntary activation of laryngeal muscles involved in voice production.

1.4.2.2 Causes

In the literature, several possible sources of tremor are identified [GM12] and summarized here below :

- *Mechanical oscillations* resulting from the positioning or motion of body effectors via the activation of antagonistic/agonistic muscle fibers or a group of muscles. Tremor may occur if the involved muscle fibers are activated at a rate close to the resonance frequency of the body parts, depending on the shape and mechanical properties of these parts.
- *Reflexive muscle activation* : reflex oscillators arising from sensory feedback pathways occur if, during the activation of an agonistic muscle, stretch receptors afferents elicit a reflexive activation of the antagonistic muscle, and inversely (*short term latency spinal reflex* and *long loop transcortical reflex*, see section 1.1.2).
- *Central oscillation* : modulation of motor unit activity occurs within individual neuron or a population of neurons and are related to dysfunctional electrical transmission and/or misinterpretation of sensory and motor functions (*central feedback*, see section 1.1.2).

1.4.2.3 Pathophysiologic entities of tremor

Often, vocal tremor is present as a symptom of one of the three following pathophysiologic entities : *essential tremor* (ET), *spasmodic dysphonia*, or generalized neurologic disease such as *Parkinson's disease* (PD), *amyotrophic lateral sclerosis* (ALS), or multiple sclerosis [GM12].

- *Essential tremor* (ET) is an *action tremor* disorder. The term "essential" refers in medical context to isolated symptoms with no known specific underlying causes. Tremor severity can vary based on the activity being performed, the position of the body part, and the presence of stress or fatigue [Fou15].
- *Parkinson's disease* (PD) is a degenerative disorder of the central nervous system that causes a gradual loss of muscle control. The underlying cause of Parkinson's disease is a dysfunction of a small area in the brain stem called the *substantia nigra* that controls movement. Cells that are located in the *substantia nigra* stop making *dopamine*, a brain chemical that helps nerve cells to communicate. As a consequence, the brain does not receive the necessary messages. The symptoms of Parkinson's disease are shaking, rigidity and slowness of limb motion. During the initial stages of the tremor disorder, it is often difficult to differentiate essential tremor from Parkinson's disease. However, this kind of tremor appears during rest and thus differs from the essential tremor.
- *Amyotrophic lateral sclerosis* (ALS) is a rapidly progressive and fatal neurological disease that attacks the nerve cells responsible for controlling voluntary muscles. The earliest symptoms may include fasciculations, cramps, tight and stiff muscles (spasticity), muscle weakness affecting an arm or a leg, slurred and nasal speech, or difficulty chewing or swallowing. [NS].
- *Spasmodic dysphonia* (or laryngeal dystonia) is an action disorder that causes muscles to contract and spasm involuntarily. Following Parkinson's disease and essential tremor, dystonia is the third most common movement disorder. [Ass]

In this study, vocal tremor is analyzed for patients suffering from Parkinson's disease. Possible vocal symptoms of the disease are vocal frequency tremor and hoarseness [Cno+08]. In a study based on a large sample of patients with Parkinson disease, it has been reported that between 70% and 90% of patients have problems related to speech and voice impairments [Log+78].

1.4.2.4 Previous research on vocal tremor

(Ludlow et al., 1986) [Lud+86]	
Corpus :	Vowels [a] sustained by 9 speakers with vocal tremor and 20 control speakers.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : n/a ◦ <i>Analysis</i> : Estimation of the modulation frequency and depth by means of heuristics.
Cues :	<ul style="list-style-type: none"> ◦ Modulation frequency of F_0 and the speech signal envelope, as well as their percentages of variations relative to vocal jitter.
Results :	<ul style="list-style-type: none"> ◦ Vocal tremor affects the variations of the signal envelope as well as the variations of the vocal frequency. ◦ Parkinson speakers are characterized by larger variations than control speakers.
(Yair and Gath, 1988) [YG88]	
Corpus :	Vowels [a] sustained by 9 Parkinson speakers and 3 control speakers.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Cycle length sequence obtained via short-term analysis based on autocorrelation ◦ <i>Analysis</i> : Modulation power spectrum computed by means of a point process model of the cycle length sequence.
Cues :	<ul style="list-style-type: none"> ◦ Modulation frequency : frequency of the main power spectrum peak. ◦ Modulation depth : Energy of the spectrum in the frequency interval $[3.5Hz, 7.5Hz]$.
Results :	<ul style="list-style-type: none"> ◦ Narrow-band modulation frequency f ($4Hz \leq f \leq 6Hz$) for Parkinson speakers ◦ Vocal tremor frequency f correlated with the tremor frequency of limbs. ◦ Correlation between vocal tremor and clinically-assessed limb tremor (in frequency and depth).
(Ackermann and Ziegler, 1991) [AZ91]	
Corpus :	Sustained vowels and unvoiced fricatives produced by 1 female speaker suffering from cerebellar voice tremor.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : n/a ◦ <i>Analysis</i> : Visual inspection of the F_0 trace and its spectrum.
Results :	<ul style="list-style-type: none"> ◦ Intermittent and rhythmic oscillations at a frequency $\approx 3Hz$.
(Winholtz and Ramig, 1992) [WR92]	
Corpus :	Vowels [a] sustained by 12 control speakers, 12 patients with vocal tremor and 12 singers producing vibrato.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Based on a vocal demodulator that produces amplitude and frequency demodulated outputs from a voice recording. ◦ <i>Analysis</i> : Based on the spectrum (FFT) of the F_0 trace in the frequency interval $[2.5Hz, 25Hz]$.
Cues :	<ul style="list-style-type: none"> ◦ Low-frequency modulation frequency and depth.
Results :	<ul style="list-style-type: none"> ◦ Modulation frequencies do not differ statistically significantly between groups of speakers. ◦ Patients with vocal tremor have larger modulation depths than the control speakers. ◦ Modulation frequency range : control speakers (σ: $[5.5Hz, 8Hz]$, φ: $[4.9Hz, 6.1Hz]$), patients (σ: $[4Hz, 6.5Hz]$, φ: $[5Hz, 6.5Hz]$) ◦ Modulation depth range : control speakers (σ: $[0.9\%, 1.8\%]$, φ: $[0.8\%, 1.3\%]$), patients (σ: $[3.4\%, 6\%]$, φ: $[4.3\%, 10.7\%]$).
(Aronson et al., 1992) [Aro+92]	
Corpus :	8 patients suffering from ALS and 8 control speakers (paired in age and gender).
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Based on the vocal demodulator developed by Winholtz and Ramig. ◦ <i>Analysis</i> : Based on the spectrum (FFT) of the F_0 trace.
Cues :	<ul style="list-style-type: none"> ◦ Frequency and amplitude of the spectral peaks.
Results :	<ul style="list-style-type: none"> ◦ Speaker-dependent modulation depth and frequency. ◦ No single prominent spectral peaks in patients. ◦ Patients are characterized by larger modulation depths.
Continued on next page	

Continued from previous page

(Hirose et al., 1995) [Hir95]

Corpus :	Vowel [a] sustained by 12 Parkinson speakers and 51 control speakers.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Detection of prominent cycle peaks. ◦ <i>Analysis</i> : Based on the spectrum (FFT) of the F_0 trace.
Cues :	◦ Energy in the interval $[0.1Hz, 16Hz]$, normalized by the average F_0 .
Results :	◦ Parkinson speakers are characterized by larger perturbation levels than control speakers.

(Dromey et al., 2002) [DWI02]

Corpus :	10 speakers suffering from essential tremor.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Synchronous analysis based on the autocorrelation function ◦ <i>Analysis</i> : Estimation of the modulation frequency and depth of F_0 and the speech signal envelope on the basis of the positions of their peaks in F_0 trace.
Cues :	◦ Modulation frequency and modulation depth of F_0 and speech signal envelope.
Results :	◦ The modulation frequency increases with increasing vocal frequency.

(Schoentgen, 2002) [Sch02]

Corpus :	Vowels [a], [i] and [u] sustained by 51 moderately dysphonic patients (without pathological tremor) and 38 control speakers.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Cycle length sequence obtained via temporal peak-picking. ◦ <i>Analysis</i> : Based on the spectrum of the cycle length sequence or the auto-covariance function in the frequency interval $[0.5Hz, 25Hz]$.
Cues :	<ul style="list-style-type: none"> ◦ Modulation frequency (2 cues) related to the weighted average of significant peaks in each spectrum. ◦ Modulation depths (2 cues) related to the maximal or standard deviation of the cycle length sequence divided by the average cycle length.
Results :	◦ No statistically significant difference is observed between male and female speakers, between control and moderately dysphonic speakers, or between different timbres.

(Buder and Strand, 2003) [BS03]

Corpus :	Vowel [a] sustained by one control speaker, one speaker with amyotrophic lateral sclerosis and one speaker with adductor spasmodic dysphonia.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : F_0 contour extraction via a waveform-matching approach based on cross-correlation ◦ <i>Analysis</i> : Estimation of the modulations of F_0 (and of the sound pressure level) via a time-frequency Fourier analysis (modulogram). Three frequency bands have been considered : the <i>flutter</i> ($[10Hz, 20Hz]$), the <i>tremor</i> ($[2Hz, 10Hz]$) and the <i>wow</i> ($[0.2Hz, 2Hz]$). For each modulation, a low-frequency spectrogram is obtained.
Cues :	<ul style="list-style-type: none"> ◦ Prominent modulation frequencies and their magnitudes. ◦ Percentage of time over which the selected modulation is observed. ◦ The sinusoidal regularity of the modulations.
Results :	<ul style="list-style-type: none"> ◦ Observation of multiple irregular modulation frequencies in the <i>tremor</i> domain. ◦ Periodic and sustained modulations are observed in the <i>wow</i> domain.

(Kreiman et al., 2003) [KGG03]

Corpus :	Synthetic vowels [a] obtained on the basis of the phonation samples from 32 speakers with various dysphonia.
Method :	<ul style="list-style-type: none"> ◦ <i>Tracking</i> : Cycle length sequence obtained via temporal peak-picking or zero-crossing detection. ◦ <i>Synthesis</i> : Parameter estimates for the synthesis were derived from acoustical analyses of the original speech sounds. Vocal tremor perturbation time series has been generated via a sine wave tremor model or an irregular tremor model. Synthetic stimuli have been obtained via a formant-based synthesizer. ◦ <i>Perceptual analysis</i> : Three experiments have been carried out to examine perceptually the vocal tremor frequency pattern (5 expert judges), the effects of the vocal tremor modulation depth (10 experts judges) as well as the effects of changes in the rate of the vocal tremor (10 expert judges). For each experiment, the judges were asked to rate the similarity between the synthetic and original speech sounds.

Continued on next page

Continued from previous page

Results :	<ul style="list-style-type: none"> ○ No significant effect on perception of the tremor pattern is observed, but depends in part on the perceived severity of vocal tremor. ○ The coefficient of variation of the F_0 contour is a good predictor of the perceived severity of vocal tremor. ○ Differences in tremor rate are easily perceived for small and sinusoidal vocal tremor but not for large or complex tremor patterns.
------------------	---

(Cnockaert et al., 2008) [Cno+08]

Corpus :	Vowels [a] sustained by 37 Parkinson speakers and 35 control speakers.
Method :	<ul style="list-style-type: none"> ○ <i>Tracking</i> : F_0 trace is estimated on the basis of the instantaneous frequency obtained via a continuous wavelet transform (CWT). ○ <i>Analysis</i> : The modulation cues are computed via a second continuous wavelet transform applied to the F_0 trace. Estimates of the modulation amplitude, frequency and energy ratio, reported hereafter, are obtained.
Cues :	<ul style="list-style-type: none"> ○ Modulation amplitude obtained by summing the square of the modulus of the wavelet transform over the frequency interval [3Hz, 15Hz] and normalized by the average F_0. ○ Modulation frequency defined as the centroid of the modulation spectrum in the frequency interval [3Hz, 15Hz]. ○ Modulation energy ratio defined as the ratio of the energy in the frequency interval [3Hz, 7Hz] and the energy in the frequency interval [7Hz, 15Hz]
Results :	<ul style="list-style-type: none"> ○ F_0 is statistically significantly higher for male Parkinson speakers and the modulation amplitude is statistically significantly higher for female Parkinson speakers. ○ The modulation frequency is statistically significantly higher for Parkinson speakers and the modulation energy ratio is statistically significantly lower for Parkinson speakers compared to control speakers. ○ A statistically significant difference between Parkinsonian and control speakers is observed for male speakers, with regard to modulation frequency and phonatory frequency. For female speakers, no statistically significant difference is observed.

(Shao et al., 2010) [Sha+10]

Corpus :	Vowels [i] sustained by 24 control speakers, 15 Parkinson speakers and 10 speakers with vocal polyps.
Method :	<ul style="list-style-type: none"> ○ The sustained speech sounds have been analyzed by means of the Multi-Dimensional Voice Program (MDVP) in which the cycle length tracking method is based on a short-term autocorrelation analysis. ○ A customized non-linear dynamic analysis has been performed to measure the complexity of the speech signals by means of the correlation dimension.
Cues :	<ul style="list-style-type: none"> ○ 6 perturbation cues collected from MDVP : percent jitter, percent shimmer, amplitude tremor index (ATRI), frequency tremor intensity index (FTRI), amplitude tremor frequency (Fatr), and fundamental frequency tremor frequency (Fftr). ○ Correlation dimension D_2.
Results :	<ul style="list-style-type: none"> ○ No significant difference is observed between patients and control speakers for amplitude tremor cues (ATRI and Fatr). ○ Statistically significantly higher cycle-to-cycle perturbation sizes, tremor frequency cues (FTRI, Fftr) and nonlinear dimensionality are observed for patients compared to control speakers. ○ No statistically significant difference is observed between Parkinson speakers and speakers with vocal polyps.

(Anand et al., 2012) [Ana+12]

Corpus :	288 synthetic vowels [a] obtained on the basis of the phonation samples of 4 speakers with essential tremor.
Method :	<ul style="list-style-type: none"> ○ <i>Synthesis</i> : Manipulation of the F_0 contour by means of the STRAIGHT speech vocoder for various average F_0, modulation frequencies, modulation depths and signal-to-noise ratios (SNR). ○ <i>Perceptual analysis</i> : Six judges (three experts, 3 naives) rated the severity of the vocal tremor for each stimulus on a seven-point rating scale. Each stimulus was presented several times in random order and a single average score is computed.

Continued on next page

Continued from previous page

- | | |
|------------------|---|
| Results : | <ul style="list-style-type: none">Voices with low F_0 are perceived to have greater tremor severity.Perceived severity of tremor increases with the modulation frequency and depth.No systematic effect of SNR on perceived tremor severity is observed. |
|------------------|---|

Table 1.3 – Chronological literature review

1.5 Objectives and motivations

The general framework of the thesis is the assessment of disordered voices. The assessment of voice and laryngeal function is based on auditory ratings and acoustic analyses of speech sounds. Acoustic feature-based assessment methods are indeed popular because they are non-invasive and enable clinicians to monitor the voice of patients quantitatively.

The goal of this study is the analysis of vocal tremor and vocal jitter in Parkinson speakers and control speakers. Few studies have, indeed, been devoted to vocal tremor in human speakers in general and Parkinson speakers in particular. Also, a large majority of the existing studies have involved small corpora with tens of speakers at most. Idem, no studies have addressed jointly vocal jitter and vocal tremor (neurological & physiological) as well as declination.

Fast, small and involuntary cycle-to-cycle perturbations of vocal cycle lengths are designated as vocal jitter and involuntary low-frequency modulations of the vocal cycle lengths are referred to as vocal tremor. The latter have physiological (breathing, cardiac beat and pulsatile blood flow) or neurological causes. Conventionally, vocal jitter and tremor are tracked in sustained speech sounds in which small cycle length perturbations are less likely to be masked by intonation, accentuation or segment-specific phenomena. Disorders of phonation are often a consequence of the inability of vocal folds to vibrate regularly. Larger than normal disturbances of the periodicity of the glottal source signal are therefore observed frequently as a consequence of organic or functional disorders of the larynx.

Here, estimates of the vocal cycle length perturbation (vocal jitter, neurological tremor, physiological tremor) size, frequency and bandwidth in vowels sustained by patients suffering from neurological disorders and normal control speakers are reported. The analysis relies on the tracking of the vocal cycle lengths in sustained voiced speech sounds by means of a temporal method, called *salience-based cycle length tracking* (SCLT), that is not based on strong assumptions with regard to the regularity of the speech cycles and their lengths. Speech cycles are tracked via a multi-scale analysis that assigns a salience to each signal peak and is founded on dynamic programming.

The cycle length time series is then decomposed into a sum of oscillating components by empirical mode decomposition the instantaneous envelopes and frequencies of which are obtained via an AM-FM decomposition. According to their frequency content, these modes are then assigned to four categories (vocal jitter, neurological tremor, physiological tremor and a residual trend, which is due to intonation and declination) and added within each category. The length time series components that are so obtained are then further analyzed to measure perturbation size, perturbation frequency and perturbation bandwidth. The within-category size of the cycle length perturbations is estimated via the standard deviation of the empirical mode sum divided by the average cycle length. Additional neurological tremor cues are obtained on the basis of the instantaneous frequencies and amplitudes of the empirical modes.

Cycle length jitter and vocal tremor frequencies and depths are obtained for two corpora of vowels comprising 123 and 74 control and 456 and 205 Parkinson speaker recordings respectively.

The general scheme of the proposal is shown in Figure 1.8.

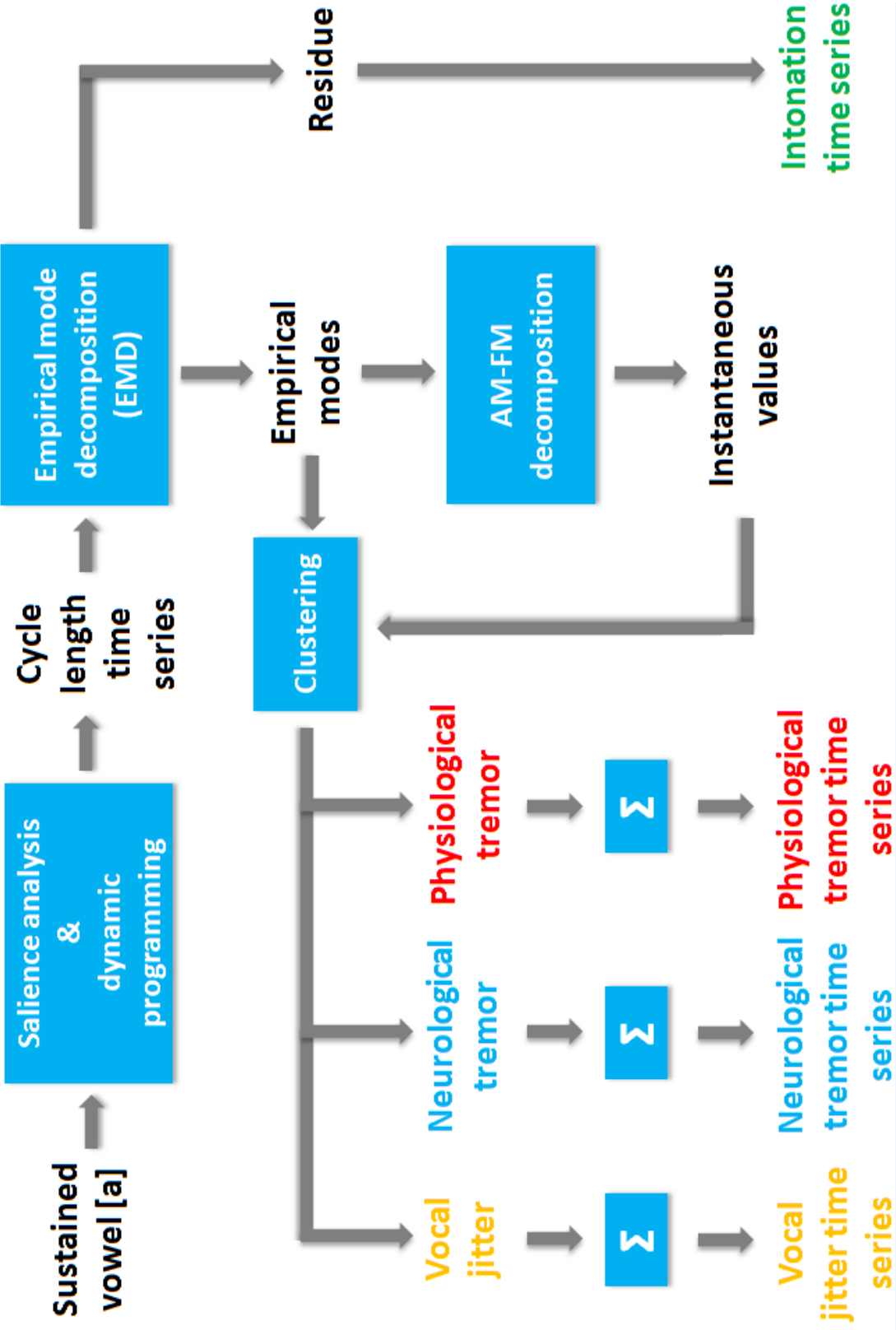


Figure 1.8 – Analysis scheme

Key points

- Voice production relies on mechanisms that involve several parts of the human body : the sub-glottal air pressure system, the glottal vibratory system and the supra-glottal resonating system
- The positioning or motion of these body parts involves the contraction of muscles, controlled by the central and peripheral nervous systems
- Disorders of phonation are often a consequence of the inability of vocal folds to vibrate regularly (organic or functional disorders of the larynx, impairment of the central nervous system)
- Vocal jitter designates small, fast and involuntary cycle-to-cycle perturbations of the vocal cycle lengths
- Vocal tremor designates small, slow, oscillatory and involuntary perturbations of the vocal cycle lengths
- Acoustic feature-based assessment is popular because it is non-invasive and enable clinicians to monitor the voice of patients numerically
- The aim of the study is to report the size, frequency and bandwidth of vocal cycle length perturbations at four distinct time scales



2. Cycle length tracking methods

Objectives of this chapter

- Describe conventional techniques used to track cycle lengths or instantaneous vocal frequency in voiced speech sounds.
- Illustrate some tracking methods by means of simple examples

Contents

2.1	Introduction	25
2.2	Short-term analysis	26
2.2.1	Temporal domain	26
2.2.2	Frequency domain	28
2.2.3	Model-based methods	31
2.3	Cycle-synchronous event analysis	33
2.3.1	Narrow-band temporal pattern analysis	33
2.3.2	Broad-band temporal pattern analysis	33

2.1 Introduction

In voice production, vocal frequency F_0 is related to the vibration rate of the vocal folds. Fundamental period T_0 is defined as the reciprocal of F_0 in a pseudo-periodic waveform. F_0 -tracking refers to the task of estimating the contours of the fundamental frequency for voiced segments or detecting individual laryngeal cycles lengths.

In the literature, several fundamental frequency/period tracking methods are proposed. These methods may be classified into two categories (Figure 2.1) : the F_0 *contour extraction via short-term analysis* and the *cycle-to-cycle length extraction via cycle-synchronous event analysis*.

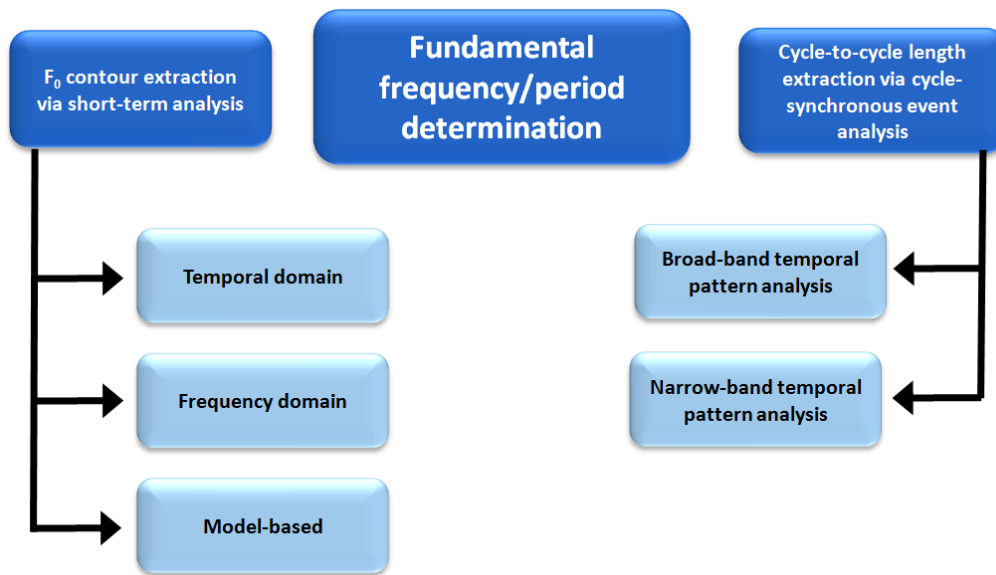


Figure 2.1 – Cycle length tracking methods

In the *short-term analysis* category are grouped the methods for which the speech signal is divided into consecutive frames which are then processed separately. The tracking of F_0 is applied frame-by-frame via a short-term transformation and aims at obtaining an estimate of F_0 which is typical of that frame. Additional processes enable the analysis of time-evolving F_0 by building a so-called F_0 contour.

In the second category, called *cycle-synchronous event analysis*, the vocal cycle lengths and the cycle length time series are obtained cycle-by-cycle by tracking temporal patterns that characterize the same glottal event. These techniques enable a more accurate analysis of time-evolving F_0 values but are directly affected by additive noise owing to the voice production or recording conditions.

2.2 F_0 contour extraction via short-term analysis

The F_0 short-term analyses are frame-based, giving one or several F_0 estimates per frame. In addition, the candidates can be connected by post-processing techniques (e.g. dynamic programming, neural networks, or hidden Markov models) to build a time-varying F_0 time series, called *pitch contour*, or to use as a priori starting F_0 value for cycle-to-cycle length extraction methods.

The frame length is usually chosen short enough to satisfy the assumption of pseudo-stationarity within the frame. As a consequence, the F_0 estimate is expected to be representative of the frame. However, that length has to be chosen large enough to guarantee that the features are measurable. A typical frame length N is equal to 2 or 3 cycles.

The frame-based analysis methods may be classified into 3 categories according to their domains of description : temporal, spectral or model-based.

2.2.1 Temporal domain

2.2.1.1 Autocorrelation function (ACF)

Let us consider a realization $x(n)$ of a stationary stochastic process. Its autocorrelation function $R_x(\alpha)$ measures the similarities between the signal $x(n)$ and its time shifted version $x(n + \alpha)$, where α is the lag. In statistics, the autocorrelation function is computed via the mathematical expectation as follows :

$$R_x(\alpha) = E[x(n)x(n + \alpha)] \quad (2.1)$$

Since the random process is a function of time, the autocorrelation function $R'_x(\alpha)$ may also be evaluated via the temporal average over some period of time, L , or over a series of events :

$$R'_x(\alpha) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{n=-\frac{L-1}{2}}^{\frac{L-1}{2}} x(n)x(n + \alpha) \quad (2.2)$$

One may show that these two formulations yield the same result if the process is stationary and ergodic. The short-term analyses based on autocorrelation use the second relation. Moreover, for non-stationary signals, the autocorrelation function r_x is computed frame-by-frame considering that the signal is stationary within the frame. For a frame of length N and origin q , the autocorrelation function is given by :

$$r_x(\alpha, q) = \frac{1}{N - |\alpha|} \sum_{n=q}^{q+N-\alpha} x(n)x(n + \alpha) \quad (2.3)$$

By assessing similarities between a signal and its time shifted version, the autocorrelation function is frequently used to detect repeating patterns, such as periodicity affected or not by noise. For instance, the autocorrelation function of a periodic signal exhibits a strong peak when the lag α equals the cycle period T_0 (in samples). This approach is frequently used by pitch tracking methods like [Boe93] [DK02]. For instance, Figure 2.2 illustrates the autocorrelation function obtained for a fragment of sustained voiced speech sound. One observes that the maxima of ACF are located at a lag equal to T_0 and its integer multiples.

Nevertheless, the accuracy of this global peak position tracking is not always guaranteed in presence of signals with rapid F_0 changes (spread global peak shape) or with a strong formant at the second or third harmonic (detection of harmonics rather than the fundamental frequency).

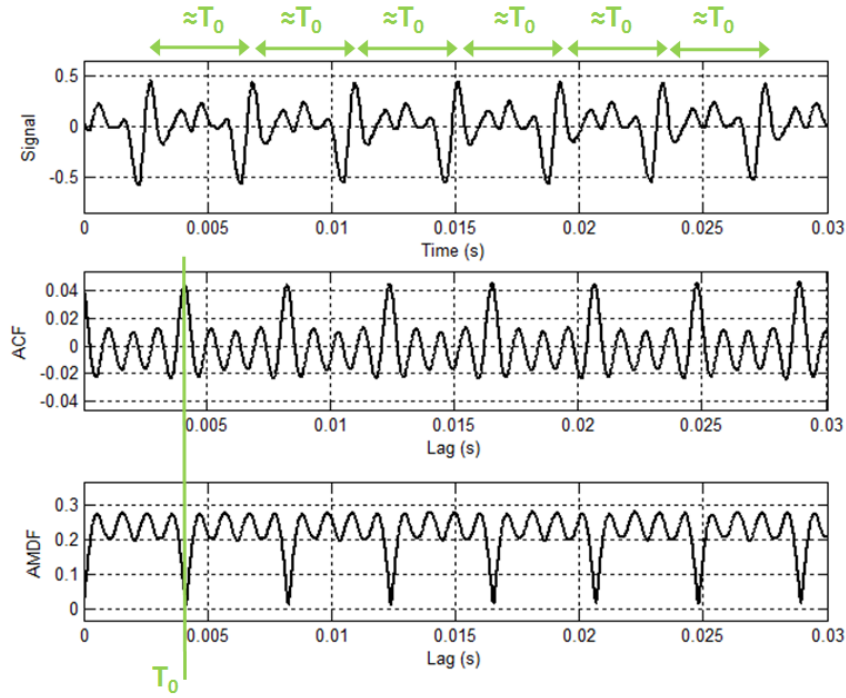


Figure 2.2 – Sustained vowel, autocorrelation function (ACF) and AMDF

2.2.1.2 Average magnitude difference function (AMDF)

Here, the average cycle length is determined frame-by-frame by comparing the values of a signal and its time shifted version. For a lag α , the distance function $D(\alpha, q)$, applied to a frame of length N and origin q is given by :

$$D(\alpha, q) = \frac{1}{N - |\alpha|} \sum_{n=q}^{q+N-|\alpha|} |x(n) - x(n + \alpha)| \quad (2.4)$$

As a counterpart of the autocorrelation function, this distance function is expected to have a minimum when the lag α equals the cycle period T_0 (in samples). Some tracking methods based on AMDF are explained in [Ros+74] [Hes83]. Figure 2.2 illustrates the AMDF obtained for a fragment of a sustained voiced speech sound.

2.2.2 Frequency domain

The spectrum of a sustained voiced speech sound is characterized by a fundamental frequency F_0 and its harmonics. The spectral contour is related to the vocal tract that filters the excitation signal causing formants and sometimes anti-formants to appear.

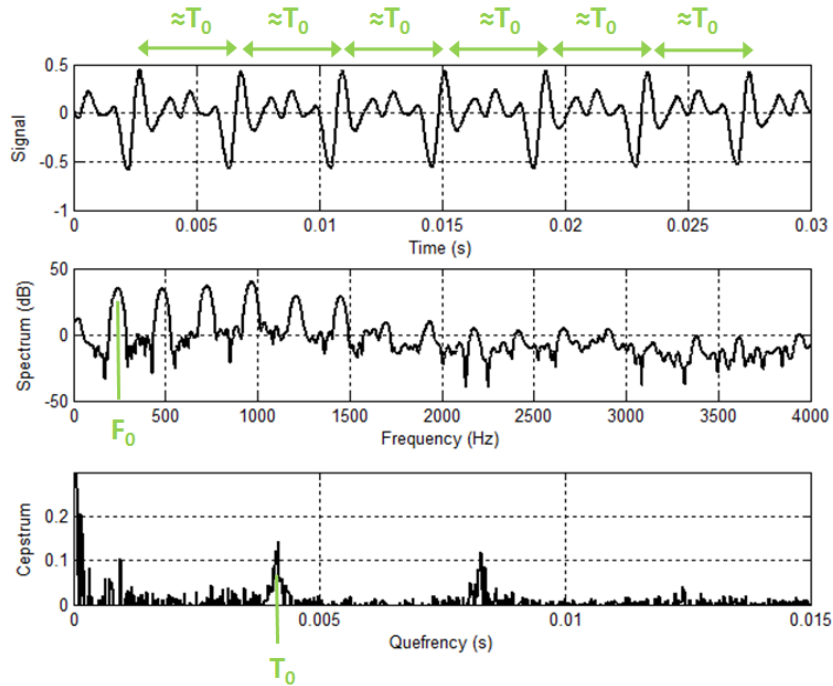


Figure 2.3 – Cepstrum of a voiced sound

Figure 2.3 illustrates a fragment of voiced speech and its spectrum. The direct determination of F_0 as the location of the first peak in the spectrum is often unreliable and inaccurate. For this reason, conventional tracking methods in the frequency domain are based on the investigation of the harmonic structure of the signal so that many harmonics contribute to the estimate of F_0 .

2.2.2.1 Cepstrum

One technique investigating harmonic structure consists in computing the cepstrum [OS04] of the speech signal. The cepstrum is defined as the inverse Fourier transform of the logarithm of the speech signal power spectrum. Assuming that the spectrum consists in harmonics spaced F_0 apart, the cepstrum highlights a prominent cepstral component at the quefrency $T_0 = 1/F_0$. Moreover, this transformation enables the separation of the effects related to the vocal source (harmonics \rightarrow high quefrencies) and the vocal tract (slow frequency-varying spectral contour \rightarrow low quefrencies). Figure 2.3 illustrates the cepstrum of a fragment of voiced speech.

2.2.2.2 Harmonic product spectrum

Another technique, illustrated in Figure 2.4, consists in investigating the harmonic structure by means of the *harmonic product spectrum (HPS)*. HPS is obtained by multiplying a set of signal amplitude spectra the frequency axes of which are compressed by an integer factor k . Assuming that the fundamental and its harmonics are equally spaced on the frequency axis, HPS displays a prominent spectral peak at F_0 .

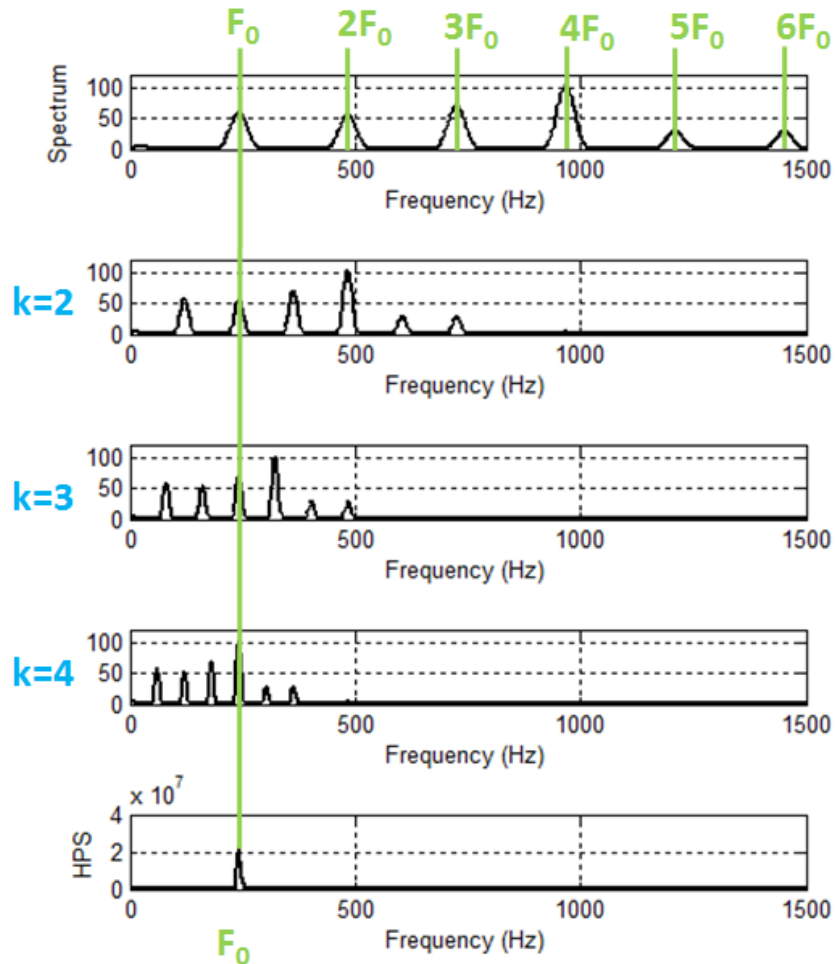


Figure 2.4 – Compressed spectra and harmonic product spectrum

2.2.2.3 Wavelet transform

Another F_0 short-term analysis technique, based on a continuous wavelet transform (CWT), is proposed in [Cno07]. The F_0 estimate is obtained by means of the phase derivative of the wavelet coefficients obtained via a continuous complex wavelet transform (in the frequency band corresponding to the maximal CWT modulus). The phonatory frequency trace is then analyzed by another wavelet transform to enable tracking local vocal frequency perturbations. Figure 2.5 illustrates the two continuous wavelet transforms applied to a synthetic speech sound.

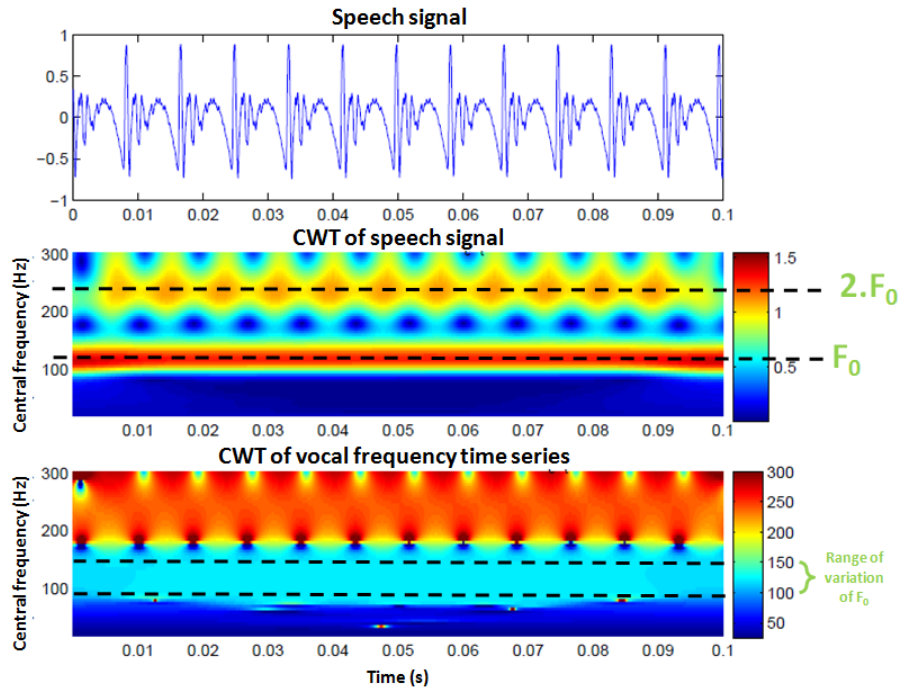


Figure 2.5 – Continuous wavelet transforms

2.2.3 Model-based methods

Model-based tracking methods involve a discrete-time system the purpose of which is estimating either the transfer function of the vocal tract or the harmonic structure directly. Often, linear prediction is used. These models rely on an autoregressive (AR) model to describe the speech signal $x(n)$. They attempt to predict the present signal value $x(n)$ via a linear combination of previous values $x(n-k)$:

$$x(n) = \sum_{k=1}^p a_k x(n-k) + e(n) \quad (2.5)$$

where $\{a_k\}$ are the p^{th} order linear predictor coefficients and $e(n)$ is the residual prediction error. The transfer function of this all-pole autoregressive model is :

$$\frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)} \quad (2.6)$$

For instance, Figure 2.6 illustrates the application of linear prediction for estimating the spectral contour of the speech signal. For that, the speech signal has been decimated to a sampling frequency equal to $8kHz$ and a 10th-order AR filter has been determined by minimizing the prediction error in the least squares sense.

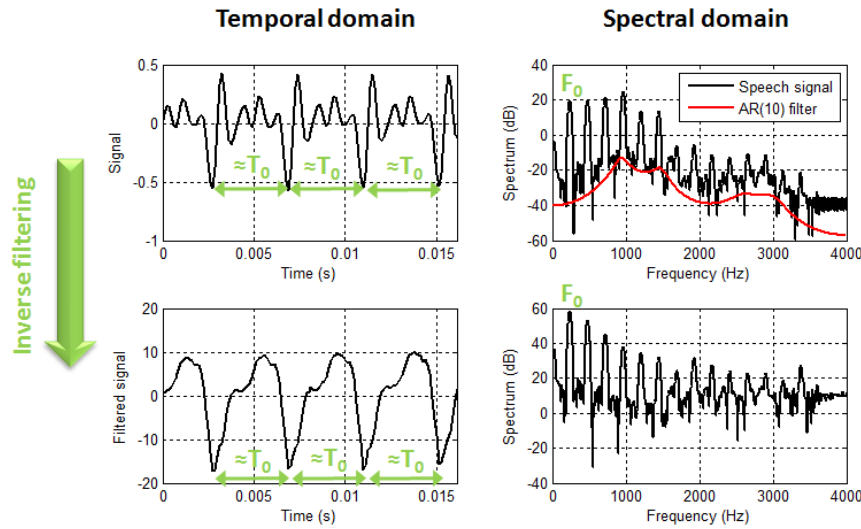


Figure 2.6 – Linear prediction and inverse filtering

One observes that the filter frequency response (red curve) highlights several vocal tract resonance frequencies. By inverse filtering, the influence of the vocal tract can be removed, which yields the excitation time series. This latter enables more accurate tracking of T_0 via previously illustrated techniques. Moreover, as shown in Figure 2.7, the residual displays peaks in the vicinity of glottal closure. These events, that are related to the inability of (linear) system (2.5) to model local non-linearities, may also be used for tracking.

Another example is given in Figure 2.8. Here, the speech signal has been decimated to a sampling frequency $F_s = 2kHz$. Therefore, the spectrum of this decimated signal reports information

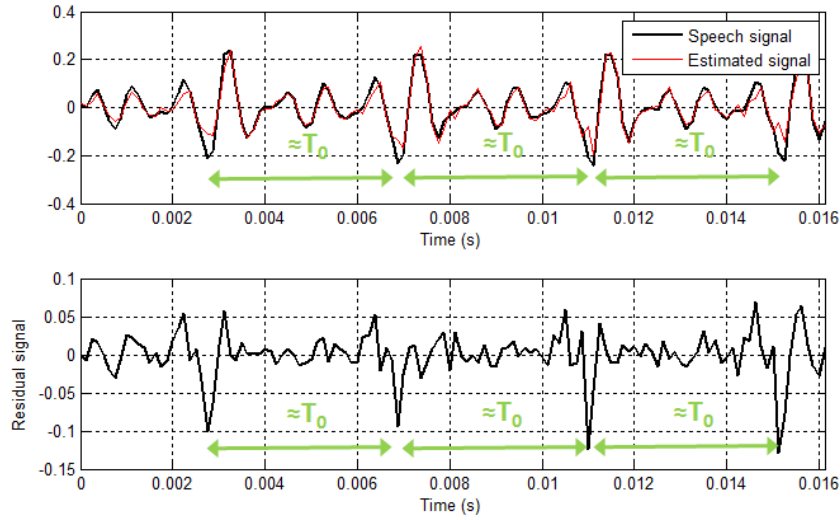


Figure 2.7 – Residual signal after linear prediction

in the frequency range $[0\text{Hz}, 1000\text{Hz}]$ so that high frequency formants are discarded. The harmonic structure is then assessed by means of a 41st-order linear prediction filter. Assuming that two complex conjugate poles are required to track prominent spectral peaks, the frequency response of this filter highlights spectral peaks at the fundamental frequency F_0 and its harmonics. The fundamental period T_0 may also be determined on the base of the impulse response of the filter.

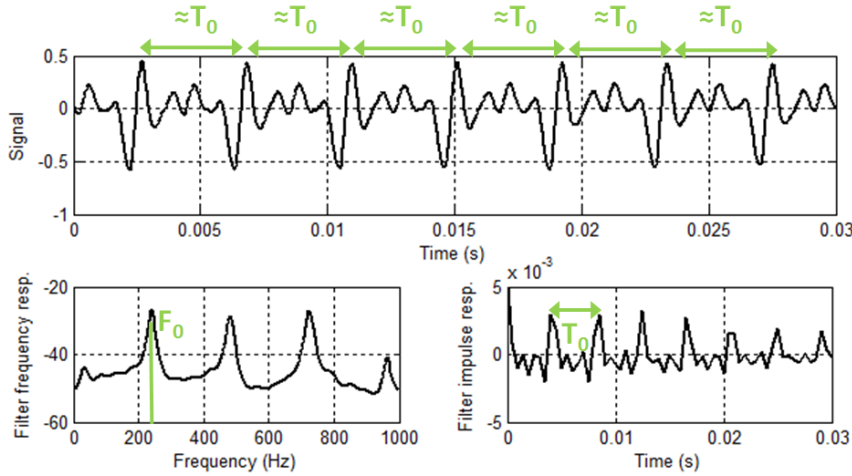


Figure 2.8 – Linear prediction of the decimated signal

2.3 Cycle length extraction via cycle-synchronous event analysis

The goal of these methods consists in tracking the speech cycles one by one. A sequence of cycle boundaries, called *markers*, are obtained. Since the instantaneous fundamental frequency is obtained for each cycle individually, these techniques are sensitive to local perturbations and voice breaks, and are thus less reliable than the previously considered short-term analyses. However, the advantage of the cycle-to-cycle tracking is its ability to estimate cycle lengths even when the signal is not perfectly periodic (but still cyclic).

2.3.1 Narrow-band temporal pattern analysis

The narrow-band temporal pattern analysis is based on the detection of signal events reporting glottal cyclicity after low-pass filtering, so that the harmonic structure as well as the vocal tract formants are removed. Simple tracking rules may therefore be applied. Figure 2.9 illustrates event detection on the basis of two different thresholds. A marker is set if the threshold is crossed in an arbitrarily chosen direction. This requires that the time series has only one threshold crossing per cycle, which is a severe drawback of the approach.

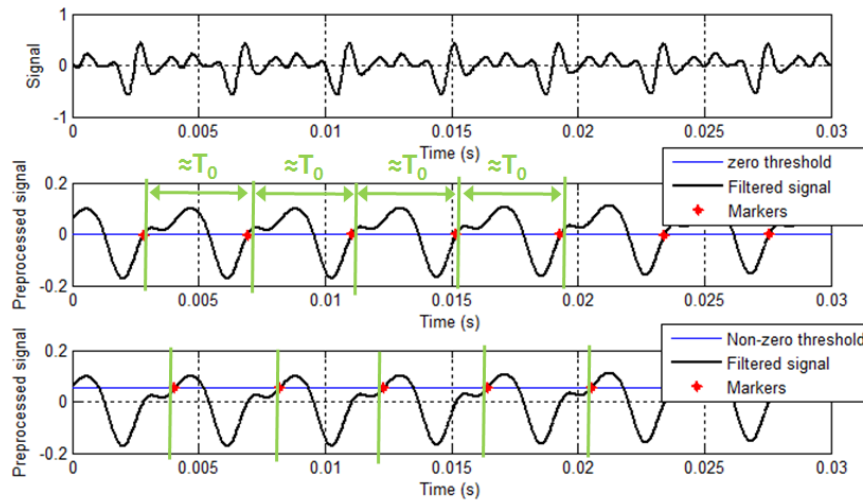


Figure 2.9 – Cycle tracking by threshold crossing

2.3.2 Broad-band temporal pattern analysis

The goal of broad-band temporal pattern analysis consists in detecting in each cycle the same glottal event. A typical glottal event is the glottal closure instant that corresponds to a discontinuity in the derivative of glottal airflow rate caused by the abrupt closure of the glottis. As a consequence, cycle detection rests on the recursive discovery and storage of speech signal extrema that occur in the vicinity of the instants of the maximal glottal excitation.

To enable this selection, one often assumes that voiced speech segments are pseudo-periodic so that the maxima can be selected one by one on the base of a prior estimation of the typical fundamental period T_0 . Conventionally, this typical cycle length estimate involves one of the previously described short-term analyses. Most tracking methods are based on the residual obtained after applying autoregressive linear prediction modelling that displays peaks in the vicinity of the glottal closure [SD83] [MY08] [SY95] [Nay+07]. Others exploit the pseudo-periodicity property of the speech signal in the adjacent cycles (autocorrelation, cross-correlation, ...) [Boe93] [PS14] [Tal95] [Hes83]. Post-processing optimization paradigms, like dynamic pro-

gramming or neural networks, are then applied to select relevant speech cycle peaks.

However, the required assumption of quasi-equal peak spacing is valid for modal voices only and not for pathological ones, which may be characterized by large cycle-to-cycle fluctuations in length or amplitude. Cycle insertion or omission errors may therefore occur, which bias the acoustic cues of cycle regularity.

In this study, a broad-band temporal pattern analysis, called *salience-based cycle length tracking* (SCLT), is proposed to track the cycle length in voiced speech sounds. The speech cycles are detected via the saliences of the speech signal samples, defined by the length of the temporal interval over which a sample is a maximum. The tracking of the cycle lengths is based on a dynamic programming algorithm which does not request that the signal is locally periodic and the average period length known a priori.

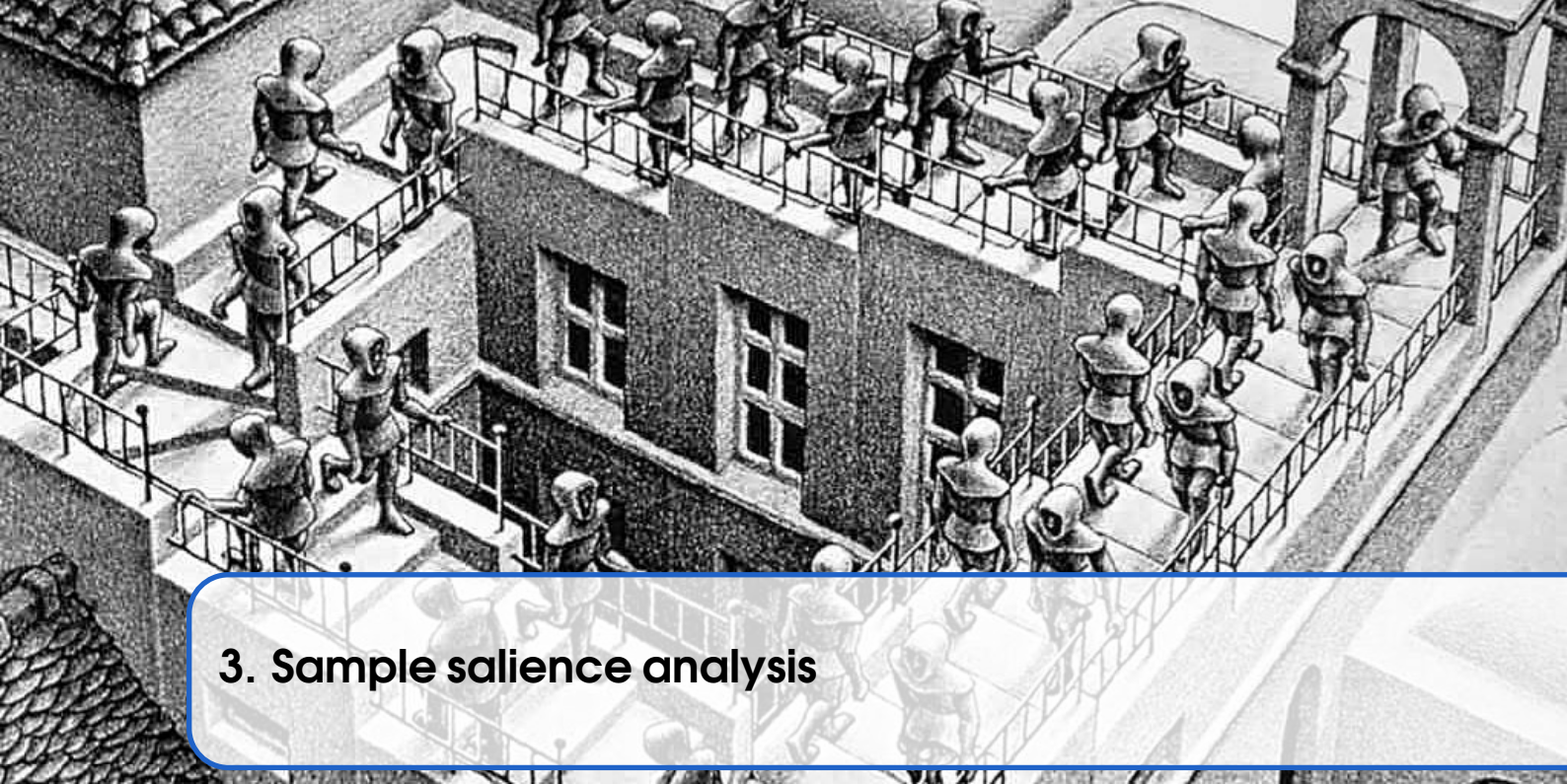
Key points

- Tracking methods may be classified in two categories : the F_0 *contour extraction via short-term analysis* and the *cycle-to-cycle length extraction via cycle-synchronous event analysis*
- *Short-term analysis* tracking methods are frame-based and deliver an estimate of the vocal frequency per frame. Their main advantage is their reliability in background noise
- *Cycle-synchronous event analysis* methods track speech cycles individually. These methods enable an accurate analysis of the time-evolving glottal cycle lengths but are easily influenced by local signal shape anomalies and voice breaks
- In this study, the proposed vocal cycle length tracking method is an example of a *cycle-synchronous event analysis* method



Methods

3	Sample salience analysis	39
3.1	Introduction	
3.2	Definition	
3.3	Salience allocation methods	
3.4	Reduction of computational load	
3.5	Validation	
3.6	Example	
3.7	Conclusions	
4	Tracking of cycle lengths via dynamic programming	61
4.1	Introduction	
4.2	Overview	
4.3	Problem formulation	
4.4	Application to an example array	
4.5	Conclusions	
5	Application of the SCLT method to sustained speech sounds	77
5.1	Introduction	
5.2	Preprocessing	
5.3	Speech sample salience analysis	
5.4	Cycle length tracking	
5.5	The vocal cycle length time series	
5.6	Conclusions	
6	The wonderful story of the rolling wheel	91
6.1	Introduction	
6.2	Fourier analysis	
6.3	Time-frequency analysis	
6.4	Instantaneous frequency and amplitude	
7	Time-frequency analysis via Empirical mode decomposition	109
7.1	Introduction	
7.2	Empirical Mode Decomposition	
7.3	Extraction of the instantaneous frequencies and envelopes	
7.4	Discussion	
8	Analysis of vocal cycle length perturbations	125
8.1	Introduction	
8.2	Time-frequency analysis of the vocal cycle length time series	
8.3	Categorization	
8.4	Average vocal frequency and cycle length perturbation size	
8.5	Neurological tremor frequency	
8.6	On the choice of the weights	
8.7	Conclusions	



3. Sample salience analysis

Objectives of this chapter

- Introduce the concept of sample salience to characterize signal samples with regard to their neighbourhood
- Propose and discuss several salience allocation methods

Contents

3.1	Introduction	41
3.2	Definition	42
3.3	Salience allocation methods	43
3.3.1	Basic algorithm	43
3.3.2	Sliding analysis window	45
3.4	Reduction of computational load	50
3.5	Validation	52
3.5.1	Preliminary remarks	52
3.5.2	Theoretical developments	52
3.6	Example	55
3.6.1	Application to an arbitrary array	55
3.6.2	Application to a voiced speech sound	58
3.7	Conclusions	58

3.1 Introduction

Often, the goal of signal processing is to discover and quantify prominent features of sequential data. Often, these features are easily perceived by the human brain and its sensory receptors like the eyes and ears. However, high-level computational techniques have to be implemented to quantify these same features.

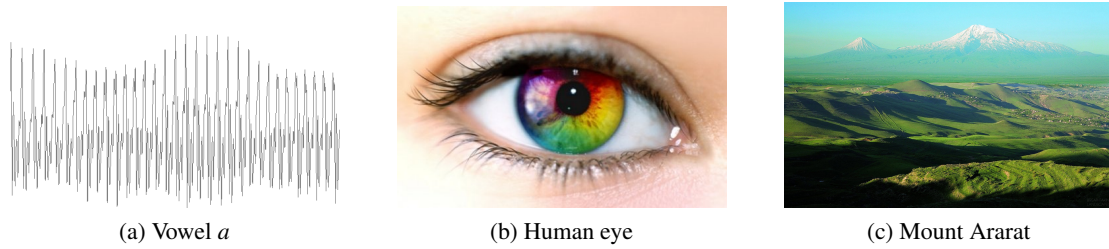


Figure 3.1 – Signal observation

For instance, Figure 3.1a reports the recorded signal of a sustained vowel $[a]$. One easily observes a succession of cycles even when several peaks occur in one cycle and the cycle amplitudes and/or durations differ widely. Another example is illustrated in Figure 3.1c. It is a photo of the landscape around Mount Ararat. One observes easily the two highest peaks (*Greater Ararat* and *Lesser Ararat*) but the other hills in front of these peaks may also be clearly discerned even though their heights are smaller.

In topography, mountains are characterized by their *elevation*, *prominence* and *isolation* (Figure 3.2a). The first one is a measurement of a summit's height relative to the (fixed) mean sea level. The others refer respectively to the height and dominance of a summit's peak relative to its surroundings :

- The *topographic prominence* measures a summit's vertical distance from the lowest contour line that encircles it and no higher peak.
- The *topographic isolation* is the great circle radius to the nearest point of equal elevation.

These features tend to summarize what the human eyes and brain perceive. Peaks with high prominences and isolations often have impressive summit views even when their elevations are comparatively modest.

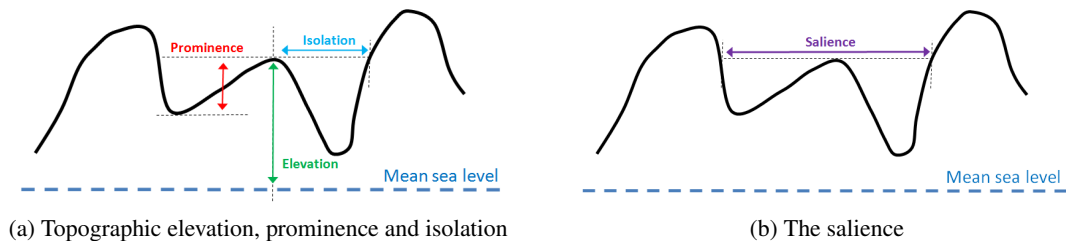


Figure 3.2 – Topographic characterization of a mountain summit

Here, one introduces the concept of *signal salience*. Salience may be defined from a topographic point of view as a generalization of a summit's isolation. It reports the length of a line segment

(parallel to the mean sea level) over which the summit has the highest elevation (Figure 3.2b). That salience is also a measure of dominance relative to the surroundings. For instance, Figure 3.3a illustrates a tent in the Marocco desert. The highest part of its roof is expected to have a high salience because that desert is a region with no or a few sparse buildings, dunes or vegetation. However, the same tent in the streets of New-York (Figure 3.3b) is expected to have a smaller salience.



(a) Desert of Marocco



(b) New York

Figure 3.3 – Topographic salience

In this study, the salience is defined for unidimensional time series. In that case, the salience corresponds to the duration of the longest temporal interval over which a signal sample is a maximum. Its unit will be expressed in seconds (or number of discrete samples).

The saliences are used to detect automatically voiced speech cycles. Salience is a relevant signal feature because one observes that signal peaks that are similarly positioned in vocal cycles may have similar saliences even if the peak amplitudes differ widely. This also applies to peaks in cycles the durations of which are perturbed moderately. The salience can therefore be used to detect automatically voiced speech cycles because they display a prominent peak in the vicinity of glottal closure.

3.2 Definition

One considers a signal array of length M . Each sample is characterized by its position and value. Salience $s(k)$ of sample k is defined as the length of the temporal interval over which that sample is a maximum. Therefore, the salience of the global maximum is M , if there is no other sample with the same value.

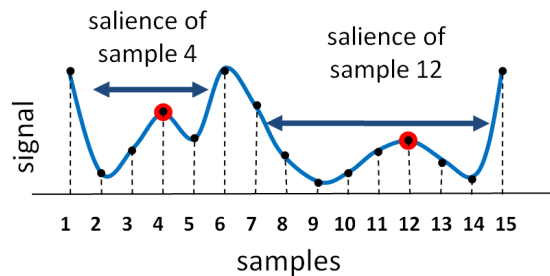


Figure 3.4 – Salience definition

For instance, Figure 3.4 illustrates the saliences assigned to the 4th and 12th samples. One observes that the 4th sample, in spite of the fact that it has a larger value than the 12th, has a smaller saliency. A sample with a large value has not necessarily a high saliency, and vice-versa. The saliency is determined by the value of a sample relative to its neighbours.

One also defines a right saliency $s_r(k)$ and left saliency $s_l(k)$ as the number of samples over which a sample k is a maximum to the right and left. Total saliency, left and right saliences are related as follows :

$$s(k) = s_l(k) + s_r(k) + 1 \quad (3.1)$$

For example, the saliency values assigned to the 4th and 12th samples are reported hereafter :

k	$s(k)$	$s_l(k)$	$s_r(k)$
4	4	2	1
12	7	4	2

3.3 Saliency allocation methods

In this section, several methods are proposed to assign a saliency to each sample of an array of length M . The raw approach consists in considering one-by-one each array sample, computing to its right and left the lengths of the longest temporal interval over which that sample is a maximum and summing these. The computational load is important due to several redundant steps.

The basic proposal consists in considering all possible within-array analysis intervals and recording the length of the largest interval over which a sample is a maximum. As shown later, one of the weaknesses of that approach is the effect of the array boundaries, which bias the saliency values. As an alternative, a window-based approach is proposed. The window is smaller than the array length, and moves sample by sample to the right. It has the advantage that each sample occupies different positions within the window and its saliency value is therefore independent of the array origin. This method also uses some computational artifices to avoid redundancy.

3.3.1 Basic algorithm

The basic approach for estimating saliences consists in considering all possible within-array analysis intervals and recording the length of the largest interval over which a sample is a maximum (Figure 3.5a).

First, all sample saliences are put to 1 (because each sample is a local maximum with regard to itself) and the length of the analysis interval is put to $n = 2$.

The signal array of length M is then subdivided into analysis intervals of length n . The rightmost interval stops at the right array boundary whatever its length (i.e. the rightmost interval length is comprised between 1 and n). Within each interval, the maximum is determined and a saliency equal to n is assigned to the interval maximum. The length of the analysis interval n is then increased by one and a new array subdivision and saliency allocation is carried out. At the end, all samples have at least one and maximally M saliency values. Only the highest saliency value is kept for each sample.

As an example, Figure 3.5b illustrates the steps of the algorithm for an array of length $M = 16$. One observes that the position of the analysis array within the signal affects the results. For instance, in Figure 3.5b, the 2nd sample has a saliency equal to 1. However, if the second sample of the array

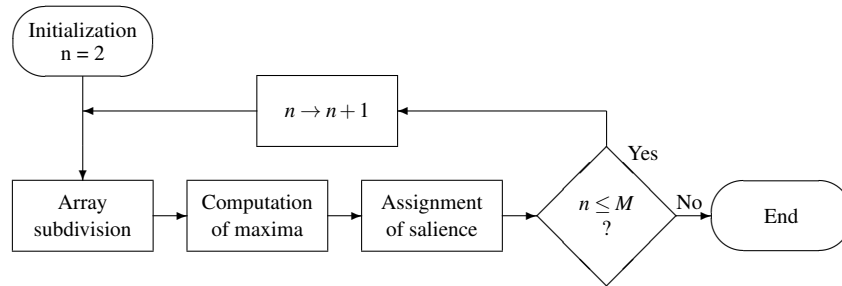
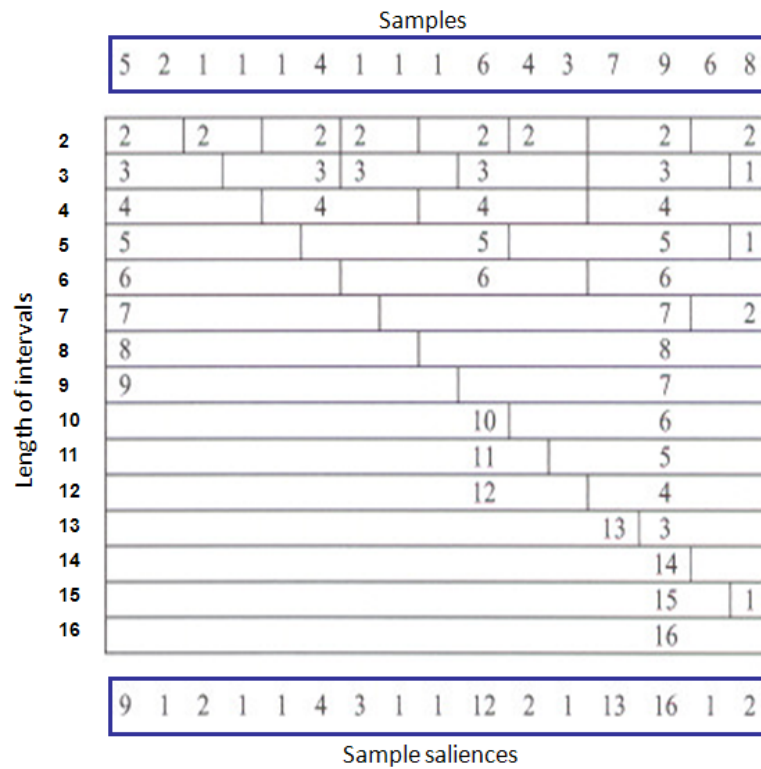
(a) Flow chart : salience allocation applied to an array of length M (b) Example : salience allocation applied to an array of length $M = 16$

Figure 3.5 – Basic algorithm for salience computation

was taken as the origin, it would have a saliency equal to 4.

Moreover, the saliency of the samples in the rightmost interval is affected by the anomalous interval length ($< n$). Also, no information is available about the samples outside the array. As a consequence, spurious saliency values may be assigned to samples near the array boundaries.

To obtain sample saliencies that are position-independent, one may rotate M times the M samples in the analysis array so that each sample occupies once the left and right boundary positions. The sample saliency is calculated for each within-array rotation. The final saliency is the average of the saliencies computed for each rotation. The average sample saliency may therefore be considered to be independent of the position with regard to the array boundaries. However, when rotating, samples that are distant in time may be put into contact, which may be a different cause of bias. Rotation also increases processing time.

Finally, the saliency is globally defined so that a sample with a higher value than its neighbours has a high saliency value $\leq M$. So, if the length of the analyzed signal is high, the saliency values differ widely. These differences may not be useful, and the knowledge of the saliency within a limited temporal interval may be preferred. Therefore, an alternative is proposed to obtain sample saliencies that are less dependent on position and valid within a limited temporal interval.

3.3.2 Sliding analysis window

To overcome the problems related to boundary effects and choice of origin, a sliding analysis window is used for saliency allocation. That analysis window $w_N(i)$ (of origin i , and length N smaller than the array length) is placed at the beginning of the array and moves sample-by-sample to the end. Therefore, each sample occupies different positions within the window and its saliency value is independent of the array origin.

Moreover, each array sample is ensured to get a valid saliency value within a limited temporal interval and the processing time may also be significantly decreased. The processing time improvement relies on several computational artifices that decrease step redundancy. As seen previously, the raw approach consists in considering each sample individually and determining the longest left and right interval lengths over which the sample is a maximum.

Hereafter, allocation relies on a partial saliency assigned at each position of the analysis window and an update of the partial saliencies. For the sake of clarity, one introduces hereafter three kinds of saliencies : the local (window-based values), running (updated values) and final saliency. The partial saliency allocation as well as the updating (called global saliency allocation) are described hereafter.

3.3.2.1 Local, running and final saliencies

Local saliencies :

The term “local” refers to the saliency values which have been obtained for one position of the sliding analysis window $w_N(i)$, independently of the previous window positions. Symbols $s(k, i)$, $s_l(k, i)$ and $s_r(k, i)$ designate respectively local total saliency, local left saliency and local right saliency values assigned to sample k when the window origin is sample i . For an analysis window of length N , the saliency ranges satisfy the following relations :

$$\begin{cases} 1 \leq s(k, i) \leq N \\ 0 \leq s_l(k, i) \leq N - 1 \\ 0 \leq s_r(k, i) \leq N - 1 \end{cases} \quad (3.2)$$

These values are defined within a temporal interval relative to the analysis window. They are interdependent. The left and right salience values may be $\leq N - 1$, but their sum is bounded and has to satisfy relation (3.1) :

$$s(k, i) = s_l(k, i) + s_r(k, i) + 1 \quad (3.3)$$

Running saliences :

The term “running” refers to the values of saliences that have been assigned to a sample taking into account the previous positions of the analysis window. Running values increase monotonically when the window slides to the right. Symbols $s^*(k, i)$, $s_l^*(k, i)$ and $s_r^*(k, i)$ designate respectively running salience, running left salience and running right salience values assigned to sample k with i being the present origin of the window. Each running salience is the maximum of the corresponding local salience obtained for the previous (and present) window positions :

$$\begin{cases} s^*(k, i) = \max_{i' \leq i} (s(k, i')) & 1 \leq s^*(k, i) \leq N \\ s_l^*(k, i) = \max_{i' \leq i} (s_l(k, i')) & 0 \leq s_l^*(k, i) \leq N - 1 \\ s_r^*(k, i) = \max_{i' \leq i} (s_r(k, i')) & 0 \leq s_r^*(k, i) \leq N - 1 \end{cases} \quad (3.4)$$

Their values are defined within the analysis window ($\leq N$) but the maximal running saliences (left, right or total) are generally obtained for different positions of the sliding analysis window. As a consequence, equation (3.1) is not satisfied in general : $s^*(k, i) \neq s_l^*(k, i) + s_r^*(k, i) + 1$.

Final saliences :

The term “final” refers to the salience values obtained for each sample at the end of the process. For an array of length M and an analysis window of length N , the origin i of the rightmost analysis window $w_N(i)$ is located at $i = M - N + 1$. Symbol s_f designates the final salience values that are linked to the running left and right saliences by equation (3.1) :

$$\begin{aligned} s_f(k) &= s_l^*(k, i) + s_r^*(k, i) + 1 \quad \text{for } i = M - N + 1 \\ 1 &\leq s_f(k) \leq 2N - 1 \end{aligned} \quad (3.5)$$

Local and running values of the total saliences $s(k, i)$ and $s^*(k, i)$ are intermediaries only that are used to reduce the computational load (see section 3.4).

3.3.2.2 Partial within-window saliency allocation

Let us consider the analysis window $w_N(i)$ of length N and origin i . The goal of this partial saliency allocation consists in determining the local saliencies of samples within the analysis window. By default, all local saliencies are initialized with their minimal value (0 for left and right local saliencies and 1 for local total saliencies). The term “partial” indicates that all samples are not guaranteed to obtain a saliency value after that process.

Partial saliency allocation exploits a feature of the tabular representation in Figure 3.5b. Instead of computing maxima over intervals of increasing length 2, 3, ..., one determines the maximum over frames of decreasing lengths. The steps of that partial saliency allocation algorithm are summarized in the flow diagram of Figure 3.6a and detailed hereafter :

1. Initialize : Let n be the length of the interval over which the maximum is computed. Initially, $n = N$.
2. Compute the position j of the maximum : one determines the absolute position j of the maximum within the interval $i \leq j \leq i + n - 1$.
3. Assign local saliencies to the maximum : The local saliencies of this maximum are respectively $s(j, i) = n$, $s_l(j, i) = j - i$ and $s_r(j, i) = i + n - j - 1$.
4. Test : if the maximum is the first sample of the window, the procedure stops. Otherwise, a new maximum computation is carried out. For that, notice that j would remain the position of the maximum as long as the considered interval starting from i has a length $n > j - i$. Therefore, at the next step, only the maximum over the $j - i$ first windows sample is computed. So, $n \rightarrow j - i$ and looping back to step 2 until $j = i$.

Figure 3.6b illustrates the application of these rules to the same array as illustrated in Figure 3.5b. One observes that partial saliency allocation always assigns local saliency values to the origin i of the sliding window. Moreover, that origin i is guaranteed to obtain at this stage its maximal right saliency.

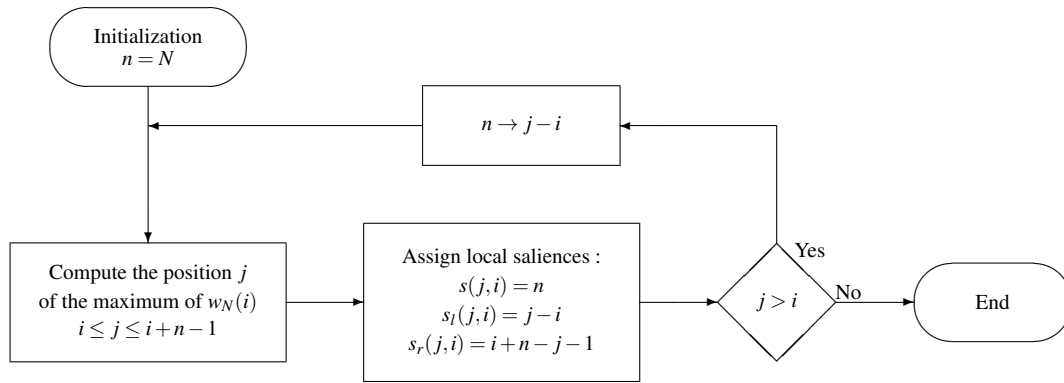
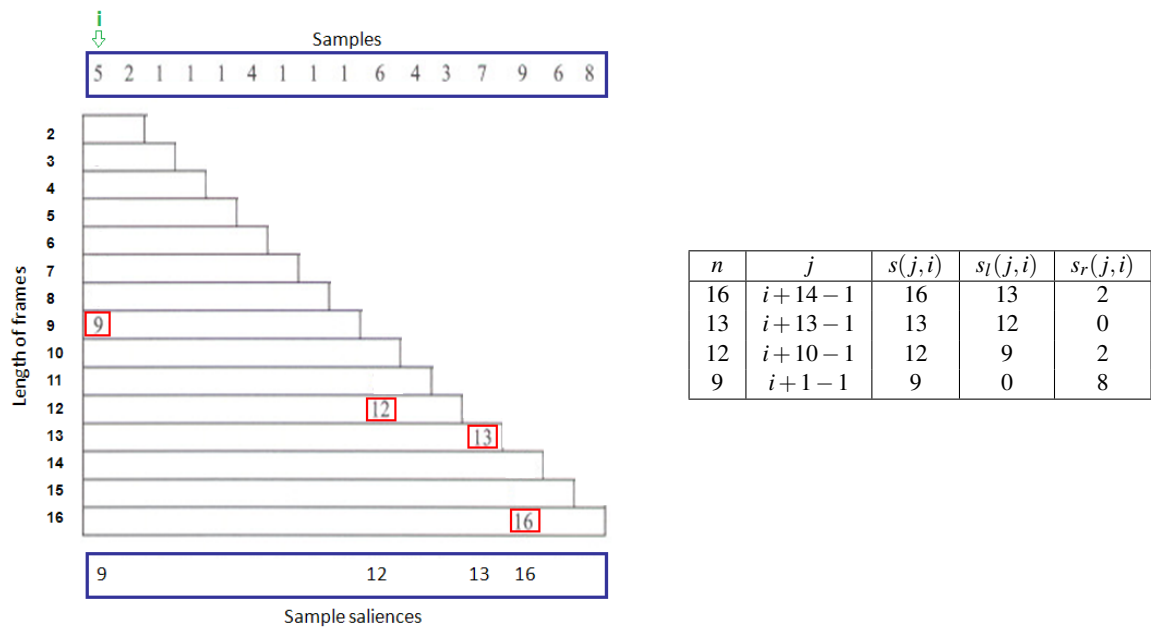
(a) Flow chart : Partial salience allocation applied to an analysis window of length N (b) Example : Partial salience allocation applied to an analysis window of length $N = 16$

Figure 3.6 – Partial salience allocation

3.3.2.3 Global saliency allocation

All samples of the array of length M are not assigned a saliency value via the partial saliency allocation algorithm and the problems related to the array origin are not solved either. Therefore, the window $w_N(i)$ (of length $N < M$ and origin i), is moved sample-by-sample along the array. Local saliencies $s(k, i)$, $s_l(k, i)$ and $s_r(k, i)$ are then obtained with reference to the sliding window and running saliencies s^* , s_l^* and s_r^* are updated for each window position. The steps, summarized in Figure 3.7, are the following:

1. Initialization : all saliencies s^* are put equal to 1 and all saliencies s_l^* and s_r^* equal to zero.
2. Application of the partial saliency allocation for position i of the sliding window $w_N(i)$, obtaining the local $s(k, i)$, $s_l(k, i)$ and $s_r(k, i)$ saliencies, $k = i, \dots, i + N - 1$.
3. For each sample k , updating of the running saliencies as follows $s^* = s(k, i)$, $s_l^* = s_l(k, i)$, $s_r^* = s_r(k, i)$ if their values increase. Otherwise, no updating.
4. Shifting of the window to the right by one sample ($i \rightarrow i + 1$) and looping to step 2.

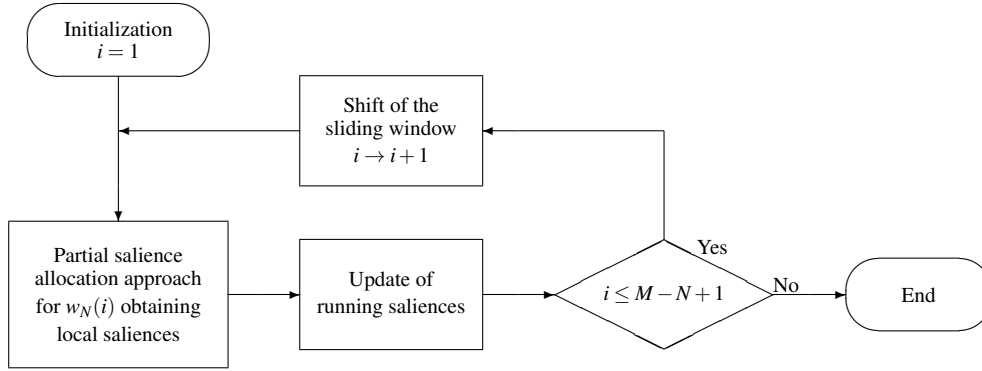


Figure 3.7 – Global saliency allocation (flow chart)

When the right-hand side of the sliding window has reached the right boundary of the array ($i = M - N + 1$), the final sample saliency values s_f are given by :

$$s_f(k) = s_r^*(k, i) + s_l^*(k, i) + 1 \quad \text{for } i = M - N + 1 \quad (3.6)$$

At this stage, each sample has been assigned a final saliency because the sliding window is hopped sample-by-sample and because the partial saliency allocation always updates the saliency value assigned to the left-most sample in the sliding window. Each sample will also obtain its maximum left saliency for the window position for which it is a maximum over all its left neighbours in the window. The relevance of the final saliencies, obtained from the left and right running saliencies, is discussed in section 3.5.

It is recommended to discard the $N - 1$ first and the $N - 1$ last samples of the saliency analysis array, because the values are conditioned by the array boundaries.

3.4 Reduction of computational load

One may show that the previous algorithm comprises steps that are redundant and that increase the computation time. Indeed, until now, all the steps of the partial salience allocation method have been applied for each position of the sliding window. The salience allocation may therefore be speeded up further by carrying out additional tests on the saliences obtained at previous sliding window positions. These additional steps will be reported hereafter.

A speeding up follows from the property of the partial salience allocation, for which the first sample of the analysis window is guaranteed to have an updated salience value. Let i be the sample index of the current window origin and $s^*(i, i-1)$ its running salience value obtained for all previous positions of the sliding window, so that :

$$s^*(i, i-1) = \max_{j < i} (s(i, j)) \quad (3.7)$$

By definition, that running salience can only increase monotonically. Therefore, one knows that this sample i has been previously a maximum over an interval of length $s^*(i, i-1)$. As a consequence, the computation of the maximum over intervals of lengths $1, \dots, s^*(i, i-1)$ is not required.

The final salience allocation method, illustrated in Figure 3.8, is thus modified as follows :

1. Initialization : all saliences s^* are put equal to 1 and all saliences s_l^* and s_r^* equal to zero.
2. Determination of the running salience value $s^*(i, i-1)$ assigned to the origin sample i of the sliding window $w_N(i)$.
3. Application of the partial salience allocation for position i of the sliding window $w_N(i)$, obtaining the local $s(k, i)$, $s_l(k, i)$ and $s_r(k, i)$ saliences, $k = i, \dots, i+N-1$. However, the partial salience allocation method is applied only for frames of length $n = N, N-1, \dots, s^*(i, i-1) + 1$
4. For each sample k , updating of the running saliences as follows $s^* = s(k, i)$, $s_l^* = s_l(k, i)$, $s_r^* = s_r(k, i)$ if their values have increased. Otherwise, no updating.
5. Shifting of the window to the right by one sample ($i \rightarrow i+1$) and looping to step 3.
6. At the end, determination of the final sample salience values and discarding of the $N-1$ first and $N-1$ last salience array samples.

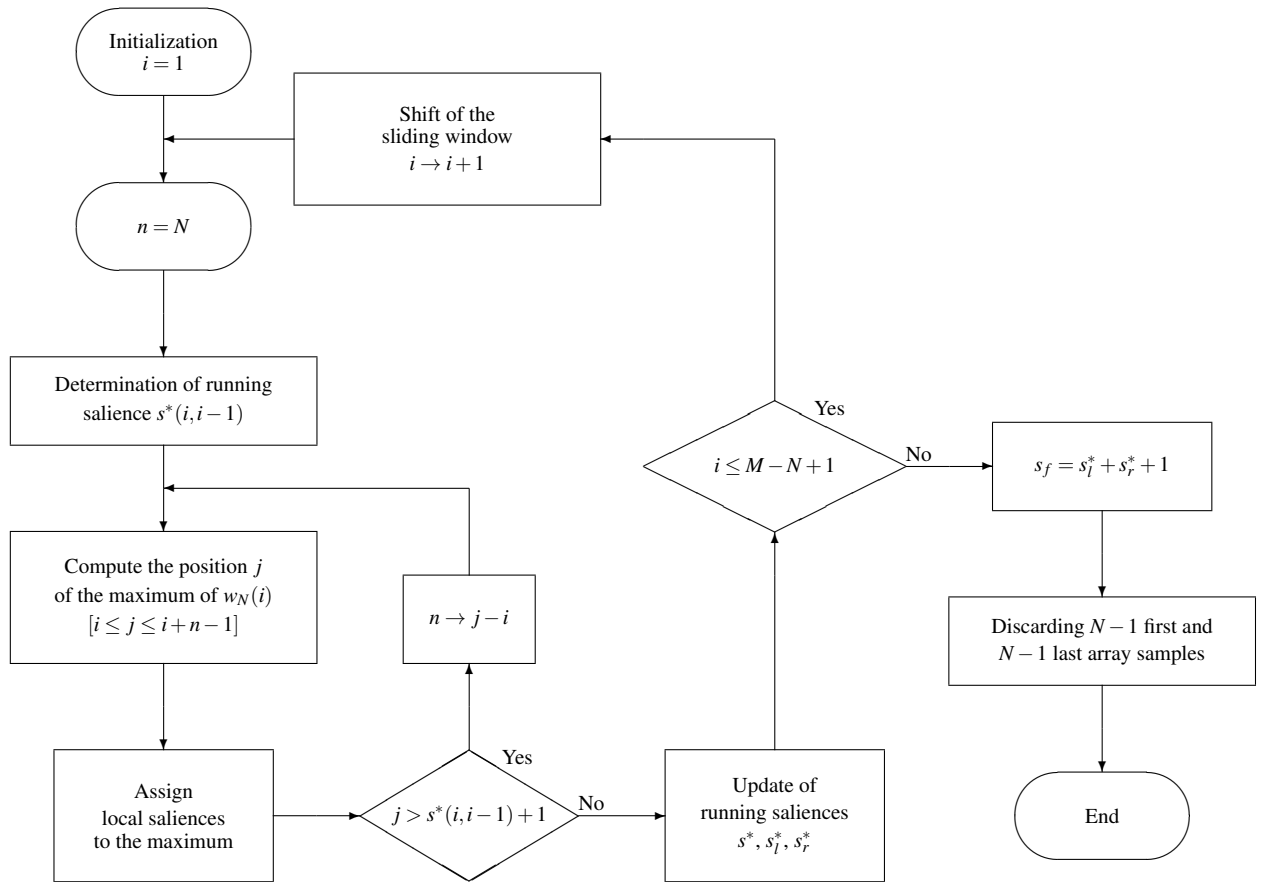


Figure 3.8 – Saliency allocation method (flow chart)

3.5 Validation

3.5.1 Preliminary remarks

The sample salience allocation relies on the computation of array maxima. Therefore, the results may be directly affected by the implementation rules used for the computation of maxima. In this study, a maximum is defined as the array sample the value of which is larger than the others. If several identical array maxima exist, the left-most maximum only is taken into account.

3.5.2 Theoretical developments

Let us consider a signal of length M and a sliding window of length $N < M$. The goal of the following developments consists in determining the salience values assigned to a sample k on the basis of the window-based salience allocation method described in section 3.3.2. For that, one assumes that the sample k is located far from the array boundaries ($N \leq k \leq M - N + 1$) so that these do not condition the assigned salience values.

For the sake of simplicity, consider the signal $x(n)$ of length M where δ is a unit impulse, $j < k < l$ and A_j, A_k, A_l the respective amplitudes. By considering different amplitudes and positions of these three samples, one may illustrate all possible cases that the algorithm has to take into account during salience allocation.

$$x(n) = A_j \cdot \delta(n - j) + A_k \cdot \delta(n - k) + A_l \cdot \delta(n - l) \quad (3.8)$$

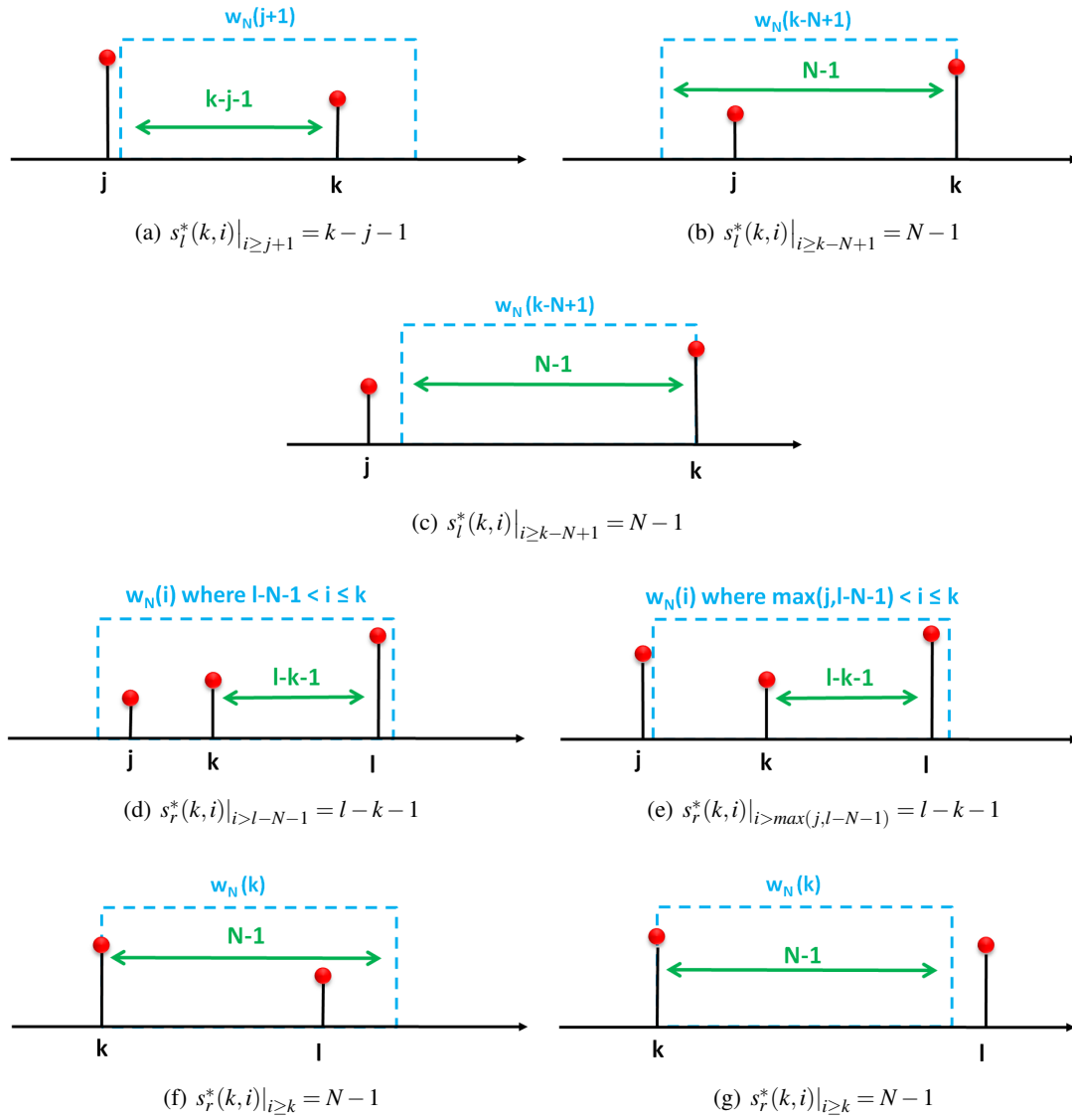
By using a sliding analysis window $w_N(i)$ of length N and origin i , the running salience values are updated if their values are increased for a window position. Below are given all possible maximal values of running left and right saliences $s_l^*(k, i)$ and $s_r^*(k, i)$ obtained for sample k for an critical position of the sliding analysis window. The term “critical” refers here to the window position required to update the corresponding running saliences.

$$s_l^*(k, i) = \left\{ \begin{array}{c|c|c} \text{Salience values} & \text{Conditions} & \text{Critical } w_N(i) \\ \hline k-j-1 & (A_j \geq A_k) \text{ and } (k-j+1 \leq N) & i = j+1 \\ \hline N-1 & (A_j < A_k) \text{ or } (k-j+1 > N) & i = k-N+1 \end{array} \right. \quad (3.9)$$

$$s_r^*(k, i) = \left\{ \begin{array}{c|c|c} \text{Salience values} & \text{Conditions} & \text{Critical } w_N(i) \\ \hline l-k-1 & (A_l > A_k) \text{ and } (l-k+1 \leq N) & l-N-1 < i \leq k \quad \text{if } A_j < A_k \\ \hline N-1 & (A_l \leq A_k) \text{ or } (l-k+1 > N) & \max(j, l-N-1) < i \leq k \quad \text{either} \\ & & i = k \end{array} \right. \quad (3.10)$$

Figure 3.9 illustrates all possible situations.

One observes that the maximal values of running left and right saliences are not obtained for the same analysis window position. The final salience value $s_f(k)$ is thus obtained at the end of the procedure by summing the obtained running left and right salience values, as explained in section 3.3.2.3. Hereafter are reported all the possible values of $s_f(k)$ in each situation :

Figure 3.9 – Running left and right salience values obtained for sample k

$$s_f(k) = \left\{ \begin{array}{c|c} \text{Values} & \text{Conditions} \\ \hline l-j-1 & [(A_j \geq A_k) \text{ and } (k-j+1 \leq N)] \\ & [(A_l > A_k) \text{ and } (l-k+1 \leq N)] \\ \hline 2N-1 & [(A_l \leq A_k) \text{ or } (l-k+1 > N)] \\ & [(A_j < A_k) \text{ or } (k-j+1 > N)] \\ \hline k+N-j-1 & [(A_j \geq A_k) \text{ and } (k-j+1 \leq N)] \\ & [(A_l \leq A_k) \text{ or } (l-k+1 > N)] \\ \hline l+N-k-1 & [(A_l > A_k) \text{ and } (l-k+1 \leq N)] \\ & [(A_j < A_k) \text{ or } (k-j+1 > N)] \end{array} \right. \quad (3.11)$$

The salience values obtained via the sliding analysis window are hereafter compared with the theoretical salience values $s(k)$, obtained on the basis of the salience definition.

$$s(k) = \left\{ \begin{array}{c|c} \text{Values} & \text{Conditions} \\ \hline l-j-1 & A_j \geq A_k \\ & A_l > A_k \\ \hline M & A_j < A_k \\ & A_l \leq A_k \\ \hline M-j & A_j > A_k \geq A_l \\ \hline l-1 & A_j < A_k < A_l \end{array} \right. \quad (3.12)$$

One observes that the algorithm yields saliences values $s_f(k)$ identical to the expected salience values $s(k)$ as long as the true left and right sample saliences are $< N$ (i.e. the analysis window length). When one of the true left or right running saliences $\geq N$, then the sample salience acquires an in-between value. When both true saliences $\geq N$, then the sample salience equals $2N-1$.

To sum up, the sliding window length N must be choosen so as to minimize the loss of information owing to the array boundaries and maximize the relevance of the salience(s) with regard to the goal of the analysis.

3.6 Example

3.6.1 Application to an arbitrary array

One considers a basic signal x of length $M = 16$.

$$x = [6 \ 2 \ 1 \ 3 \ 4 \ 5 \ 3 \ 1 \ 6 \ 7 \ 8 \ 9 \ 10 \ 8 \ 1 \ 3]$$

The sample salience allocation method is applied with a sliding analysis window of length $N = 5$. Figure 3.10 illustrates the $M - N + 1$ positions of the sliding window and the running salience allocation.

A color code summarizes the algorithm steps :

- White : Default. Ignored frame.
- Green : Frame which has been used for the computation of a maximum. (black dot indicates the maximum position)
- Orange : Redundant step. Frame which has not been considered during partial salience allocation to decrease computational time.

The salience allocation results are illustrated in Figure 3.11 and summarized below :

$$\begin{aligned} s_l^* &= \left[\begin{array}{cccc|cccccccc|cccc} 0 & 0 & 0 & 2 & 3 & 4 & 0 & 0 & 4 & 4 & 4 & 4 & 4 & 4 & 0 & 0 & 0 & 0 \end{array} \right] \\ s_r^* &= \left[\begin{array}{cccc|cccccccc|cccc} 4 & 1 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \end{array} \right] \\ s_f &= \left[\begin{array}{cccc|cccccccc|cccc} 5 & 2 & 1 & 3 & 4 & 7 & 2 & 1 & 5 & 5 & 5 & 5 & 5 & 5 & 8 & 1 & 1 & 1 \end{array} \right] \end{aligned}$$

Figure 3.11 illustrates also the true values which have been obtained by visual inspection. One observes that, in the central part of the graph (where the sample saliences are not affected by the boundaries), theoretical left, right and final salience values have been assigned to samples 5, 6, 7 and 8. However, samples 9, 10, 11, 12 have decreased final salience values. This is explained by the thresholding of the left saliences to a value equal to $N - 1$.

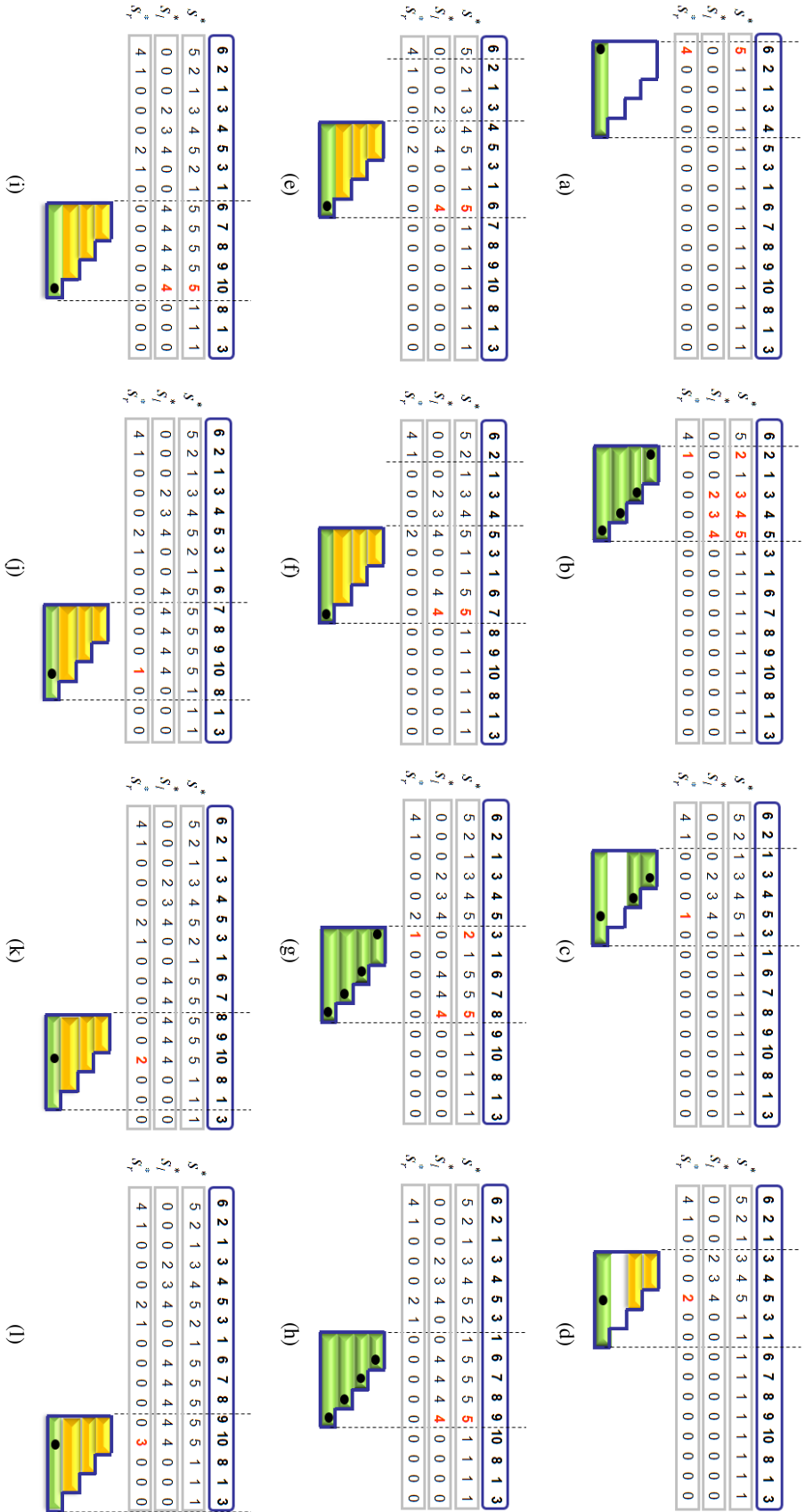


Figure 3.10 – Running salience allocation to an example signal for successive positions of the sliding window

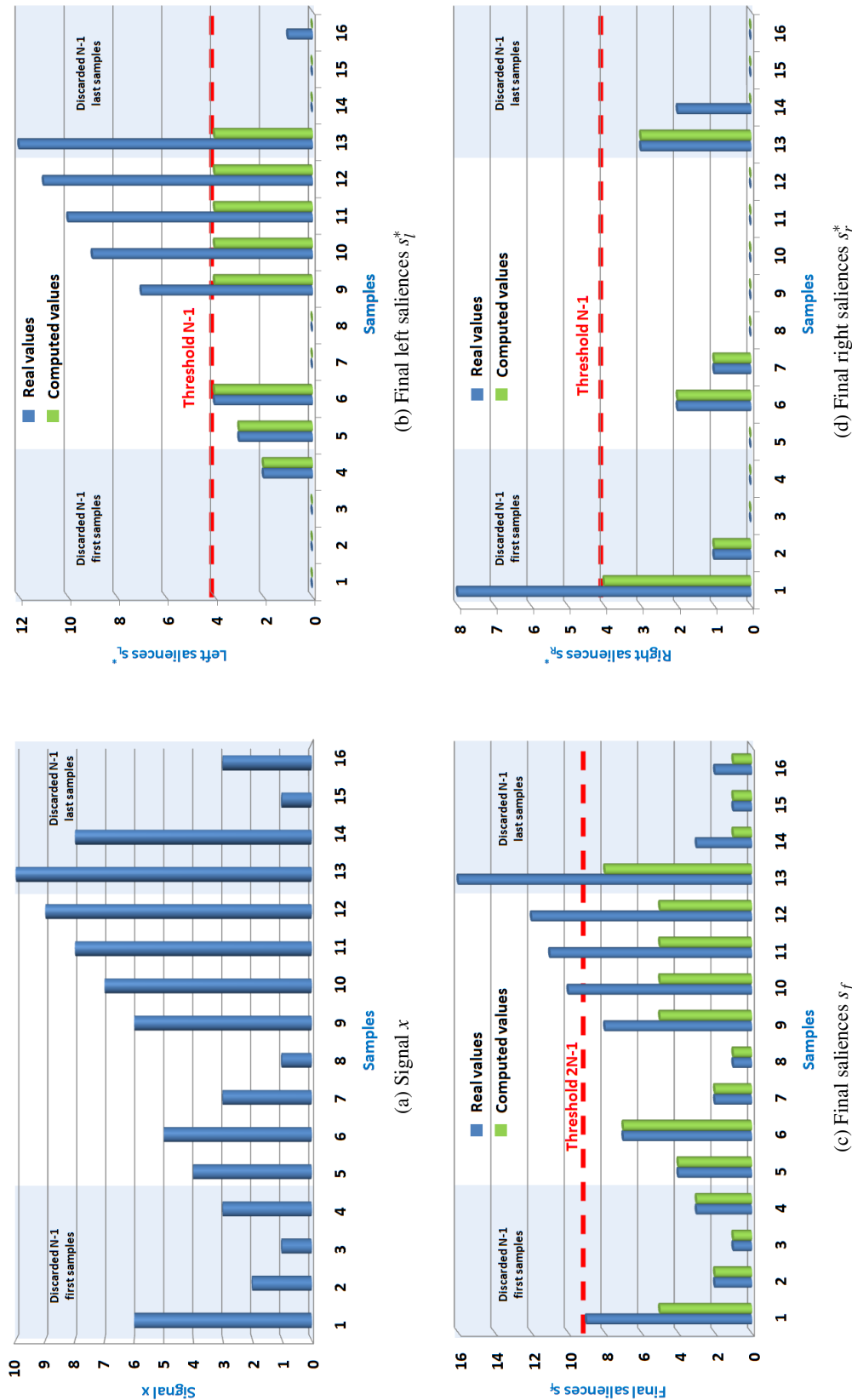


Figure 3.11 – Saliency allocation : application to an example signal

3.6.2 Application to a voiced speech sound

Figure 3.12 illustrates the results of salience allocation for a vowel [a] sustained by a normophonic or a dysphonic speaker.

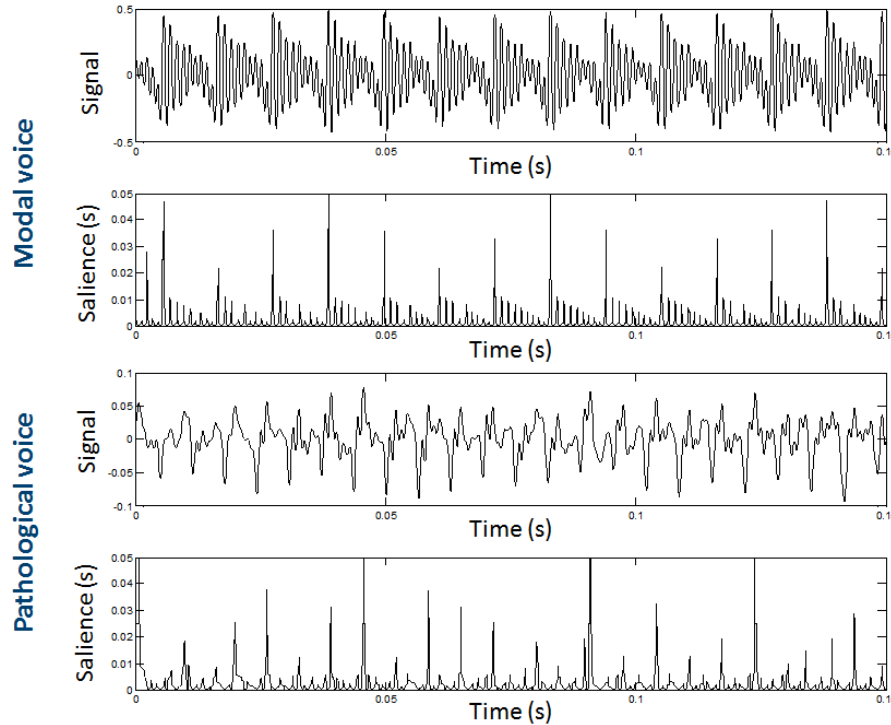


Figure 3.12 – Salience allocation : application to a voiced speech sound

For each speaker, the upper picture illustrates the speech sound and the lower picture reports the final salience values (expressed in seconds) assigned to each sample. These results have been obtained with a sliding window of 25ms. (More details will be given in chapter 5). One observes that the speech cycle peaks that are located in the vicinity of the maximal glottal excitation are characterized by large salience values. The salience can therefore be used to detect voiced cycles automatically.

3.7 Conclusions

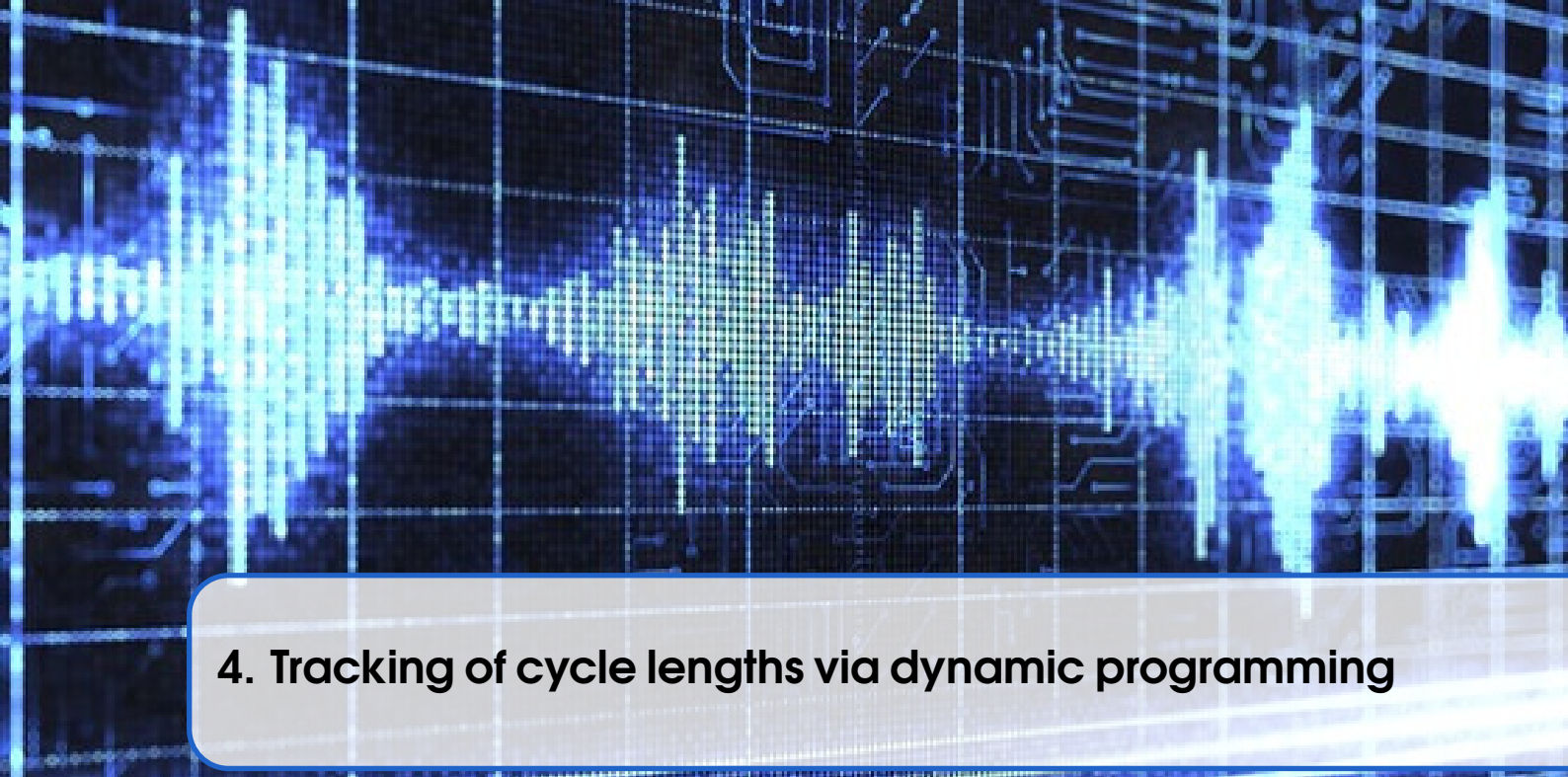
In this chapter, the concept of *sample salience* has been introduced to characterize the prominence of a signal sample relative to its surrounding. Different salience allocation methods have been proposed. The retained method is frame-based and relies on a partial salience allocation applied for each position of the analysis window and an update of the sample salience. It has been shown that this algorithm yields salience values identical to the theoretical values as long as the distance between a signal sample and its neighbour with higher amplitude (or the array boundaries) is lower than the analysis window length. If this is not the case, the saliences may be clipped and acquire a lower value. Therefore, the length of the sliding analysis window has to be chosen so as to minimize the loss of information owing to the array boundaries and maximize the relevance of the saliences with regard to the goal of the further analysis. In addition, the proposal comprises several improvements to speed up the salience allocation process.

The concept of *salience* will be used in Chapters 4 and 5 to track cycle lengths via dynamic programming, and more specifically the glottal cycle lengths. Indeed, in voiced speech fragments,

speech cycles are often characterized by a prominent signal peak that is the effect of the glottal excitation. The salience of that peak is expected to be high irrespective of the evolving signal amplitude.

Key points

- The sample salience is defined as the length of the longest interval over which a sample is a maximum
- The salience is determined by the value of a sample relative to its neighbours. A sample with a large value has not necessarily a high salience, and vice-versa
- A frame-based sample salience allocation method is proposed to assign a salience value to each array sample
- The obtained salience values agree with the theoretical values as long as the distance between a signal sample and its neighbour with larger amplitude (or the array boundaries) is lower than the analysis window length



4. Tracking of cycle lengths via dynamic programming

Objectives of this chapter

- Propose a temporal method for the tracking of cycle lengths of which no strong assumptions are made with regard to their regularity
- Describe and illustrate the proposed method that relies on a dynamic programming algorithm based on peak saliences and inter-peak durations

Contents

4.1	Introduction	63
4.2	Overview	63
4.3	Problem formulation	63
4.3.1	Topology	63
4.3.2	Stages	64
4.3.3	States	65
4.3.4	Initialization	66
4.3.5	Optimal path search	67
4.3.6	Backtracking	67
4.4	Application to an example array	68
4.4.1	Initialization of the optimization network	68
4.4.2	Optimal path search	70
4.5	Conclusions	74

4.1 Introduction

A method is proposed to track cycles in the temporal domain via a multi-scale analysis that assigns a salience to each signal peak, and which is founded on dynamic programming. The goal consists in selecting cycle peaks that characterize the same event. The method does not rest on the assumptions that the signal is locally periodic and the average period length is known a priori. The only assumption is that the average cycle frequency F_0 is comprised between two fixed limits, $F_{0,min}$ and $F_{0,max}$.

4.2 Overview

Dynamic programming is an optimization method that transforms a complicated problem into a sequence of simpler sub-problems. The topology of dynamic programming comprises several elements, usually called *stages* and *states* [BHM77].

A general characteristic of dynamic-programming is the development of a recursive optimization procedure, which arrives at a solution of the overall problem by first solving a one-stage problem and sequentially including additional stages, one at a time, and solving one-stage problems until the overall optimum has been found. This approach refers to the principle of optimality that stipulates that any optimal policy has the property that, whatever the current state and decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the current decision.

The optimization problem is structured into multiple stages which are solved sequentially. Each stage is defined as a point of the structure where an ordinary optimization problem has to be solved and a decision made. That decision helps to define the characteristics of the next one-stage problem in the sequence. Often, the stages represent different temporal events within the problem's planning horizon. Associated with each stage are the states of the process. The states reflect the information required to fully assess the consequences that the current decision has upon future actions. They should convey enough information to enable future decisions without regard to how the process reached the current state.

Here, the optimization consists in considering several candidate cycle length time series obtained by means of the retained peak distances and discovering via dynamic programming the length series that has the smallest overall cycle duration perturbation. The candidate cycle length series are built by taking into account several signal peak sub-sequences on the base of the local inter-peak durations and the peak salience values, assuming that speech cycle peaks owing to the glottal excitation are characterized by large salience values.

4.3 Problem formulation

The cycle length tracking is based on speech signal peaks, that are characterized by their positions and their saliences. The goal consists in selecting a subset of speech signal peaks that characterize the same glottal events. A typical cycle length sequence is expected to report quasi constant distances between peaks characterized by high saliences.

4.3.1 Topology

The dynamic programming approach used for the tracking of cycle lengths consists in building a network of stages in which each node is a triplet of peaks. Each peak triplet may be isolated or

joined to other peak triplets, forming thus a sequence of peaks. The optimization consists in starting from a peak triplet located at the beginning of the array and determining the optimal path through the network to arrive at a peak triplet located at the end of the array. The decision rules involve the second order difference of inter-peak durations as well as the peak saliences.

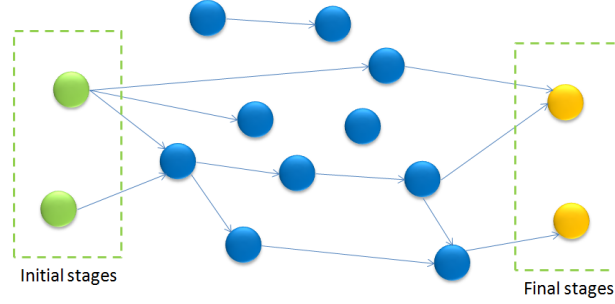


Figure 4.1 – Dynamic programming approach : Optimization network, each stage being a triplet of signal peaks

4.3.2 Stages

The stages are successions of 3 peaks g , h and i , called a triplet and denoted hereafter as $(g - h - i)$. The inter-peak distances have to satisfy the following conditions :

$$\begin{cases} d_{min} \leq d_{(g,h)} \leq d_{max} \\ d_{min} \leq d_{(h,i)} \leq d_{max} \\ \frac{1}{(1+\alpha)}d_{(g,h)} \leq d_{(h,i)} \leq (1+\alpha)d_{(g,h)} \end{cases} \quad (4.1)$$

Distances $d_{min} = \frac{F_s}{F_{0,max}}$ and $d_{max} = \frac{F_s}{F_{0,min}}$ designate respectively the minimal and maximal expected cycle lengths in samples. Symbol $d_{(g,h)}$ designates the distance between peaks g and h , and symbol α refers to the maximal local length perturbation the tracker is able to take into account. Figure 4.2 illustrates two examples of triplets ending at the same peak i .

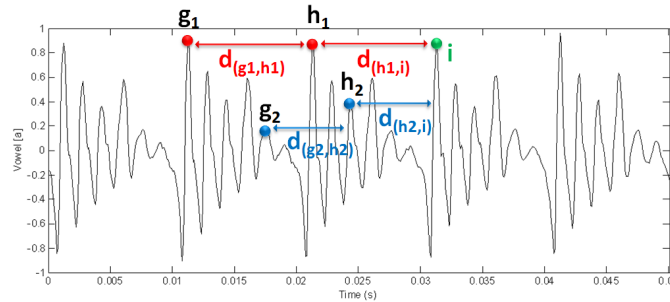


Figure 4.2 – Two triplets $(g_1 - h_1 - i)$ and $(g_2 - h_2 - i)$ ending at peak i

During optimization, several triplets are concatenated to build a triplet sequence. A triplet sequence is a succession of consecutive triplets which have two peaks in common (example : $(e - f - g)$, $(f - g - h)$, $(g - h - i)$, ...). The optimal triplet sequence therefore corresponds to the path involving peak triplets that give rise to the smallest overall cycle length perturbation.

As explained later, the stage decision consists in determining the optimal preceding triplet to build a triplet sequence. The use of peak triplets rather than peak couples is motivated by the fact that the first-order perturbations of the inter-peak distances may be assessed on the base of at least three local peak positions. The concatenation of two adjacent triplets informs about the positions of 4 peaks and thus 3 inter-peak distances. Therefore, the local second-order perturbations may be assessed via the second-order difference of these 3 distances. By using the second order difference within a decision rule, the tracking of the peaks is not affected by a linear trend of the inter-peak distance time series owing to intonation/declination.

4.3.3 States

For each triplet $(g - h - i)$, the optimization problem consists in determining the best preceding triplet $(f^* - g - h)$ among a set of admissible preceding triplets $T_{(g,h,i)}$. Figure 4.3 illustrates several admissible preceding triplets $(f - g - h)$.

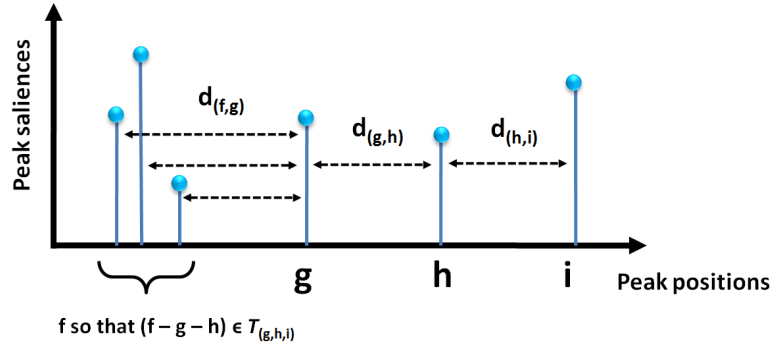


Figure 4.3 – Set of preceding triplets

The state, associated to the triplet (g, h, i) and denoted $C_{(g,h,i)}$, is a measurement of the overall length perturbation of the sequence ending at triplet $(g - h - i)$. The state variables are the second order inter-peak difference of the sub-sequence $[(f - g - h), (g - h - i)]$ and the peak salience $s(f)$ assigned to peak f . Let $(f^* - g - h)$ be the optimal preceding triplet, the state $C_{(g,h,i)}$ evolves as follows :

$$C_{(g,h,i)} = (C_{(f^*,g,h)} + c_{(f^*,g,h,i)}) \Big|_{(f^*-g-h) \in T_{(g,h,i)}} \quad (4.2)$$

$$\text{where : } c_{(f^*,g,h,i)} = \frac{(|2d_{(g,h)} - d_{(f^*,g)} - d_{(h,i)}| + 1)^{\gamma_1}}{s(f^*)^{\gamma_2}}$$

$c_{(f^*,g,h,i)}$ is the absolute length perturbation increment of the triplet sequence. Positive parameters γ_1 and γ_2 assign different weights to the perturbations and saliences. If $\gamma_1 > \gamma_2$, the second-order perturbation of the sequence is a key factor for the choice of the best preceding triplet. Otherwise, a greater weight is assigned to the peak salience. In this study, the cycle length tracking relies on the selection of peaks that characterize the same glottal event, i.e. the peaks that are located in the vicinity of the maximal glottal excitation that are expected to have higher saliences. Moreover, pathological voices are expected to have higher cycle-to-cycle length perturbations. As a consequence, a higher weight is assigned to the salience so that $\gamma_1 = 1$ and $\gamma_2 = 2$. In the case of vanishing second-order perturbations, the “+1” in the numerator of c enables to take into account the contribution of the peak salience.

Additional state variables, denoted hereafter $L_{(g,h,i)}$ and $\mu_{(g,h,i)}$, are also introduced :

- $L_{(g,h,i)}$ reports the number of cycles in the sequence ending at triplet (g,h,i) . L is important because a triplet sequence with long inter-peak distances is expected to have a smaller overall perturbation C than a triplet sequence with short inter-peak distances, because of the variable number of cycles. Moreover, proper use of L in the path search enables the tracking of vocal cycle lengths in the presence of moderate diplophonia.
- $\mu_{(g,h,i)}$ reports the average inter-peak distance of the sequence ending at triplet (g,h,i) . This state variable is a safeguard that avoids that distant triplets in the sequence have inter-peak distances that are too different. As an example, Figure 4.4 illustrates a triplet sequence for which the inter-peak distance $d_{(d,e)}$ is much higher than the distance $d_{(a,b)}$ although each triplet satisfies condition (4.1).

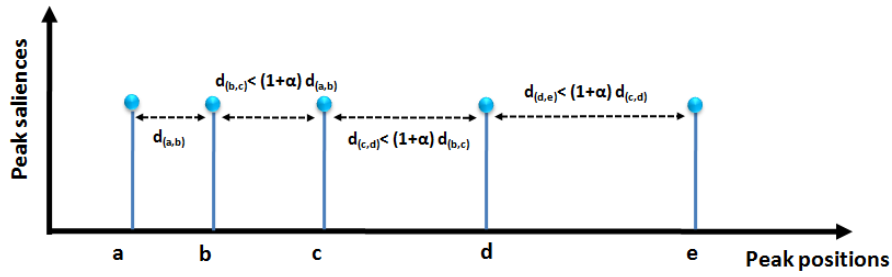


Figure 4.4 – Triplet sequence with increasing inter-peak distances

Here, the average inter-peak distance $\mu_{(g,h,i)}$ is used during the optimal path search to limit the set of admissible preceding triplets $T_{(g,h,i)}$ as follows :

$$(f - g - h) \in T_{(g,h,i)} \Leftrightarrow \frac{1}{1 + \gamma_3} \mu_{(f,g,h)} \leq d_{(h,i)} \leq (1 + \gamma_3) \mu_{(f,g,h)} \quad (4.3)$$

Parameter γ_3 is typically chosen in the interval $[0.5, 1]$. A value < 0.5 is not recommended when cycle lengths are tracked in presence of intonation/declination.

4.3.4 Initialization

Let's consider an array of length M . First, one determines the positions and saliences of the signal peaks. The initialization consists in determining all the peak triplets which satisfy conditions (4.1) and identifying its preceding candidate triplets that have 2 peaks in common with the peak triplet. On the basis of the triplet position and the number of candidate predecessors, the triplet states are then initialized as follows :

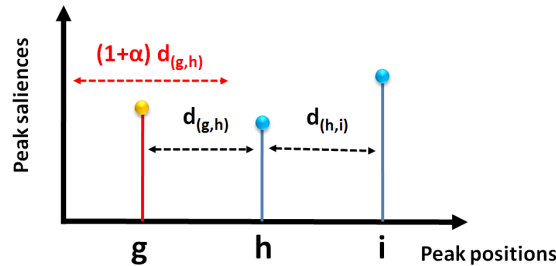


Figure 4.5 – Initialization of triplet states

- If the triplet $(g - h - i)$ has no candidate predecessor and is located at the beginning of the array so that the peak position g satisfies the condition $g \leq (1 + \alpha)d_{(g,h)}$ (Figure 4.5), its state variables are initialized as follows :

$$\begin{cases} C_{(g,h,i)} = 0 \\ L_{(g,h,i)} = 2 \\ \mu_{(g,h,i)} = \frac{d_{(g,h)} + d_{(h,i)}}{2} \end{cases} \quad (4.4)$$

These triplets constitute the possible initial stages of the triplets sequences.

- Otherwise, the states are initialized at their default values: $C_{(g,h,i)} = \infty$, $L_{(g,h,i)} = 0$ and $\mu_{(g,h,i)} = \infty$

Moreover, all the triplets $(g - h - i)$ that are located at the end of the array so that $i \geq M - (1 + \alpha)d_{(h,i)}$ and have no successor are marked as “final”. They constitute the possible final stages of the candidate triplet sequences.

4.3.5 Optimal path search

Optimal search via dynamic programming here consists in finding the path involving peak triplets that give rise to the smallest perturbations of the inter-peak durations. For that, all peak triplets are ranked by increasing position of their last peak and then considered sequentially. For each peak triplet $(g - h - i)$, the selection of its best preceding triplet $(f^* - g - h)$ among the set of admissible preceding triplets $T_{(g,h,i)}$ relies on inter-peak distances, saliences as well as the state variables $L_{(g,h,i)}$. The one-stage decision consists here in determining the triplet sequence ending at triplet $(g - h - i)$ which minimizes the overall length perturbation considering the length of the sequence. Let $(f^* - g - h)$ be the best preceding triplet of $(g - h - i)$ which belongs to this sequence. This best predecessor has to satisfy the following condition :

$$(f^* - g - h) = \arg \min \left(\frac{C_{(f,g,h)} + c_{(f,g,h,i)}}{(L_{(f,g,h)} + 1)^{\gamma_4}} \right) \Big|_{(f-g-h) \in T_{(g,h,i)}} \quad (4.5)$$

The purpose of the denominator is to disfavor cycle omissions. Parameter γ_4 has been fixed to 3 to enable tracking cycle length time series characterized by high cycle-to-cycle length perturbations.

The link to this optimal preceding triplet is kept in memory and the states of $(g - h - i)$ are then updated on the basis of that optimal predecessor as follows :

$$\begin{cases} C_{(g,h,i)} = (C_{(f^*,g,h)} + c_{(f^*,g,h,i)}) \\ L_{(g,h,i)} = L_{(f^*,g,h)} + 1 \\ \mu_{(g,h,i)} = \frac{(\mu_{(f^*,g,h)} L_{(f^*,g,h)} + d_{(h,i)})}{L_{(f^*,g,h)} + 1} \end{cases} \quad (4.6)$$

A triplet (initial or not) with no predecessor will keep its default initial state values.

4.3.6 Backtracking

When all triplet states have been updated, the peak triplet $(g^* - h^* - i^*)$ giving rise to a minimal ratio C/L^{γ_4} is kept and marked as “final”.

The peak triplet sequence corresponding to this optimal final triplet is then recovered by backtracking, starting from the optimal triplet ($g^* - h^* - i^*$) and adding the memorized optimal preceding peak triplet, and so on.

4.4 Application to an example array

Let's consider an array of length 85 for which 14 array peak positions (labelled a, b, c, \dots, n) have been determined :

Position label	a	b	c	d	e	f	g	h	i	j	k	l	m	n
Position value	1	11	20	22	29	31	33	40	46	51	60	71	77	81

4.4.1 Initialization of the optimization network

The parameter values for the triplet determination have been fixed to : $d_{min} = 5$, $d_{max} = 21$ and $\alpha = 0.35$. Table 4.1 reports the inter-peak distances.

	a	b	c	d	e	f	g	h	i	j	k	l	m
b	10												
c	19	9											
d	21	11	2										
e	28	18	9	7									
f	30	20	11	9	2								
g	32	22	13	11	4	2							
h	39	29	20	18	11	9	7						
i	45	35	26	24	17	15	13	6					
j	50	40	31	29	22	20	18	11	5				
k	59	49	40	38	31	29	27	20	14	9			
l	70	60	51	49	42	40	38	31	25	20	11		
m	76	66	57	55	48	46	44	37	31	26	17	6	
n	80	70	61	59	52	50	48	41	35	30	21	10	4

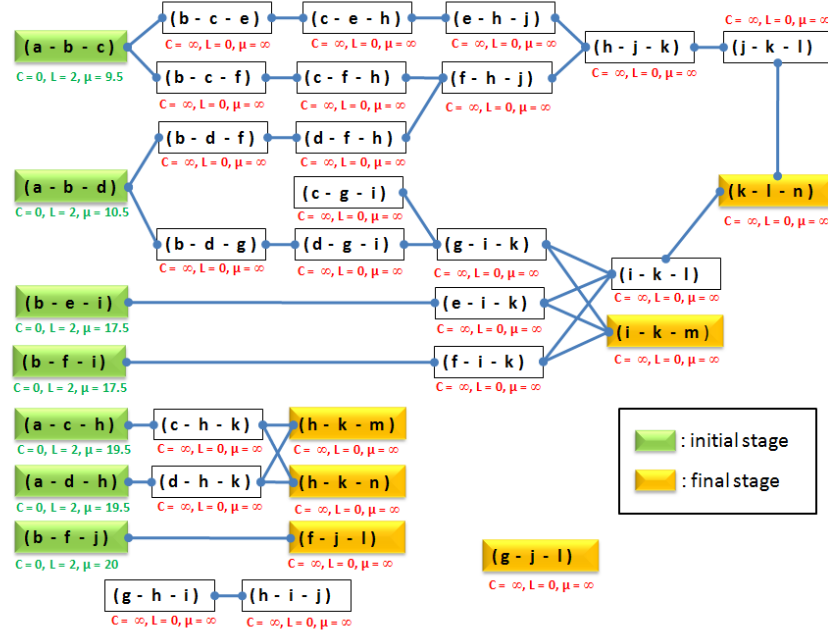
Table 4.1 – Example : Inter-peak distances

The pair of peaks (marked in red in the table) the inter-peak distance of which is $< d_{min}$ or $> d_{max}$ are discarded. For the remaining peak doublets, one determines all the possible doublet sequences to build a triplet satisfying conditions (4.1). Table 4.2 reports the so-obtained triplets as well as their candidate preceding triplets. On the basis of the triplet positions and the number of triplet predecessors, the initial and final stages of the interconnection network are identified and appropriate initial state values are assigned.

After initialization, the multistage interconnection network associated with the tracking problem is illustrated in Figure 4.6. One observes that there are several candidate sequences involving peak triplets that join the initial and final stages. Moreover, the number of triplets in these sequences is different. Some triplet sequences like $(g - h - i)$ $(h - i - j)$ will never belong to any candidate sequence because the first triplet of the sequence is not marked as initial. The same applies to isolated triplets like $(g - j - l)$.

Triplets	Status	Candidate predecessors	C	L	μ
$(a-b-c)$	Initial		0	2	9.5
$(a-b-d)$	Initial		0	2	10.5
$(b-c-e)$		$(a-b-c)$	∞	0	∞
$(b-c-f)$		$(a-b-c)$	∞	0	∞
$(b-d-f)$		$(a-b-d)$	∞	0	∞
$(b-d-g)$		$(a-b-d)$	∞	0	∞
$(a-c-h)$	Initial		0	2	19.5
$(a-d-h)$	Initial		0	2	19.5
$(c-e-h)$		$(b-c-e)$	∞	0	∞
$(c-f-h)$		$(b-c-f)$	∞	0	∞
$(d-f-h)$		$(b-d-f)$	∞	0	∞
$(b-e-i)$	Initial		0	2	17.5
$(b-f-i)$	Initial		0	2	17.5
$(c-g-i)$			∞	0	∞
$(d-g-i)$		$(b-d-g)$	∞	0	∞
$(g-h-i)$			∞	0	∞
$(b-f-j)$	Initial		0	2	20
$(e-h-j)$		$(c-e-h)$	∞	0	∞
$(f-h-j)$		$(c-f-h)$ $(d-f-h)$	∞	0	∞
$(h-i-j)$		$(g-h-i)$	∞	0	∞
$(c-h-k)$		$(a-c-h)$	0	2	20
$(d-h-k)$		$(a-d-h)$	0	2	19
$(e-i-k)$		$(b-e-i)$	∞	0	∞
$(f-i-k)$		$(b-f-i)$	∞	0	∞
$(g-i-k)$		$(c-g-i)$ $(d-g-i)$	∞	0	∞
$(h-j-k)$		$(e-h-j)$ $(f-h-j)$	∞	0	∞
$(f-j-l)$	Final	$(b-f-j)$	∞	0	∞
$(g-j-l)$	Final		∞	0	∞
$(i-k-l)$		$(e-i-k)$ $(f-i-k)$ $(g-i-k)$	∞	0	∞
$(j-k-l)$		$(h-j-k)$	∞	0	∞
$(h-k-m)$	Final	$(c-h-k)$ $(d-h-k)$	∞	0	∞
$(i-k-m)$	Final	$(e-i-k)$ $(f-i-k)$ $(g-i-k)$	∞	0	∞
$(h-k-n)$	Final	$(c-h-k)$ $(d-h-k)$	∞	0	∞
$(k-l-n)$	Final	$(i-k-l)$ $(j-k-l)$	∞	0	∞

Table 4.2 – Example : Triplets, candidate predecessors and initial state values

Figure 4.6 – Example : Multistage interconnection network ($d_{min} = 5$, $d_{max} = 21$ and $\alpha = 0.35$)

4.4.2 Optimal path search

Optimal path search via dynamic programming is applied to the signal of length 85. For that, a salience value is assigned to each peak. In the first experiment, these salience values are considered constant and equal to $S = 10$. Let's remark that this approach is basic. Indeed, by definition, the salience value assigned to an array peak depends on the positions and values of samples (peaks or not) located in its vicinity. Therefore, here, the salience is considered to be a simple weight rather than a sample characteristic. The tracking involves parameters γ_i that are fixed as follows : $\gamma_1 = 1$, $\gamma_2 = 2$, $\gamma_3 = 0.6$, $\gamma_4 = 3$.

Figure 4.7 illustrates the optimal extracted peak sequence as well as the state values associated with each triplet in the network. The red lines point to the best predecessor of each triplet. The state values associated with the 6 final triplets are given in Table 4.3.

	$(g-j-l)$	$(f-j-l)$	$(h-k-m)$	$(i-k-m)$	$(h-k-n)$	$(k-l-n)$
C	∞	0.01	0.06	0.1	0.04	0.2
L	0	3	4	6	4	8
μ	∞	20	19	12.66	20	10
$\frac{C}{L^{\gamma_4}}$	/	3.7e-04	9.3e-04	4.6e-04	6.2e-04	3.9e-04

Table 4.3 – Example : State values associated to the 6 final triplets

One observes that the triplet sequence of length $L = 3$ and ending at triplet $(f-j-l)$ obtains the smallest overall length perturbation level C/L^{γ_4} . The average inter-peak duration is equal to 20 samples. One also observes that the overall length perturbation level associated with the longest sequence (of length $L = 8$) starting from $(a-b-c)$ and ending at $(k-l-n)$ is also feeble. Observe that the same results are obtained if the salience values assigned to the non-selected peaks are smaller than those of the selected peaks.

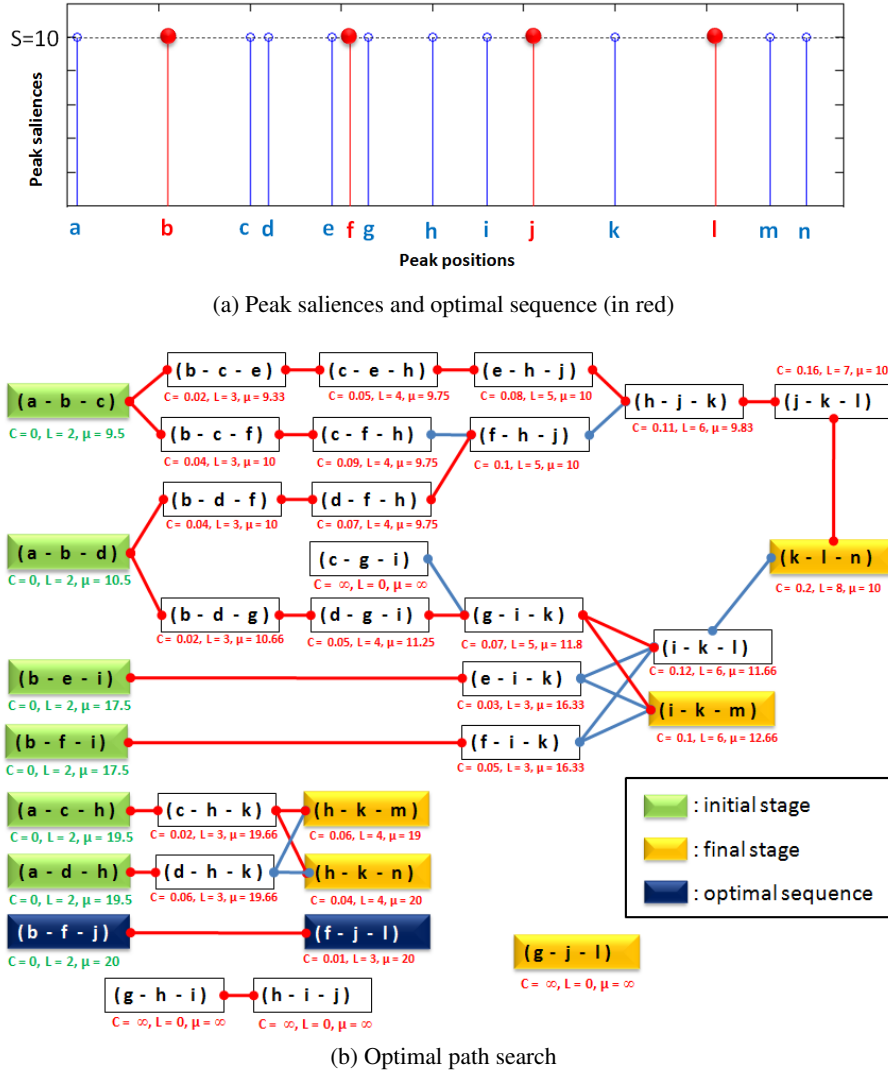
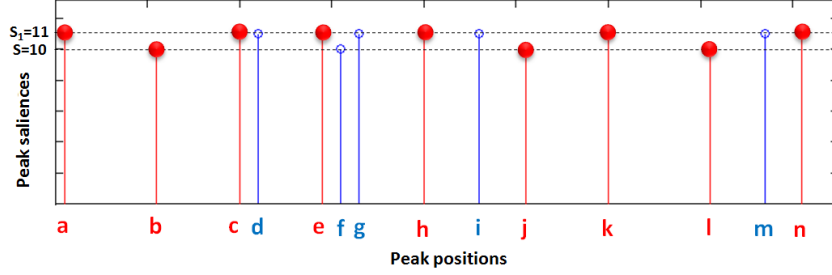
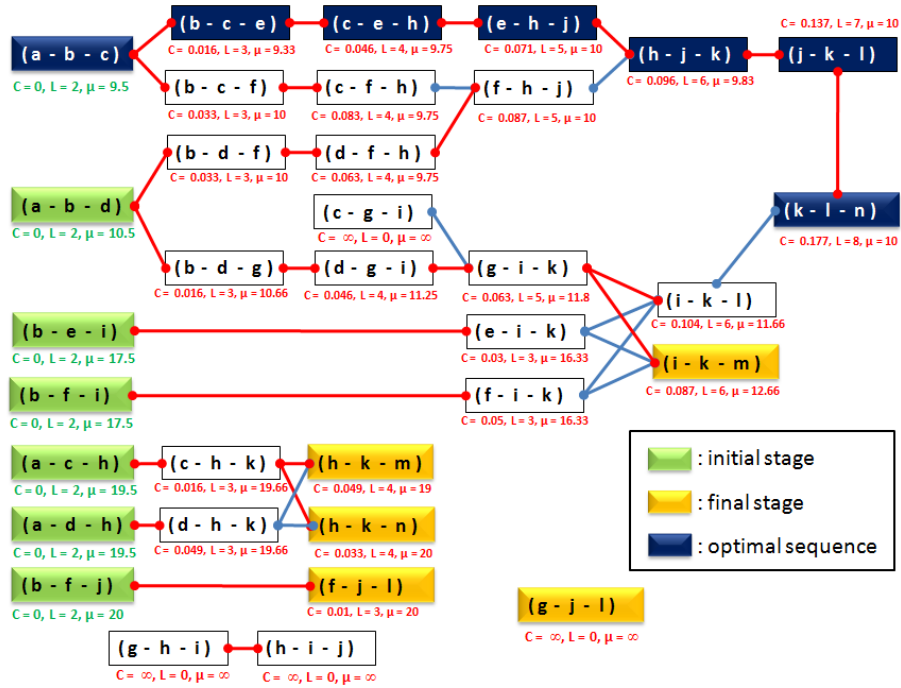


Figure 4.7 – Example : First experiment

The second experiment consists in assigning a salience $S_1 = 11$ to the previously unselected peaks. The results are illustrated in Figures 4.8a and 4.8b. One observes that the state values C associated with the triplets characterized by high saliences are smaller so that the previous triplet sequence of length $L = 8$ is now discovered.



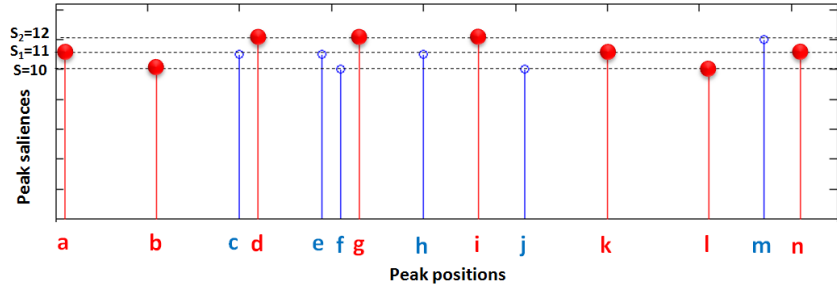
(a) Peak saliences and optimal sequence (in red)



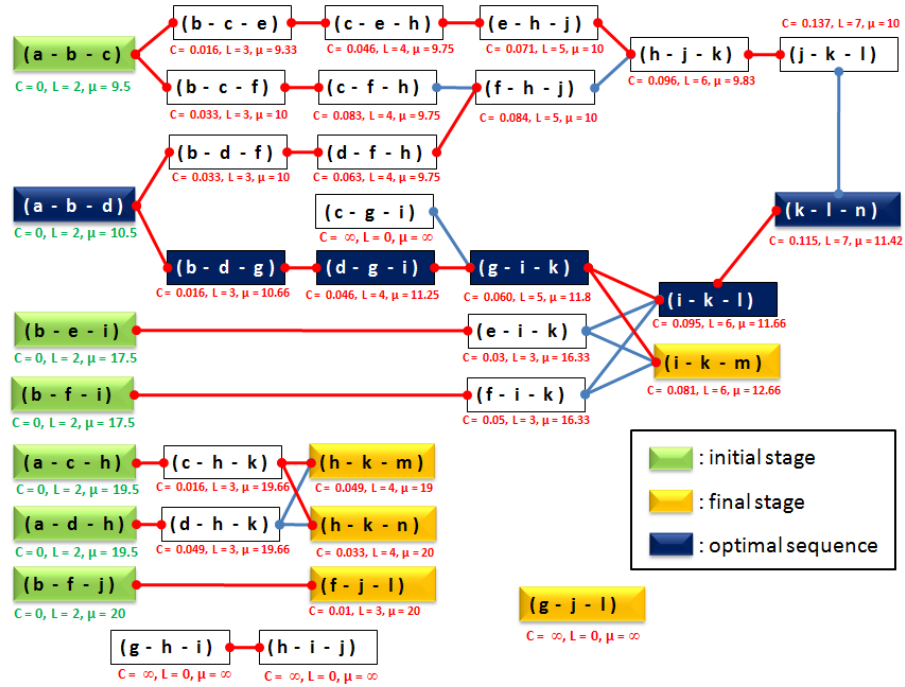
(b) Optimal path search

Figure 4.8 – Example : Second experiment

Finally, the third experiment consists in assigning a salience $S_2 = 12$ to all array peaks that have not been selected during the two preceding experiments. Figures 4.9a and 4.9b illustrate the results. In experiment 3, a peak re-affiliation occurs so that a triplet sequence of length $L = 7$ is considered optimal.



(a) Peak saliences and optimal sequence (in red)



(b) Optimal path search

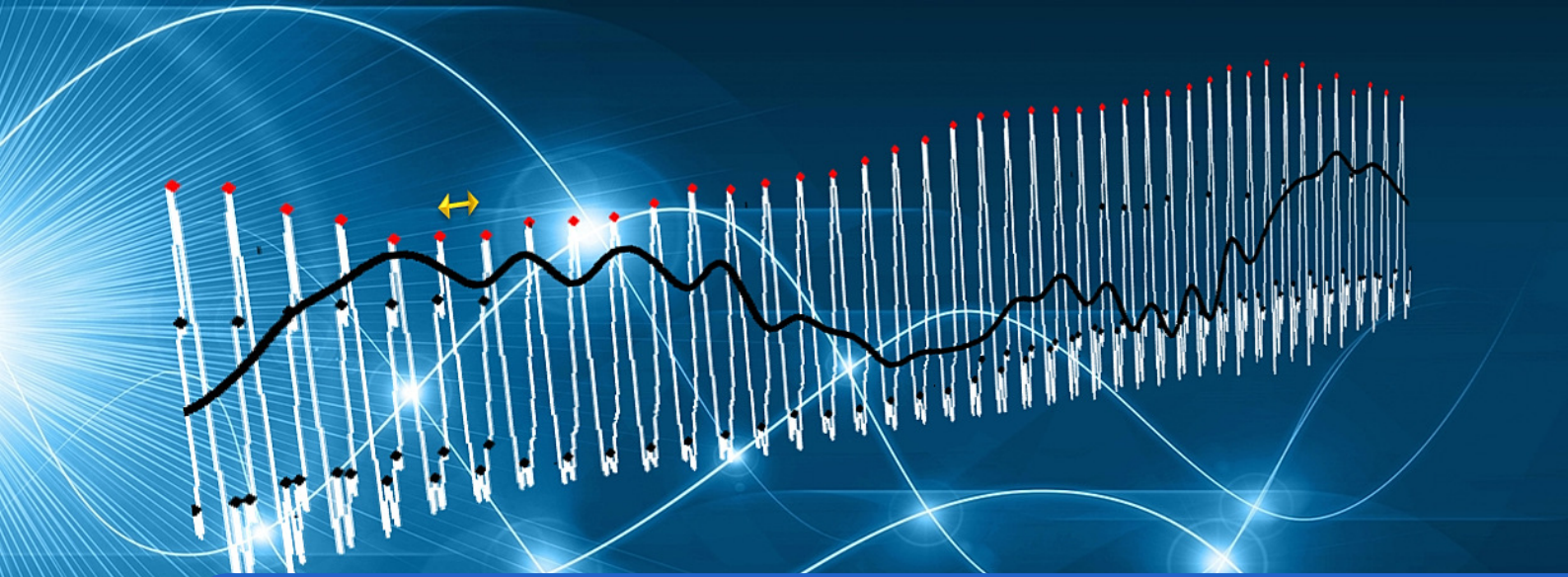
Figure 4.9 – Example : Third experiment

4.5 Conclusions

A *salience-based cycle length tracking* method has been proposed to track cycles in the temporal domain via a multi-scale analysis that assigns a salience to each signal peak and that is founded on dynamic programming. The method delivers a cycle peak sequence obtained by concatenating several peak triplets and giving rise to the smallest overall cycle length perturbation. The tracker does not rely on estimates of the typical cycle length, as opposed to existing proposals involving dynamic programming in the extraction of the cycle lengths. In the next chapter, the method will be applied to sustained speech sounds and the vocal cycle length tracking method will be validated in Chapter 9.

Key points

- The proposed *salience-based cycle length tracking* method (SCLT) is an example of *cycle-synchronous event analysis* method
- The tracking relies on a dynamic programming paradigm based on peak saliences and inter-peak durations
- No strong a priori assumptions are made with regard to the cycle length regularity
- The tracker delivers a cycle peak sequence giving rise to the smallest overall cycle length perturbation



5. Application of the SCLT method to sustained speech sounds

Objectives of this chapter

- Apply and illustrate the *saliency-based tracking* (SCLT) of the vocal cycle lengths in sustained voiced speech sounds
- Describe the post-processing used to obtain a constant-step interpolated cycle length time series

Contents

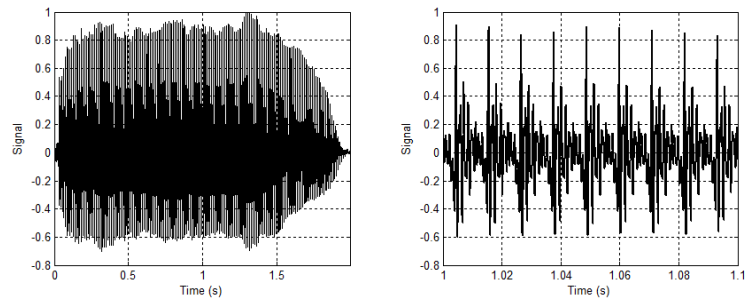
5.1	Introduction	79
5.2	Preprocessing	80
5.3	Speech sample salience analysis	82
5.4	Cycle length tracking	84
5.5	The vocal cycle length time series	86
5.6	Conclusions	88

5.1 Introduction

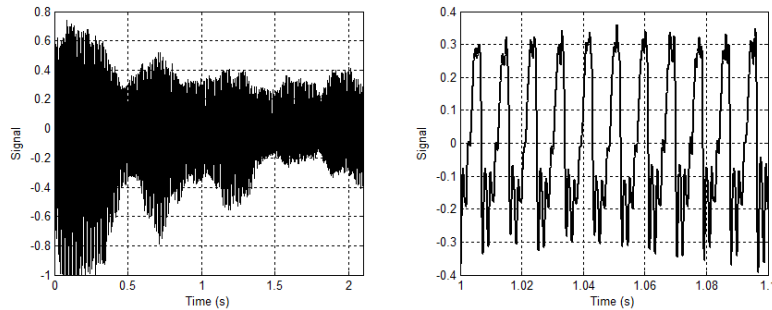
Here, the salience analysis as well as the tracking via dynamic programming are applied to voiced speech sounds to determine the glottal cycle lengths. The only assumption is that vocal frequency F_0 is comprised between $F_{0,min} = 60Hz$ and $F_{0,max} = 400Hz$, which covers the typical speaking range. Indeed, the typical F_0 range is between $50Hz$ and $250Hz$ for adult man, while for adult woman the range is between $120Hz$ and $500Hz$.

These methods are illustrated on the basis of 3 sustained vowels [a] produced by :

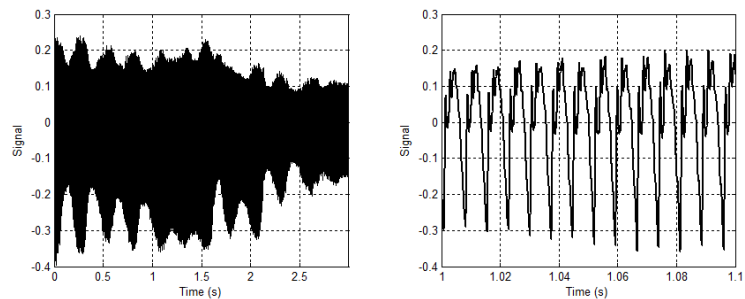
1. a normophonic male speaker with $F_0 \approx 88Hz$ (Figure 5.1a)
2. a male Parkinson speaker with $F_0 \approx 123Hz$ (Figure 5.1b)
3. a female speaker with $F_0 \approx 132Hz$ affected by essential tremor (Figure 5.1c)



(a) Modal voice



(b) Pathological voice (speaker with Parkinson's disease)



(c) Pathological voice (speaker with essential tremor)

Figure 5.1 – Fragment of vowel [a] produced by 3 speakers

5.2 Preprocessing

The sustained speech sound is preprocessed before applying sample salience analysis. For that, the speech signal is decimated to $F_s = 8000\text{Hz}$ and then band-pass filtered by means of a window-based finite impulse response (FIR) filter. Figure 5.2 illustrates the filter characteristics in the temporal and frequency domains.

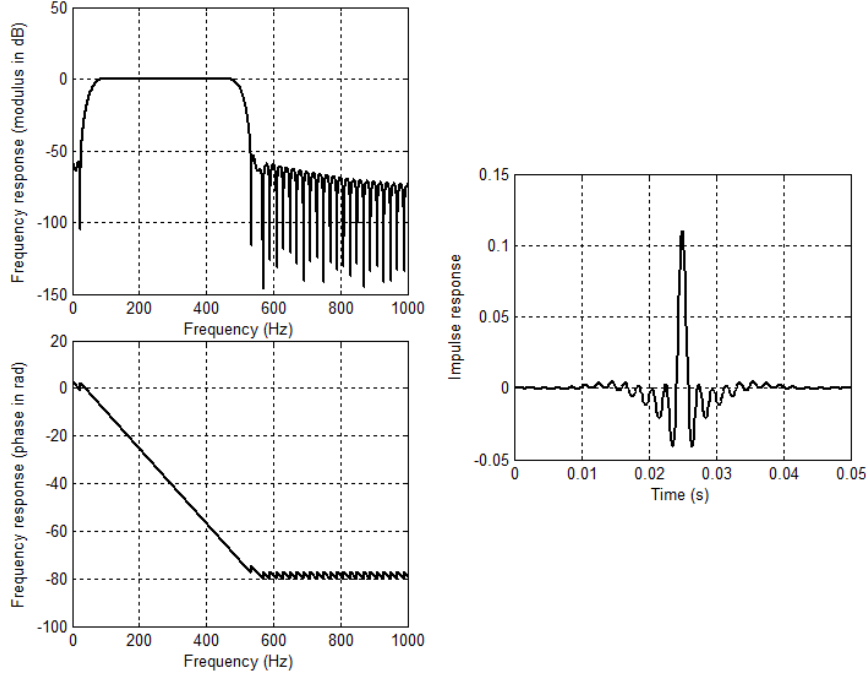
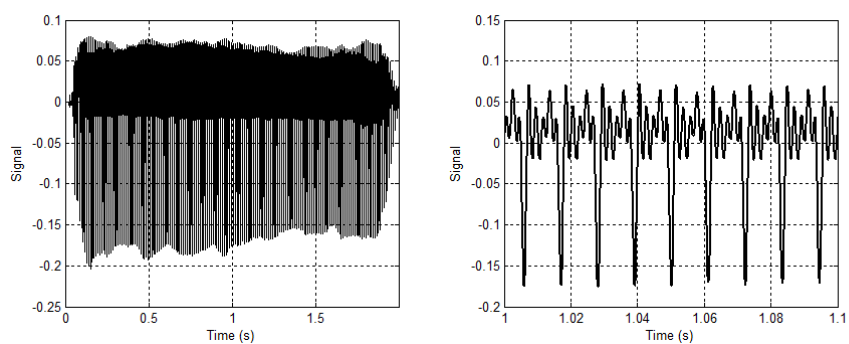


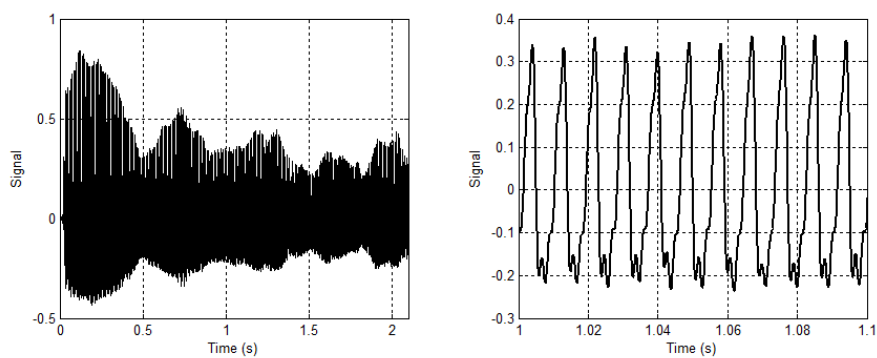
Figure 5.2 – FIR filter characteristics : frequency response (in modulus and phase) and impulse response [filter order $2N = 400$, $F_s = 8\text{kHz}$]

- The filter order has been chosen high enough to have a narrow transition band and reduce ripples in the pass-band. The filter order is equal to $2N$, with N the length (in samples) of the sliding analysis window used for salience analysis. Moreover, during salience analysis, it is recommended to discard the $N - 1$ first and last samples. The transitory filter response is therefore also discarded.
- The filter is non-recursive. Therefore, it is always stable. An important spectral characteristics of a symmetric FIR filter is its linear phase. As a consequence, a frequency-independent delay is introduced between the input and output. For a symmetric FIR filter, that delay (in samples) is equal to half of the filter order. That property may also be easily observed in the filter impulse response for which the prominent output impulse appears in the vicinity of N/F_s s.
- The cut-off frequencies of the filter have been chosen equal to 60Hz and 500Hz to remove additive low-frequency hum, high-frequency additive noise owing to turbulence as well as high-frequency formants.

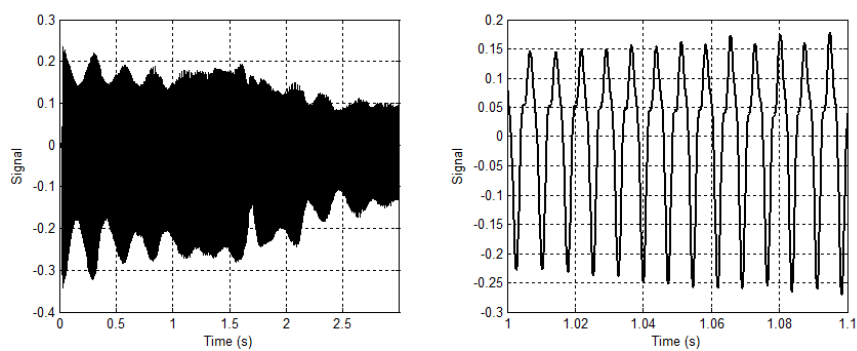
Figure 5.3 illustrates the band-pass filtered signals for the same 3 speakers.



(a) Modal voice



(b) Pathological voice (speaker with Parkinson's disease)



(c) Pathological voice (speaker with essential tremor)

Figure 5.3 – Band-pass filtered speech signals

5.3 Speech sample salience analysis

The salience s of a signal sample (which may be a signal peak or not) is defined as the length of the temporal interval over which the signal sample is a maximum. In voiced speech fragments, speech cycles are often characterized by a prominent signal peak that is the effect of the glottal excitation. The salience of that peak is expected to be high irrespective of the evolving signal amplitude.

The length of the sliding analysis window used for the salience allocation is 50% larger than the longest expected vocal cycle length, i.e. :

$$\frac{1.5}{F_{0,min}} = 25ms \quad (5.1)$$

Expressed in number of samples, the window length is :

$$N = 1.5 \frac{F_s}{F_{0,min}} = 200 \text{ for } F_s = 8kHz \quad (5.2)$$

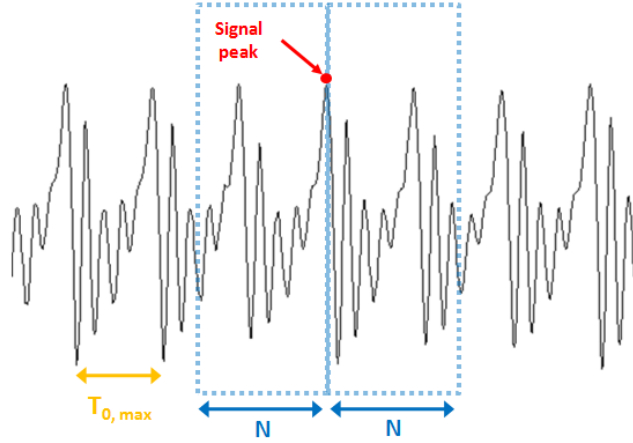
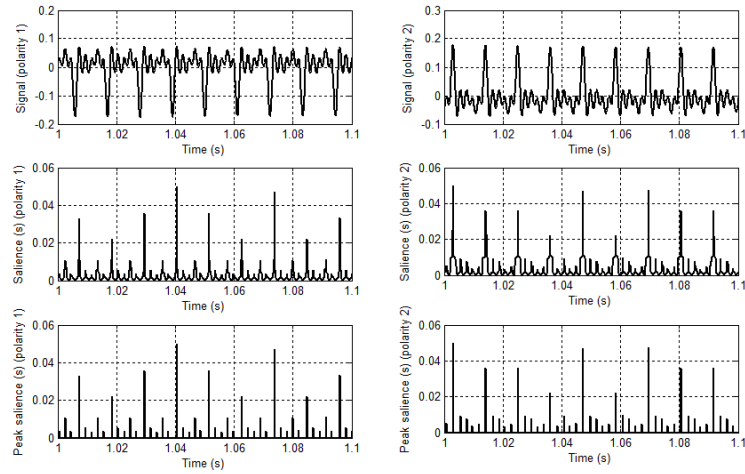


Figure 5.4 – Length N of the sliding analysis window used for salience allocation

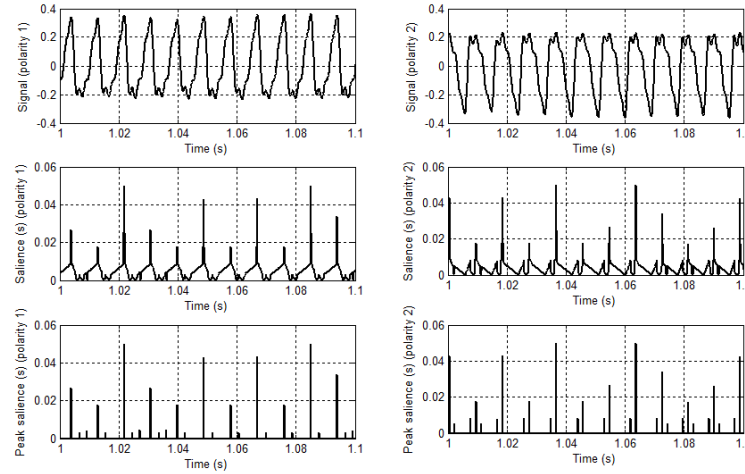
As a consequence, the sliding analysis window is guaranteed to contain at least one glottal cycle, so that the obtained final saliences (sum of left and right saliences) will be valid within a temporal interval containing at least 2 successive glottal cycles. Indeed, Figure 5.4 illustrates a voiced speech sound sustained by a speaker whose vocal frequency equals $F_{0,min}$. One observes that the salience of the central peak (marked in red) is valid within the temporal interval of length $2N - 1$ around that peak. Moreover, that interval comprises a minimum of two cycles.

The signal peaks are then ranked according to decreasing salience. Only the peaks the salience values of which are greater than 50% of the length of the shortest possible cycle (i.e. $\geq 1.25ms$) are kept. Even so, the number of remaining peaks is in excess of the number of expected cycles because a typical salience value of a speech cycle peak is equal to twice the cycle length. With a view to detecting the optimal cycle length time series, dynamic programming is used, as detailed in the previous chapter.

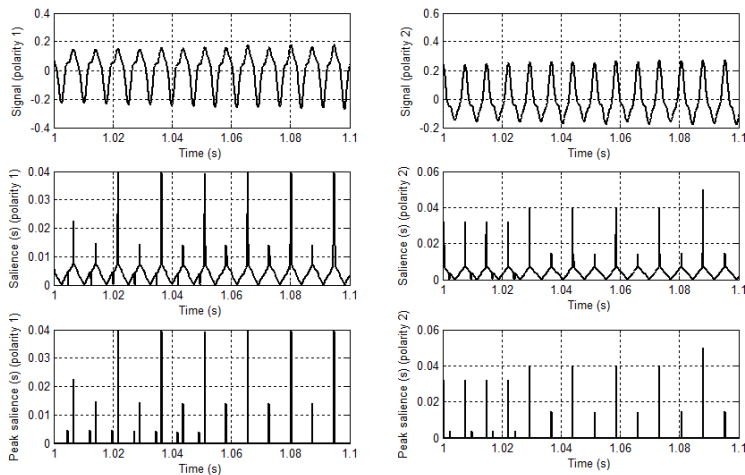
Figure 5.5 illustrates the sample saliences obtained for the 2 signal polarities. As expected, one observes that the salience (sum of left and right saliences) assigned to peaks (or valleys) are higher than the saliences of their close neighbours.



(a) Modal voice



(b) Pathological voice (speaker with Parkinson's disease)



(c) Pathological voice (speaker with essential tremor)

Figure 5.5 – Saliences assigned to speech signal samples or peaks. The salience analysis has been applied to the two signal polarities.

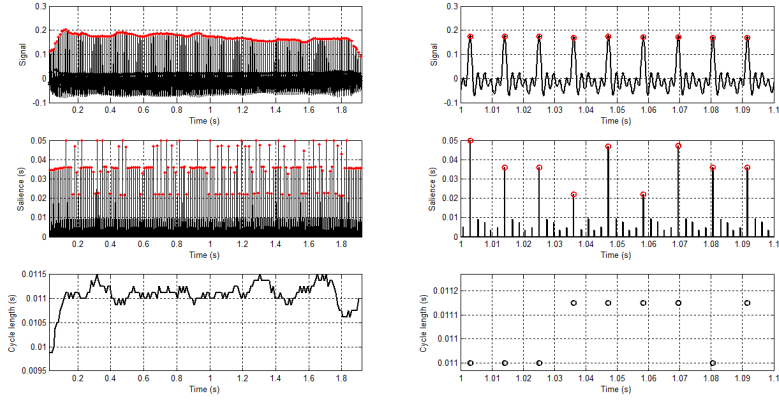
5.4 Cycle length tracking

The salience analysis and cycle length tracking via dynamic programming are carried out once for each polarity of the signal. The tracking parameters are :

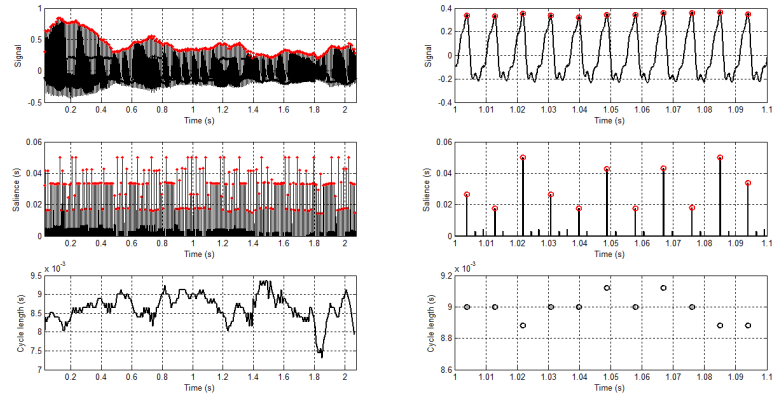
Maximal local length perturbation	:	$\alpha = 35\%$
Second-order perturbation weight	:	$\gamma_1 = 1$
Salience weight	:	$\gamma_2 = 2$
Admissible preceding triplet selection	:	$\gamma_3 = 60\%$
Triplet sequence length weight	:	$\gamma_4 = 3$

The polarity giving the smallest ratio C/L^{γ_4} (where C and L designate respectively the overall length perturbation and the number of cycles in the sequence) is retained.

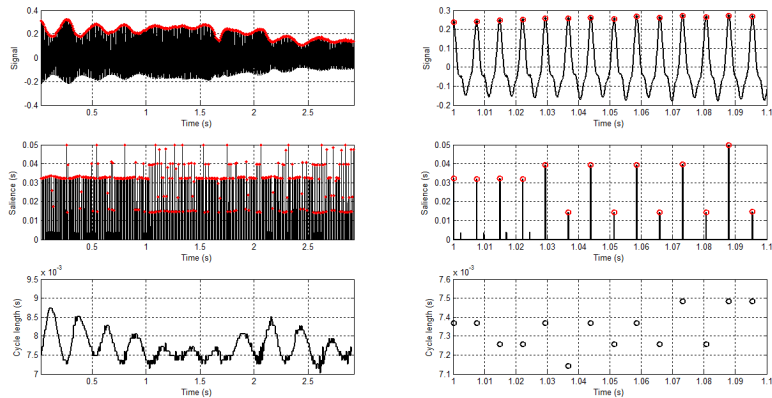
Figure 5.6 illustrates the tracking results for the same 3 speakers. For each speaker, from top to bottom, the pictures illustrate the band-pass filtered speech signal (with its chosen polarity), the saliences assigned to signal peaks and the first-order difference of selected speech cycle peak positions (red circles). One observes that the peaks with high saliences are selected. One also observes that the obtained cycle length time series reports one value per cycle and is not constant-step sampled.



(a) Modal voice



(b) Pathological voice (speaker with Parkinson's disease)



(c) Pathological voice (speaker with essential tremor)

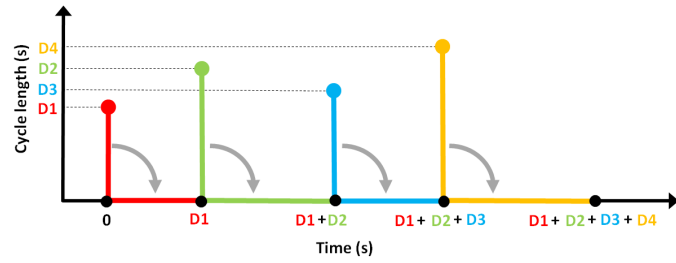
Figure 5.6 – Results of the vocal cycle length tracking : speech signal (top), peak saliences (middle), and first-order difference of selected peak positions (bottom). Left column : entire signal; right column : zoom in the middle part.

5.5 The vocal cycle length time series

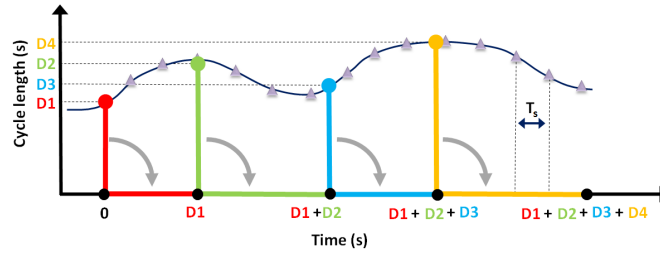
The raw cycle length time series is post-processed to obtain a constant-step sampled and more accurate cycle length time series.

The preprocessed speech signal is upsampled to a sampling frequency $\approx 200kHz$ to enable the peak positions to be measured with a higher precision requested by the size of vocal jitter, which in modal voices is expected to be $< 1\%$ of the typical cycle length. The peak positions are so refined and the increased precision cycle length time series is obtained via the first-order difference of the peak positions.

The raw vocal cycle length time series is then constant-step resampled. For that, the temporal axis is reconstructed by summing the successive vocal cycle lengths. The obtained series is interpolated by means of cubic splines and resampled to obtain a time series of lengths sampled at a constant sampling frequency equal to $F_s = 8kHz$. Figure 5.7 illustrates these steps.



(a) Reconstruction of the temporal axis by summing successive vocal cycle lengths

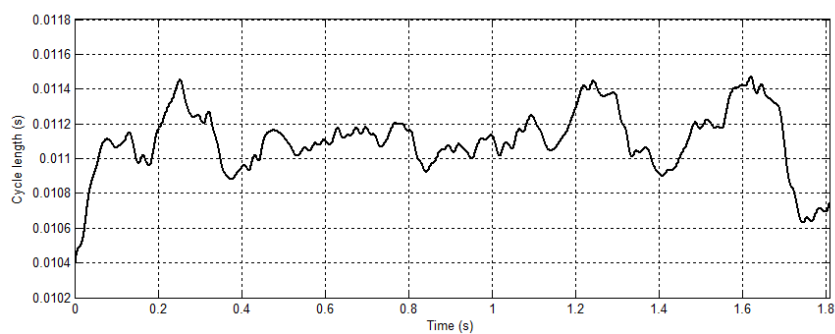


(b) Cubic spline interpolation of cycle lengths and resampling

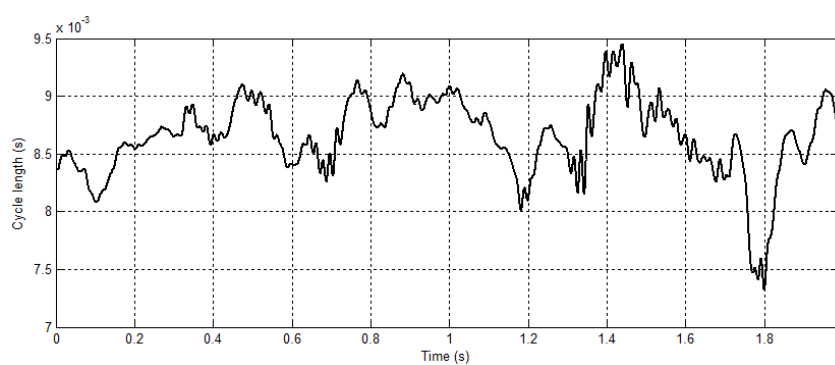
Figure 5.7 – Constant-step resampling

Notice that the frequency content of the cycle length time series is expected to be comprised between 0 and $F_0/2$. The choice of a high sampling frequency, compared to F_0 , is motivated by the fact that this choice guarantees a high temporal precision of sample positions requested by the perturbation analysis techniques explained later.

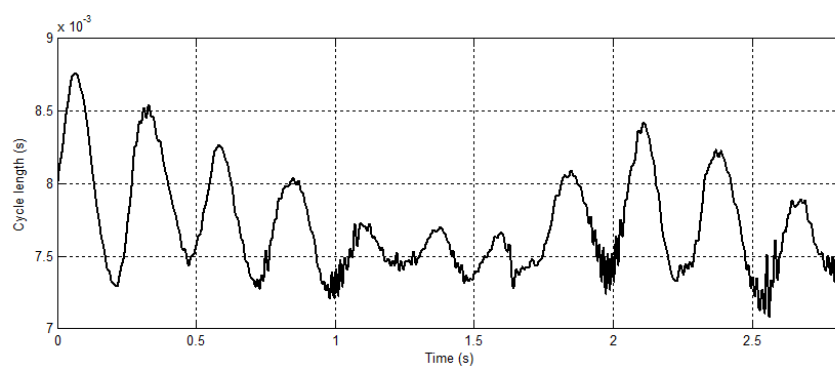
Figure 5.8 illustrates the constant-step cycle length time series obtained for 3 speakers.



(a) Modal voice



(b) Pathological voice (speaker with Parkinson's disease)



(c) Pathological voice (speaker with essential tremor)


Figure 5.8 – Vocal cycle length time series obtained for 3 speakers

5.6 Conclusions

In this chapter, the parameters that are used for the salience-based cycle length tracking in the framework of the application to sustained vowels have been specified. A time series giving the evolving cycle lengths is obtained. This time series has to be analyzed and, especially, its slow and fast perturbations. The cycle length perturbation analysis will be described in the next chapters.

Key points

- The SCLT method has been applied to sustained voiced speech sounds
- The tracking has been carried out for each polarity of the speech signal, and the raw cycle length time series giving rise to the minimal overall length perturbation is retained.
- This retained raw cycle length time series is interpolated and re-sampled to obtain a time series of lengths sampled at a constant sampling step



6. The wonderful story of the rolling wheel in a signal processing context

Objectives of this chapter

- Explain key techniques used to analyze time series
- Introduce the concept of instantaneous frequency and amplitude

Contents

6.1	Introduction	93
6.1.1	Geometry	93
6.1.2	Wheel in sustained motion	93
6.1.3	Wheel mechanism in sustained motion	94
6.2	Fourier analysis	95
6.2.1	The Fourier series and Fourier transform	95
6.2.2	Finite length time series and frequency resolution	96
6.2.3	Global analysis	98
6.3	Time-frequency analysis	100
6.4	Instantaneous frequency and amplitude	102
6.4.1	Instantaneous phase, amplitude and envelope	102
6.4.2	Instantaneous frequency	105
6.4.3	Relevance of the instantaneous values	106
6.4.4	Analysis of multi-component time series	107

6.1 Introduction

Once upon a time, in the middle of the 4th millennium BCE, a circular component, considered today as one of the six simplest mechanical devices which provide a mechanical advantage, was created : the wheel. Simple as it seems, it is the very basis of motion in various domains.

Also, in the context of signal processing, a good understanding of wheel motion and of its mathematical modelling is a key asset with a view to addressing traditional signal analysis techniques and understanding their advantages and limitations.

6.1.1 Geometry

A wheel is a circular device that rotates on an axis z_{rot} . That device is characterized geometrically by its radius, r , which is the constant length of a line segment that joins the center of a circle with any point on its circumference. In cartesian coordinates, that shape, illustrated in Figure 6.1, is generally defined on the basis of two orthogonal axes, x and y .

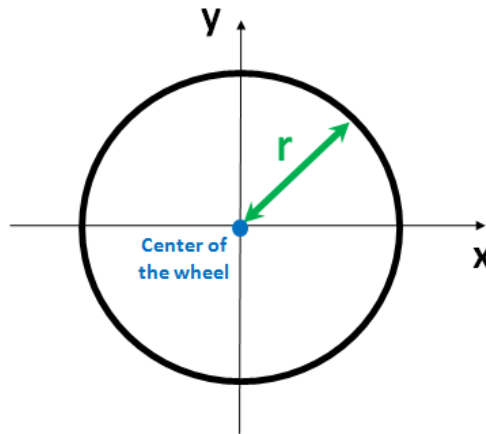


Figure 6.1 – Geometry of wheel

Assuming that the origin of these axes is located at the center of the circle, the circle equation is given by :

$$\sqrt{x^2 + y^2} = r \quad (6.1)$$

6.1.2 Wheel in sustained motion

One considers that a wheel rotates around its fixed center axis, z_{rot} . At first, one also assumes that the movement is sustained and the angular speed ω_0 is constant.

An experiment, illustrated in Figure 6.2, consists in considering a point $P(x_p, y_p)$ which is fixed to the circle circumference. The position of this point in the x, y plane is determined at each instant t by its coordinates $x_p(t)$ and $y_p(t)$ which correspond to the orthogonal projection of the point on respectively the axes x and y . In this example, the angular velocity $\omega_0 = 2\pi \text{ rad/s}$ and the radius $r = 1$.

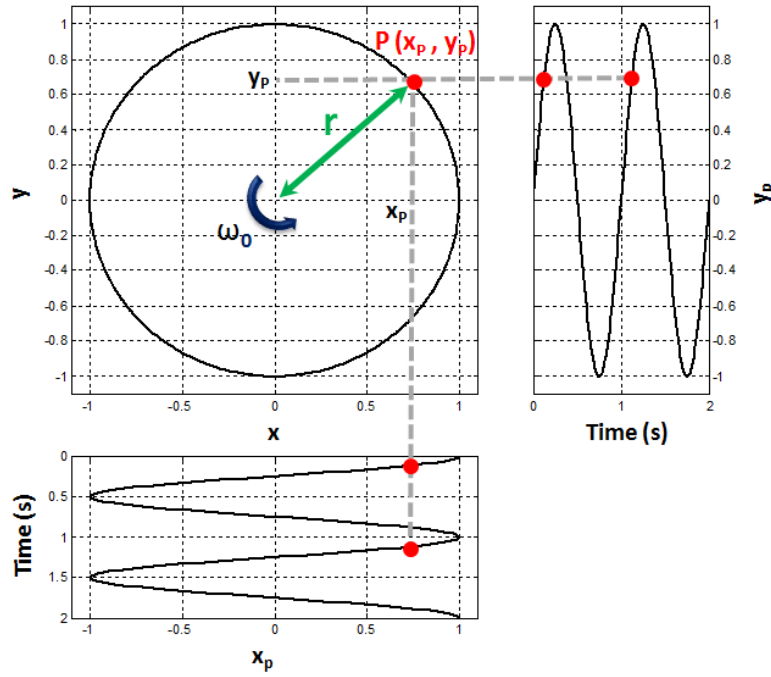


Figure 6.2 – Wheel in motion and temporal evolution of a point fixed to its circumference ($r = 1$ and $\omega_0 = 2\pi \text{ rad/s}$)

A visual inspection of the temporal evolution of these two coordinates shows that the trajectory of that point can be expressed on the basis of 2 functions $x_p(t)$ and $y_p(t)$ that have to satisfy simultaneously the circle equation (6.1). These functions are :

$$\begin{cases} x_p(t) = r \cos(\phi(t)) \\ y_p(t) = r \sin(\phi(t)) \\ \phi(t) = \omega_0 t + \phi_0 \end{cases} \quad (6.2)$$

The quantity $\phi(t) = \omega_0 t + \phi_0$ is called the phase function. Assuming that initially the point is located on the abscissa ($x_p(0) = r, y_p(0) = 0$), its initial value, ϕ_0 , is equal to zero.

The trajectory $z(t)$ may also be described via a complex exponential, as follows :

$$\begin{aligned} z(t) = z_{real}(t) + j z_{imag}(t) &= r e^{j\phi(t)} = r e^{j(\omega_0 t + \phi_0)} \\ &= r \cos(\phi(t)) + j r \sin(\phi(t)) \end{aligned} \quad (6.3)$$

where $j = \sqrt{-1}$. In this equation, the real $z_{real}(t)$ and imaginary $z_{imag}(t)$ parts are directly related to the time series $x_p(t)$ and $y_p(t)$.

The modulus $|z| = \sqrt{z_{real}^2 + z_{imag}^2} = r$ is the circle radius and the argument $\arg(z(t)) = \arctan\left(\frac{z_{imag}(t)}{z_{real}(t)}\right) = \phi(t)$ is the phase function.

6.1.3 Wheel mechanism in sustained motion

The previous considerations with regard to the motion of one wheel can be generalized to a superposition of wheels defined as a mechanism where several wheels are concatenated so that the center of rotation of the next wheel is situated on the circumference of the preceding wheel. Each wheel has its own radius and angular velocity.

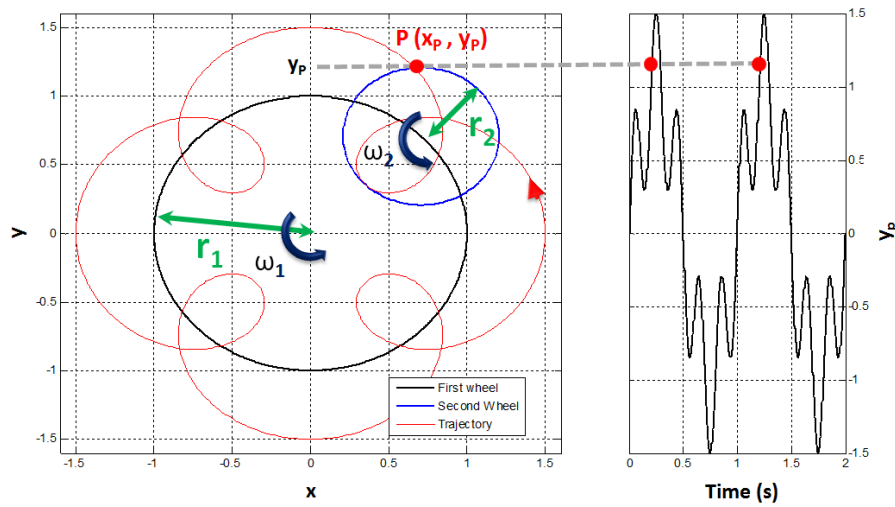


Figure 6.3 – Chain of two wheels in motion. Trajectory and temporal evolution of a point fixed to the circumference of the second wheel ($r_1 = 1$, $r_2 = 0.5$, $\omega_1 = 2\pi \text{ rad/s}$, $\omega_2 = 10\pi \text{ rad/s}$)

As an example, Figure 6.3 illustrates a concatenation of two wheels. The wheel radii are respectively $r_1 = 1$ and $r_2 = 0.5$ and the angular velocities are respectively $\omega_1 = 2\pi \text{ rad/s}$ and $\omega_2 = 10\pi \text{ rad/s}$. The trajectory in the $x - y$ plane of a point attached to the second wheel as well as the temporal evolution of its ordinate $y_p(t)$ are represented.

Assuming that the motion is sustained and the radii and the velocities are constant, time series $y_p(t)$ is periodic, provided that ω_2/ω_1 is a rational number. This time series can be expressed as the sum of two functions which describe the behaviour of individual wheels where $\phi_{0,1}$ and $\phi_{0,2}$ are the initial phases :

$$\begin{aligned} y_p(t) &= y_{p,1}(t) + y_{p,2}(t) \\ &= r_1 \sin(\omega_1 t + \phi_{0,1}) + r_2 \sin(\omega_2 t + \phi_{0,2}) \end{aligned} \quad (6.4)$$

Therefore, a periodic time series can be analyzed by means of a set of oscillating basis functions the radius and frequency of which are time-independent. That traditional analysis technique is known as Fourier analysis. The lesson to be learned here is that a Fourier series has an interpretation in terms of a device formed of a concatenation of wheels rotating on the circumference of another wheel (called epicycles). The wheel that has a fixed axle is called the deferent.

6.2 Fourier analysis

In time-frequency analysis, traditional analysis methods are mostly based on the well-known Fourier analysis. Even though that method is frequently used in academia and industry, its field of applications is limited. The method has been initially proposed for the analysis of periodic signals, but several generalizations have been introduced. Its theoretical concepts are often considered to be founding principles of more sophisticated signal processing techniques.

6.2.1 The Fourier series and Fourier transform

The goal of the Fourier analysis consists initially in expressing a periodic signal $x(t)$ via a linear combination of a set of basis functions which have been chosen here to be complex exponentials, $e^{jn\frac{2\pi}{T_0}t}$, where T_0 is the oscillation period.

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\frac{2\pi}{T_0}t} \quad (6.5)$$

The Fourier series coefficients c_n may be computed via the following inverse relation.

$$c_n = \frac{1}{T_0} \int_0^{T_0} x(t) e^{-jn\frac{2\pi}{T_0}t} dt \quad (6.6)$$

Complex coefficients c_n are the amplitude and phase of each oscillating function. In other words, their modulus $|c_n|$ corresponds to the radius of a wheel that rotates at a time-invariant angular speed equal to $n\frac{2\pi}{T_0}$ and its argument $\arg(c_n)$ refers to the phase function of that wheel.

As an example, Figure 6.4 illustrates the Fourier coefficients of the signal which results from the sustained motion of a deferent and an epicycle, presented in Figure 6.3. The upper illustration is the temporal evolution of the signal and the lower is the modulus of the Fourier coefficients. One observes that the signal is characterized by two components which oscillate at $1Hz$ and $5Hz$ and the amplitudes of which are respectively 1 and 0.5, as expected.

These relations may be adapted to the analysis of non-periodic signals. In that case, the frequency content of the signal $x(t)$ is investigated for a set of frequencies f comprised in the interval $[-\infty, +\infty]$, referring to the so-called Fourier transform : $x(t) \xleftrightarrow{F} X(f)$ where,

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df \quad \text{and} \quad X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \quad (6.7)$$

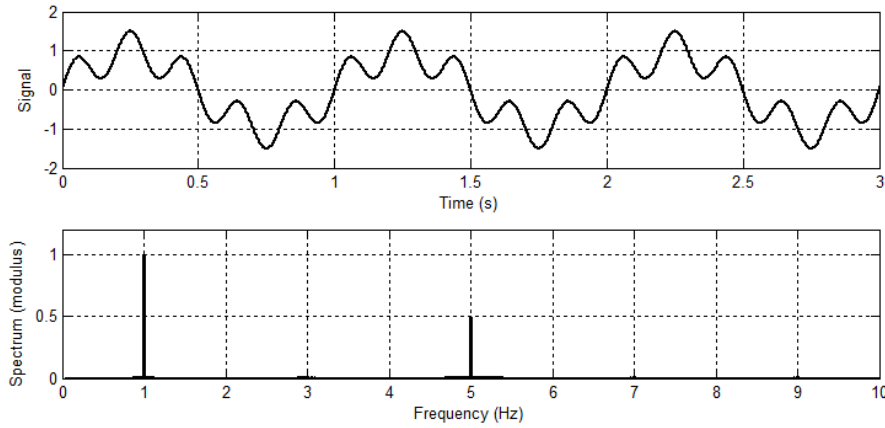


Figure 6.4 – Periodic time series and its Fourier coefficients

6.2.2 Finite length time series and frequency resolution

In practice, a signal may be recorded and analyzed over a finite temporal interval only. In that situation, the observed signal $x_o(t)$ may be assimilated to a product of the original signal of infinite duration $x(t)$ and an observation window of finite length $w(t)$.

Several windows have been defined. Figure 6.5 illustrates rectangular, triangular and Hamming windows in the temporal and frequency domains. In the temporal domain, these windows manage

boundary effects differently. The spectrum of these windows has the same general shape. It contains a prominent main lobe, centered around 0Hz , and additional secondary lobes the amplitudes of which decrease with frequency.

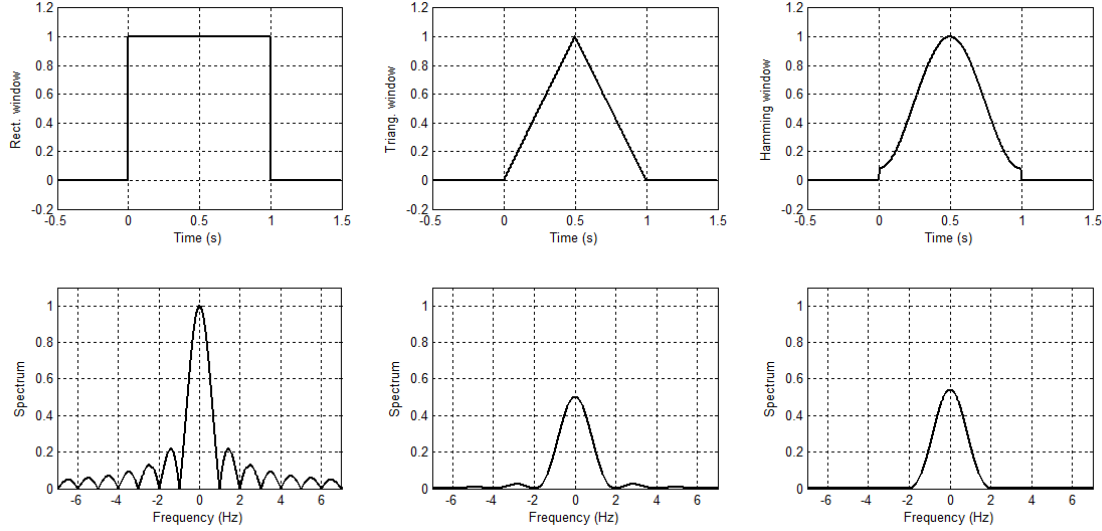


Figure 6.5 – Observation window in temporal and frequency domains

Windows are generally characterized by the width of the main lobe and the attenuation of the secondary lobes. An important property is that the main lobe width is inversely proportional to the length of the window. Window properties directly affect the spectrum of the finite length signal $x_o(t)$. The spectrum $X_o(f)$ of $x_o(t)$ is obtained via the convolution of the spectrum $X(f)$ of $x(t)$ with the spectrum $W(f)$ of the observation window. As a consequence, the window spectrum is replicated and centred on each frequency component of $x(t)$. The diffuse shape of the spectrum makes it therefore more difficult to detect the frequency components of a time series.

$$\begin{aligned} x(t) &\xleftrightarrow{F} X(f) \\ w(t) &\xleftrightarrow{F} W(f) \\ x_o(t) = x(t) \cdot w(t) &\xleftrightarrow{F} X_o(f) = X(f) \otimes W(f) \end{aligned} \quad (6.8)$$

$$X(f) \otimes W(f) = \int_{-\infty}^{\infty} X(\alpha) W(f - \alpha) d\alpha \quad (6.9)$$

A Hamming window is often used. It is computed easily and, due to its tapered shape, has reduced side-lobes at the price of a doubling of the main lobe width compared to the rectangular window. Due to its shape, an amplitude correction factor of 1.855 is applied to the amplitude of the oscillating functions.

Figure 6.6 illustrates three spectra computed for three different observation durations of the periodic time series $x(t)$ displayed in Figure 6.4.

One observes the most prominent frequency components at 1Hz and 5Hz , but other frequency components appear which are the consequence of windowing. One also observes that the frequency bandwidth of the window spectrum is inversely proportional to the duration of the observation

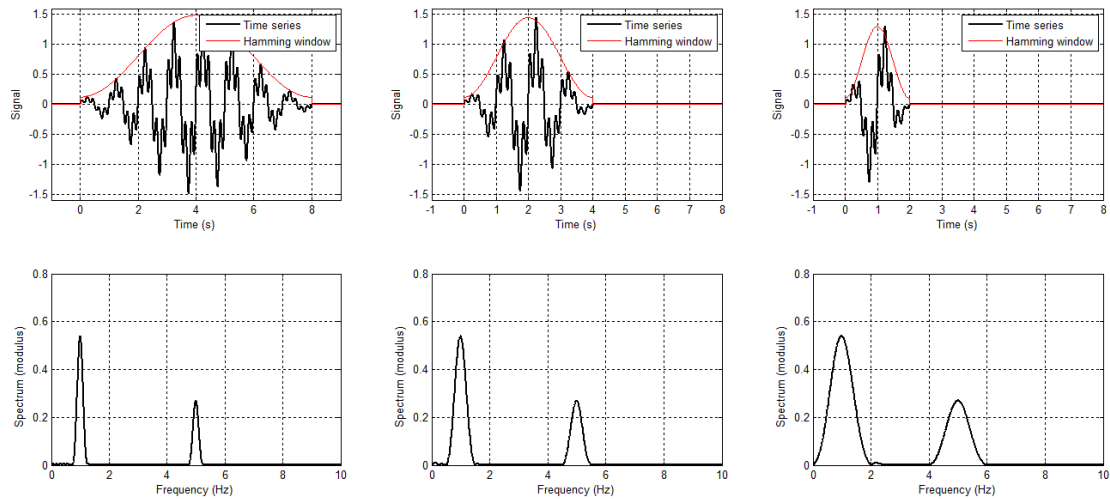


Figure 6.6 – Influence of the window duration on the Fourier spectrum

window. Therefore, if the observation duration keeps decreasing, the two components at 1Hz and 5Hz will become indistinguishable. A consequence is that two frequency components that are close are detected easily only if the frequency spacing is larger than the width of the window main lobe. In other words, the duration of observation has to be chosen as long as required with regard to the targeted analysis purpose.

6.2.3 Global analysis

The Fourier transform is global giving time-invariant amplitude and frequency values for the whole time span covering the range of integration (equation (6.7)). Until now, we have considered that the wheels move with constant angular velocity and wheel radius. What happens now if these conditions are not satisfied ? Is Fourier analysis still relevant when analyzing data that are produced via a non-stationary process ? This case is illustrated below.

Example 1 :

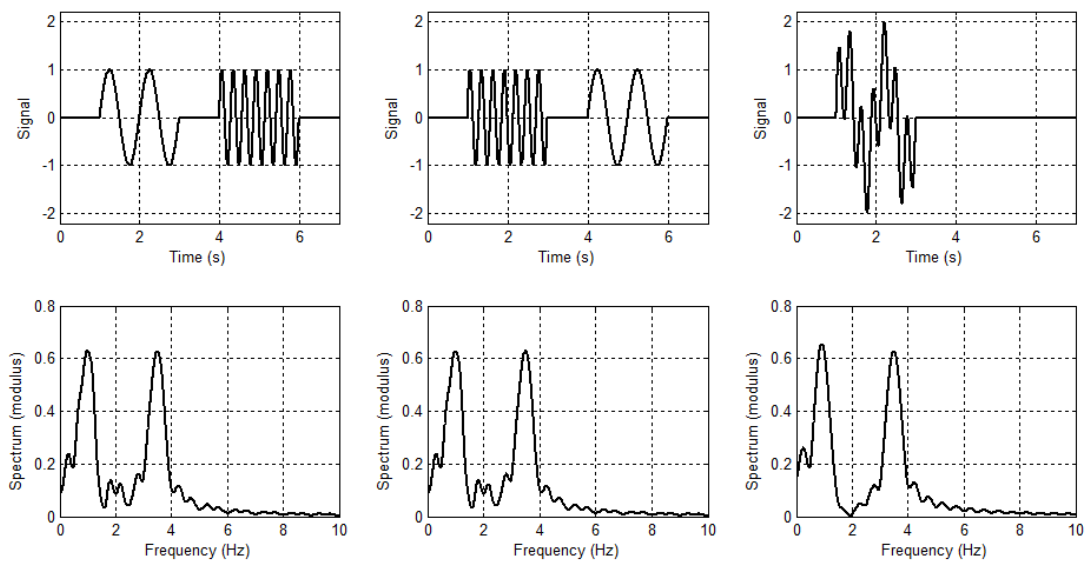
The case for which the angular velocity is time-dependent and the radius constant. Figure 6.7a illustrates three configurations : the two first situations correspond to the rotation of one wheel at two angular velocities $\omega_1 = 2\pi \text{ rad/s}$ and $\omega_2 = 7\pi \text{ rad/s}$. The third configuration is related to the motion of a superposition of two wheels with the same radius (this case is in fact stationary before windowing). In each case, the spectrum has been computed.

One observes that the three spectra give the same information. Therefore, obtaining the amplitude spectra is not sufficient to detect the instants of velocity modification.

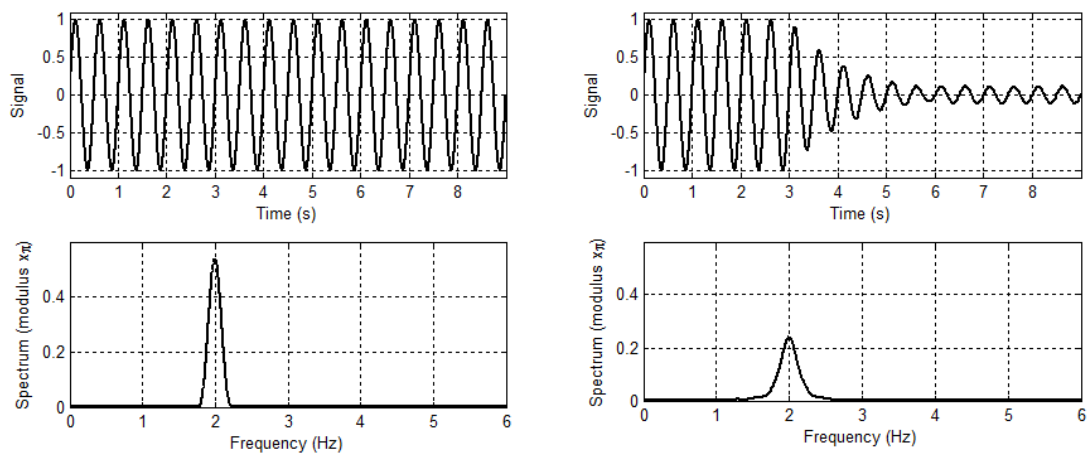
Example 2 :

One considers the case for which the radius of a wheel is decreasing according to an exponential law, but the angular velocity is maintained constant. The signal obtained by tracking the trajectory on the x axis of a point on the wheel is illustrated in Figure 6.7b. This signal has a time-varying amplitude.

The spectra show that amplitude variations may not be detected on the base of the frequency components. The right-side spectrum is indeed roughly equivalent to the spectrum of a short windowed constant-amplitude periodic time series with a frequency of 2Hz .



(a) Three quasi-identical spectra obtained on the basis of signals with (left and middle) or without (right) time-dependent frequency



(b) Time-invariant and time-variant amplitude

Figure 6.7 – Fourier analysis applied to data obtained via stationary or non-stationary processes

6.3 Time-frequency analysis

The previous considerations suggest that Fourier analysis is not the best candidate to analyse data that are samples non-stationary processes.

As shown below, the short-term Fourier analysis (STFT) is a sliding window-based Fourier analysis the goal of which is to track temporal or frequency changes, based on the assumption of local signal stationarity. The window is placed at the beginning of the signal and moves progressively to the right. For each position of the analysis window, the amplitude spectrum is computed. The concatenation of the successive spectra offers a time-frequency representation of the time series, called spectrogram.

The major limitation of STFT is related to a compromise, known as the time-frequency uncertainty principle, which stipulates that a good temporal resolution and a good frequency resolution may not be obtained simultaneously.

As an example, Figures 6.8b and 6.8c illustrate two spectrograms obtained for a signal the spectral characteristics of which evolve. The time series (Figure 6.8a) comprises several piecewise harmonic signals and an impulse. The spectrograms have been obtained on the basis of two different analysis window lengths T_w .

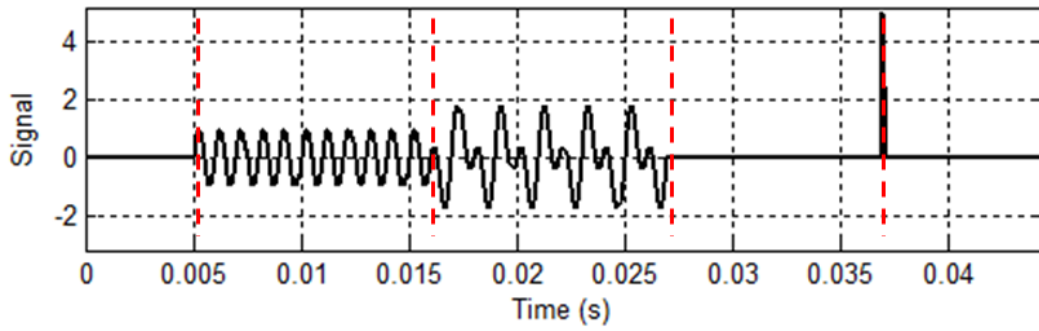
One observes that the results are affected by temporal and frequency resolutions. A long observation window decreases the temporal resolution so that temporal events cannot be precisely localized (Figure 6.8b). Inversely, transient events may be accurately tracked on the basis of short observation windows, but then the frequency resolution is feeble (Figure 6.8c).

Another way of analyzing time-varying data consists in using the continuous wavelet transform (CWT) and its time-frequency representation, called scalogram. The wavelet decomposition is a multi-scale analysis method, initially introduced with a view to overcome the temporal and frequency resolution problems of the Fourier analysis. That decomposition method is based on an a priori selection of basis functions that are called wavelets. The main difference is that wavelets are localized in both time and frequency whereas harmonics are localized in frequency only. In CWT, the analysis of a signal is carried out by the use of a special function, $h(t)$, called the mother wavelet. The wavelet transform is defined as follows :

$$CWT(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) h^*\left(\frac{t - \tau}{a}\right) dt \quad (6.10)$$

The continuous wavelet transform projects the signal on shifted (delay τ) and compressed or stretched (scale a) versions of a mother wavelet. By projecting the signal on wavelets at various scales and positions, a function of two variables is obtained. Wavelet analysis enables the use of long time intervals where one wants more precise low-frequency information, and shorter intervals where one wants high-frequency information. In other terms, for small values of a , the wavelet is a shorter version of the mother function, which is responsive to higher frequencies. For very large values of a , the wavelet is expanded and is responsive to lower frequencies. Figure 6.8d illustrates the time-frequency representation obtained via CWT.

However, the a priori selection of the basis function creates in practice additional problems (boundary effects, selection of relevant wavelet coefficients, ...). Therefore, a necessary condition with a view to analyzing data that are produced by non-stationary process is to have an adaptive basis.



(a) Temporal evolution of a time series with time-varying spectral content

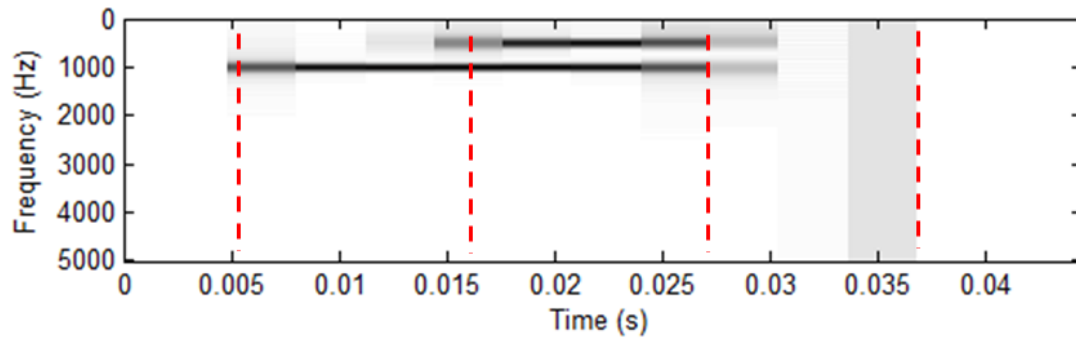
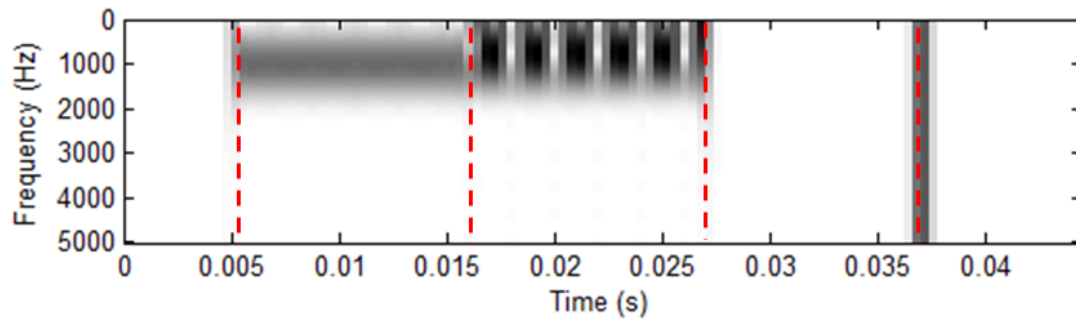
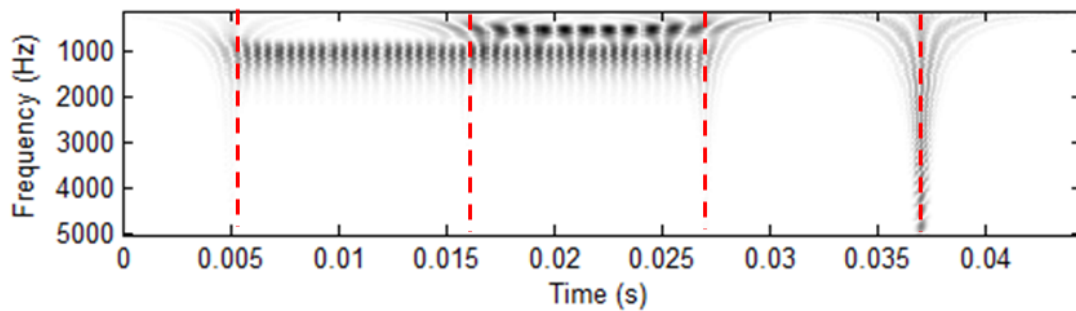
(b) Spectrogram ($T_w = 12.8ms$) : The frequency components at $500Hz$ and $1000Hz$ are clearly observed, but the instants of spectral changes are not clearly marked(c) Spectrogram ($T_w = 1.6ms$) : Modifications of the spectral content are tracked in the temporal domain, but the spectral components are not clearly marked(d) Scalogram (based on Morlet mother wavelet) : The frequency components at $500Hz$ and $1000Hz$ are clearly observed as well as the impulse instant

Figure 6.8 – Time-frequency representation of a time series with time-varying spectral content, obtained via a short-term Fourier transform and continuous wavelet transform

6.4 Instantaneous frequency and amplitude

We have shown that the Fourier analysis is meaningful only for the analysis of time series that have a spectral content that is locally constant in time. How can we analyze signals with time-varying amplitude and/or frequency ? As an example, consider the car of the Flintstones' family, illustrated in Figure 6.9. The front wheel may be damaged during travel. The car speed may depend, amongst others, on the traffic and driver health conditions. The analysis of signals with time-dependent features relies in a majority of cases on the tracking of the amplitude and frequency at each time instant t .

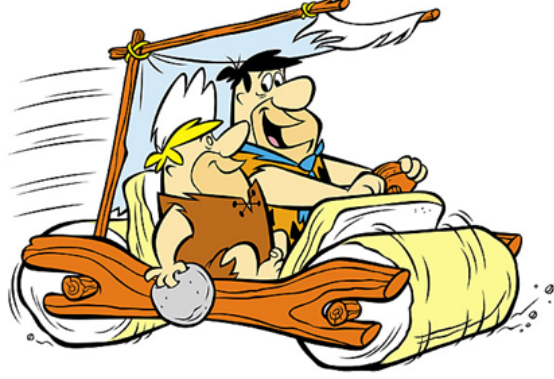


Figure 6.9 – *The Flintstones* (1960)

In this section, one introduces the concept of instantaneous frequency $f(t)$ and envelope $a(t)$ that offer a way to describe the frequency content of non-stationary signals. One considers that the frequency and envelope are functions of time and have an instantaneous value.

6.4.1 Instantaneous phase, amplitude and envelope

Figure 6.10a illustrates a perfectly circular wheel. If the wheel rotates at angular velocity ω_0 and its rotation center corresponds exactly to the axis origin, its simple harmonic movement trajectory is described by :

$$z(t) = a_0 e^{j\omega_0 t} \quad (6.11)$$

The term a_0 corresponds here to the constant radius and the phase function increases linearly with slope ω_0 .

The relation (6.11) may be generalized to describe the trajectory in the case of a time-variant wheel contour and/or angular velocity, as in Figure 6.10b.

$$z(t) = a(t) e^{j\phi(t)} = z_{real}(t) + j z_{imag}(t) \quad (6.12)$$

The instantaneous amplitude function $a(t)$ is defined as the length of a line segment that joins the origin of axes and a point of the trajectory at time t . The instantaneous phase function $\phi(t)$ is obtained by measuring the angle between the real axis (abscissa) and that line segment. The wheel movement may also be characterized by time series z_{real} and/or z_{imag} by projecting the trajectory on the real and/or imaginary axes.

As shown below, if the temporal fluctuations of the instantaneous amplitude are slower than the instantaneous changes of the carrier, the instantaneous amplitude $a(t)$ may be interpreted as the envelope of the time series.

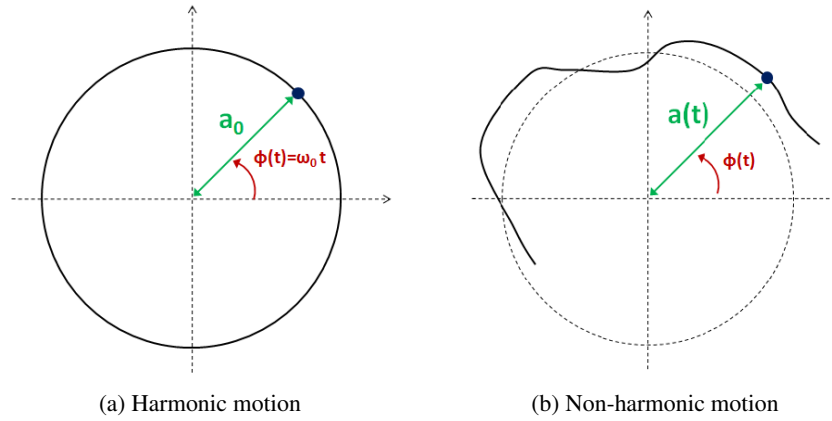


Figure 6.10 – Wheel in harmonic or non-harmonic motion

6.4.1.1 Simulations

Figure 6.11 illustrates the realisations of three different processes. Each process is characterized by a wheel which is put in rotation. The experiments consist here in tracking the trajectory on the real axis of a point which is fixed to the wheel contour. The angular velocity of the wheel and its geometry are time-dependent or time-independent. In addition, one assumes that the radius fluctuations are slower compared to the rotation velocity. The time series as well as the instantaneous values have been obtained by simulation.

- The first wheel is circular. The radius has been fixed to $r_1 = 1$ but its angular velocity $\omega_1(t)$ varies as follows :

$$\omega_1(t) = 2\pi(3 + 2 \sin(10\pi t)) \quad (6.13)$$

- The second wheel rotates at a constant angular velocity $\omega_2 = 6\pi \text{ rad/s}$ but the wheel is progressively damaged so that its radius $r_2(t)$ is time-varying.

$$r_2(t) = e^{-\frac{t^2}{4}} (1 + 0.2 \cos(\pi t)) \quad (6.14)$$

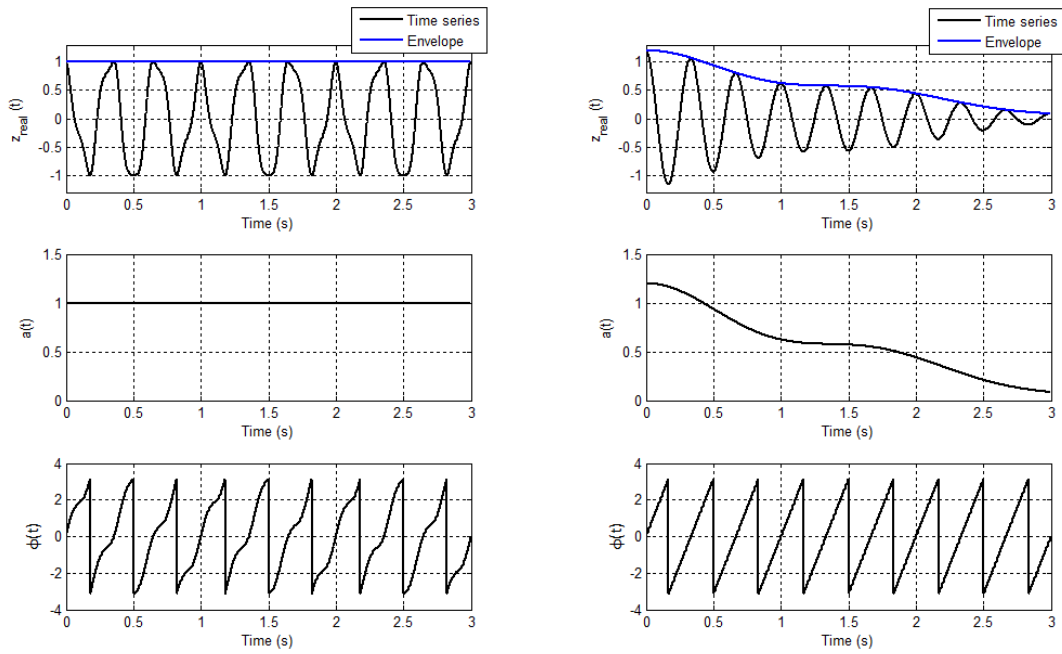
- And finally, the radius $r_3(t) = r_2(t)$ and the angular velocity $\omega_3(t) = \omega_1(t)$ of the third wheel vary both in time.

For each configuration, the trajectory on the real axis $z_{\text{real}}(t)$ as well as the instantaneous amplitude $a(t)$ and phase $\phi(t)$ time series are given.

In the first example, the instantaneous amplitude is a constant equal to the radius $|z| = r_1 = 1$. In the other examples, the instantaneous amplitude varies in time. The phase function $\phi(t)$ is a signal with values comprised between $-\pi$ and π .

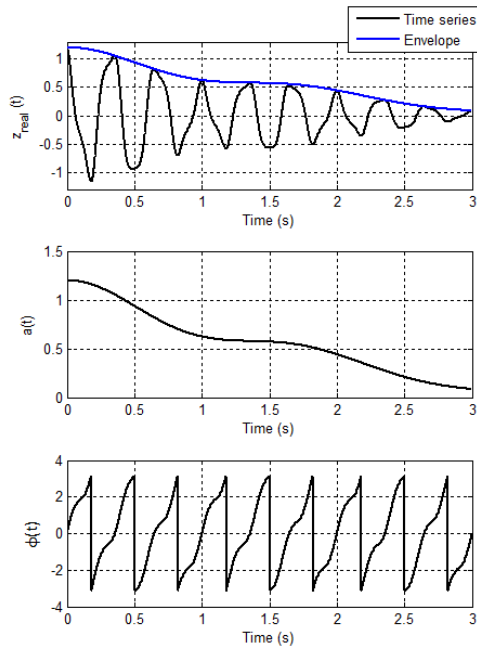
Assuming that the fluctuations of amplitude are slower compared to the oscillation frequency, the instantaneous amplitude may be interpreted as the envelope of the signal. Therefore, the time series may be expressed as the product of a slowly time-varying envelope $a(t)$ and a purely frequency modulated function with unity amplitude $\cos(\phi(t))$, called the carrier.

$$z_{\text{real}}(t) = a(t) \cos(\phi(t)) \quad (6.15)$$



(a) Time-dependent angular velocity $\omega_1(t)$ and constant radius r_1

(b) Time-dependent wheel radius $r_2(t)$ and constant angular velocity ω_2



(c) Time-dependent angular velocity $\omega_3(t)$ and wheel radius $r_3(t)$

Figure 6.11 – Trajectory on the real axis $z_{real}(t)$ as well as the instantaneous amplitude $a(t)$ and phase $\phi(t)$ time series obtained for three simulated wheels with time-dependent angular velocity and/or radius

6.4.2 Instantaneous frequency

The instantaneous frequency $f(t)$ may be obtained from the phase function. The phase function is unwrapped with a view to obtaining a monotonically increasing function, which is then differentiated :

$$f(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt} \quad (6.16)$$

As shown further, the unwrapped phase function is guaranteed to increase monotonically if a condition with regard to the time series local mean is satisfied.

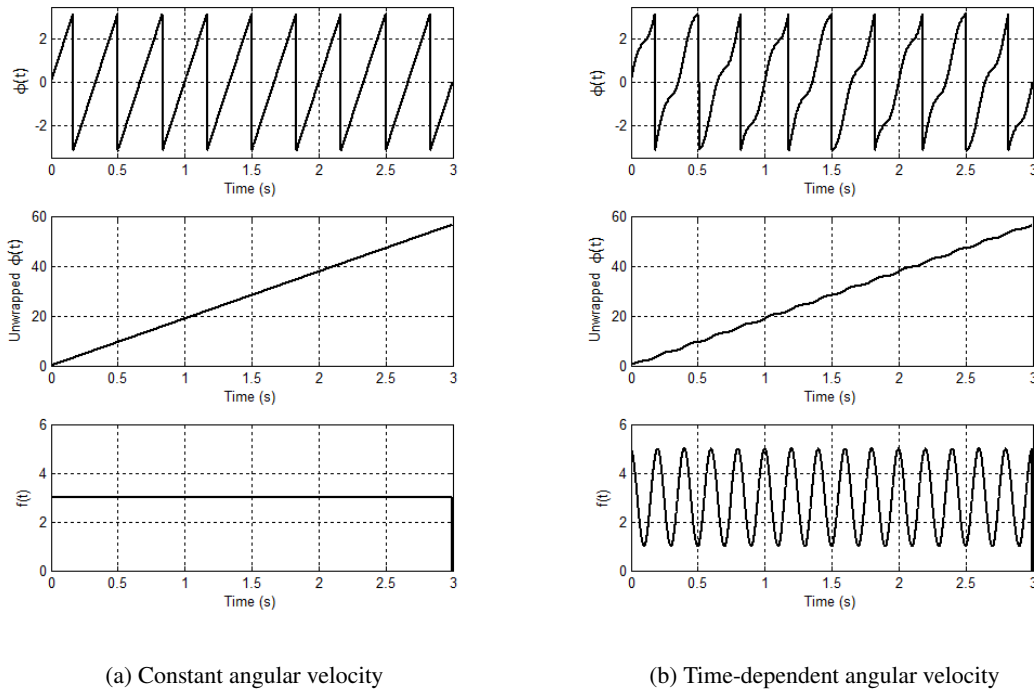


Figure 6.12 – Determination of the instantaneous frequency function : instantaneous phase function $\phi(t)$ (upper illustration), unwrapped instantaneous phase function (middle) and instantaneous frequency $f(t)$ (bottom).

Figure 6.12 illustrates these operations for the previous examples. The extracted instantaneous frequency time series corresponds exactly to the expected frequency modulation values. In the first example, the frequency value is constant (3Hz) and in the second example, the frequency of 3Hz is modulated by means of a oscillating function with a frequency of 5Hz and an amplitude equal to 2Hz (see relation (6.13)).

Notice that the existence of an instantaneous frequency value and its definition are not universally accepted by the research community. One reason is that the Fourier frequency is commonly defined as the inverse of the oscillation period. Therefore, the frequency may exist only if there is a whole oscillation cycle and that frequency should be constant over this duration.

6.4.3 Relevance of the instantaneous values

The previous considerations show that the tracking of the instantaneous values of amplitude or frequency enable analyzing signals that are produced via a non-stationary process. We have considered that the origin of axes is the center of rotation. What happens if that condition is not satisfied ?

Here, with a view to investigating the validity of the instantaneous values, three examples are considered : in each, a wheel is put in rotation, but the center of rotation, denoted z_0 , is placed at different positions on the real axis ($z_0 = 0$, $z_0 \in]0, r]$ or $z_0 > r$). The wheel radius $r = 1$ and the angular velocity $\omega_0 = 4\pi$ are constant.

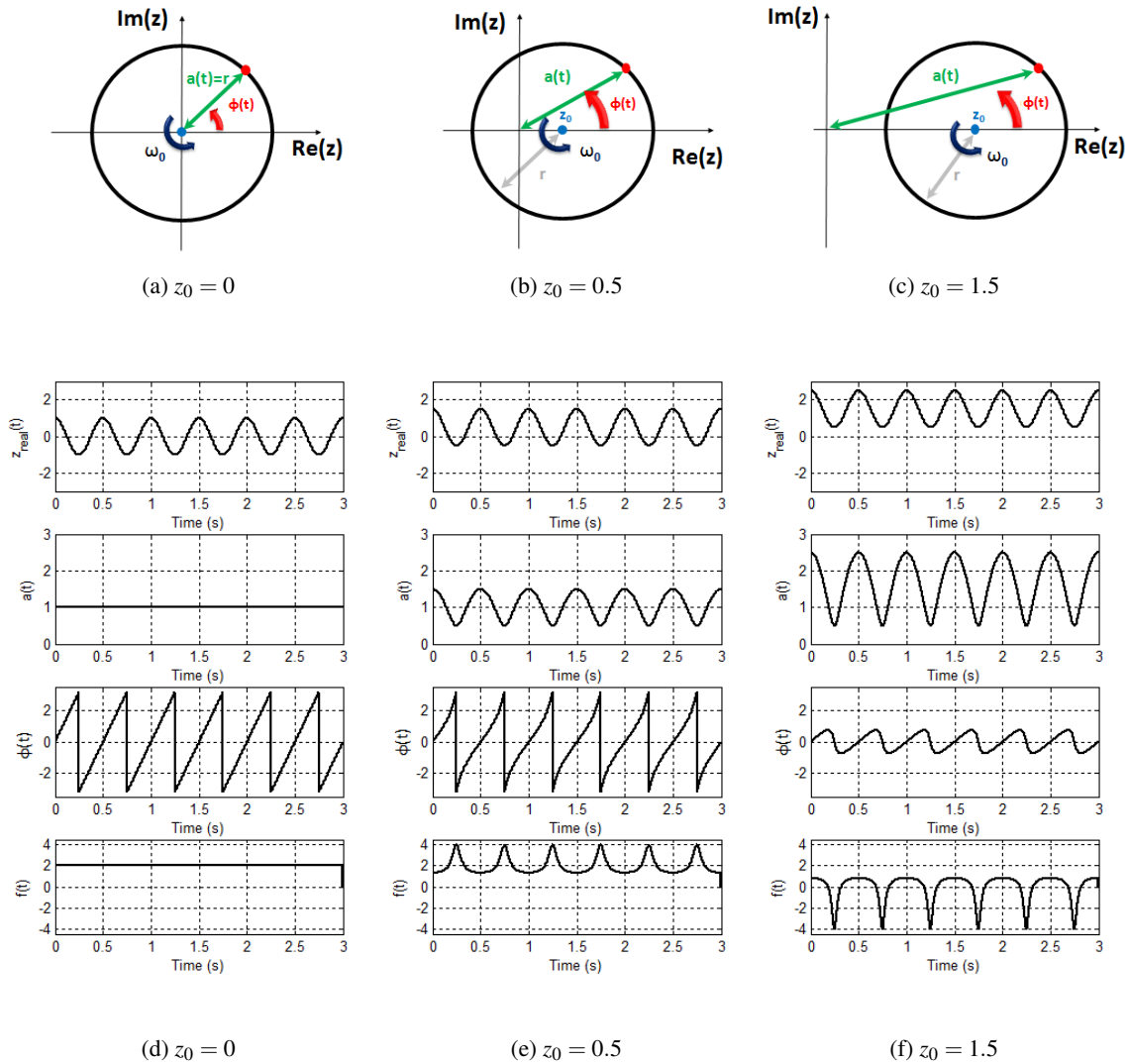


Figure 6.13 – Trajectory $z_{real}(t)$ and the corresponding computed instantaneous values for three different positions of the center of rotation, z_0 .

The results, in Figure 6.13, report the instantaneous amplitude $a(t)$, phase $\phi(t)$ and frequency $f(t)$ that have been computed via simulations. One observes that the computed values differ from the expected one if z_0 does not correspond to the origin of the axes. The frequency values are negative in the third example. Therefore, the signal has to be a locally narrow-band time-evolving function with local zero mean. Indeed, in the second and third examples, an additional constant

component appears, which biases the expected instantaneous values.

6.4.4 Analysis of multi-component time series

In several applications, the time series contains several time-varying spectral components. Therefore pre-processing has to be implemented with a view to decomposing the signal into a set of locally narrow-band time-evolving functions with local zero mean.

In this study, the signal is decomposed via a method, called Empirical Mode Decomposition (EMD). The goal of EMD consists in decomposing the time series into a sum of oscillating functions with respect to the local zero mean, called empirical modes. On the basis of these empirical modes, an empirical AM-FM decomposition is applied with a view to extracting the instantaneous envelope and frequency of each mode. The fluctuations of the instantaneous amplitude are thus assumed to evolve slowly compared to the frequency-modulated carrier. These methods are explained in the next chapter.

Key points

- Fourier analysis of time series is popular but suffers from an inherent limitation : the complex exponential basis functions are spread out over the entire time interval and are therefore not well adapted to the analysis of local events or data that are produced by non-stationary process (time-frequency resolution)
- The concepts of instantaneous amplitude and frequency offer a way to describe the frequency content of data that are produced by non-stationary processes
- A necessary condition to extract relevant instantaneous frequency and amplitude time series is that the analyzed signal is a locally narrow-band time-evolving function with local zero mean



7. Time-frequency analysis via Empirical mode decomposition

Objectives of this chapter

- Describe the time-frequency analysis of data via empirical mode decomposition
- Illustrate the advantages and drawbacks of EMD
- Describe the instantaneous value extraction via AM-FM decomposition of the empirical modes

Contents

7.1	Introduction	111
7.2	Empirical Mode Decomposition	111
7.2.1	Definition	111
7.2.2	Extraction of empirical modes	111
7.2.3	Sifting	113
7.3	Extraction of the instantaneous frequencies and envelopes	116
7.3.1	AM-FM decomposition	117
7.3.2	Computation of the instantaneous frequency of the empirical modes	119
7.4	Discussion	121
7.4.1	Preprocessing	121
7.4.2	Mode mixing : the empirical compromise	121
7.4.3	Conclusion	122

7.1 Introduction

The Empirical Mode Decomposition (EMD) algorithm is a tool for the analysis of multi-component signals. The most important property is that the basis functions are directly derived from the signal itself, i.e., the analysis method does not require a priori fixed basis functions as conventional analysis methods (e.g. Fourier or wavelet transforms). Another advantage is the perfect reconstruction of the analyzed signal.

It has been proposed initially in [Hua+98] to analyze data that are produced by a non-linear and non-stationary processes like ocean waves and has found applications in many fields such as geophysics, finance, biomedical signal processing and speech processing.

As shown in the previous chapter, the extraction of relevant instantaneous values relies on the properties of time series with regard to their local mean and their frequency content. EMD yields several locally narrow-band time-evolving functions with local zero mean, called *intrinsic mode functions (IMF)* or *empirical modes* which satisfy these requirements. The instantaneous frequency and envelope of the empirical modes are then obtained via an AM-FM decomposition.

7.2 Empirical Mode Decomposition

7.2.1 Definition

Let us consider an arbitrary signal $x(t)$. That signal may be expressed as the sum of I oscillating functions $c_i(t)$ and one residue $r(t)$ as follows :

$$x(t) = \sum_{i=1}^I c_i(t) + r(t) \quad (7.1)$$

Here, functions $c_i(t)$, called *intrinsic mode functions (IMF)* or *empirical modes*, are locally narrow-band time-evolving functions with local zero mean. They have to satisfy the two following conditions :

- Time-evolving alternating functions : the number of signal extrema and the number of zero crossings have to be equal or differ by one.
- Local zero mean : the average of the upper and lower signal envelope (obtained respectively on the base of the positions and amplitudes of minima and maxima) has to be equal to zero for each instant t .

As an example, Figure 7.1 illustrates the decomposition of a time series into three empirical modes and one residue. One observes that the first extracted mode corresponds to the fastest and the last mode to the slowest fluctuations. One also observes a desirable property of the empirical mode decomposition when compared to band-pass filtering. Namely that the original time series can be perfectly reconstructed by summing the empirical modes and the residue.

7.2.2 Extraction of empirical modes

The extraction of the empirical modes relies on an iterative decomposition, called *sifting*. The sifting consists in subtracting iteratively the local average time series of the signal to obtain a function that satisfies the IMF properties.

Let us consider a signal $x(t)$ of length T and introduce the following notations : $h_{(i,k)}(t)$, where $h(t)$ is a time series, and (i,k) refers to the application of k^{th} sifting iteration to the function $h(t)$ to

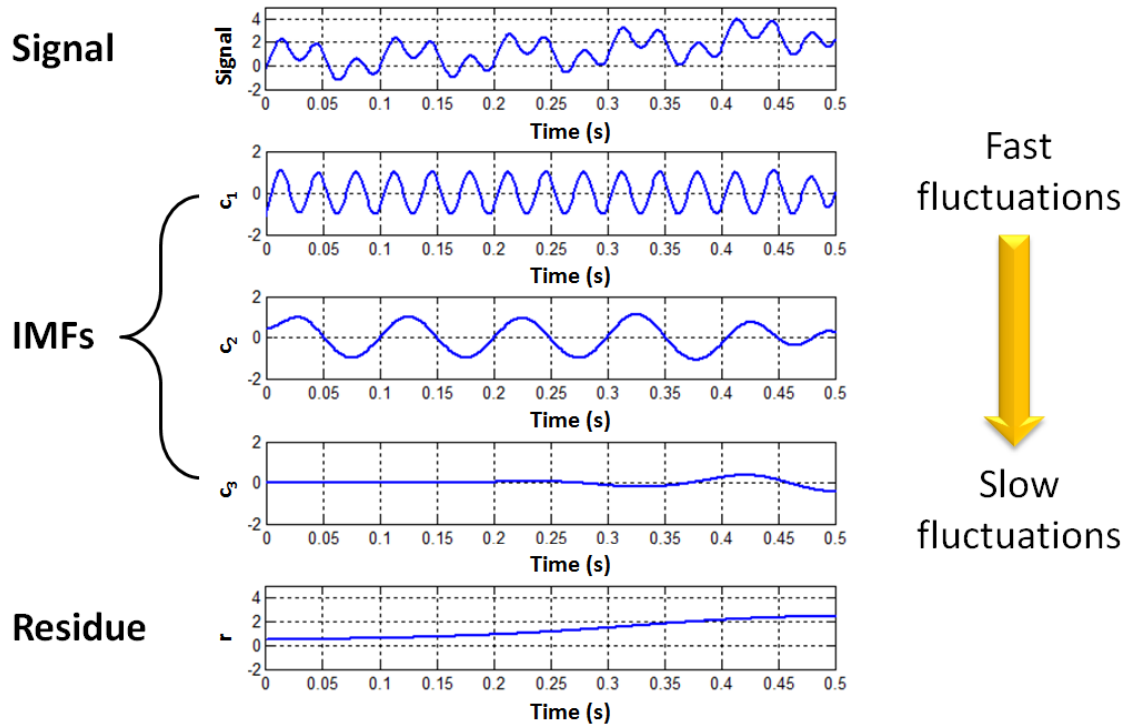


Figure 7.1 – Example of the empirical mode decomposition of a time series into three empirical modes (IMF) and one residue

extract the i^{th} empirical mode, denoted $c_i(t)$.

So, initially, $x(t) = h_{(1,1)}(t)$, and the sifting is applied to $h_{(1,1)}(t)$ to determine the first empirical mode $c_1(t)$. The signal $x(t)$ may at present be expressed as the sum of that mode $c_1(t)$ and a rest $h_{(2,1)}(t)$ obtained by subtraction, as follows :

$$x(t) = h_{(1,1)}(t) = c_1(t) + \underbrace{(h_{(1,1)}(t) - c_1(t))}_{h_{(2,1)}(t)} = c_1(t) + h_{(2,1)}(t) \quad (7.2)$$

More generally, applying the sifting to time series $h_{(i,1)}(t)$ enables the extraction of empirical mode $c_i(t)$. Let us assume that i modes have been determined, then the signal $x(t)$ may be expressed as follows :

$$\begin{cases} x(t) = \left(\sum_{n=1}^i c_n(t) \right) + h_{(i+1,1)}(t) \\ h_{(i+1,1)}(t) = h_{(i,1)}(t) - c_i(t) \end{cases} \quad (7.3)$$

A test is carried out on $h_{(i+1,1)}(t)$ to determine whether it is the residue of the decomposition. Here, one considers that a time series is the residue if its number of extrema is ≤ 3 . If the time series $h_{(i+1,1)}(t)$ satisfies this test, the process is stopped and the residue of the decomposition $r(t) = h_{(i+1,1)}(t)$ is found. Otherwise, a new sifting is applied to $h_{(i+1,1)}(t)$ to determine the next empirical mode $c_{i+1}(t)$, and so on, until obtaining a residue $r(t)$. At the end, the time series $x(t)$ is decomposed into the sum of I empirical modes and a residue $r(t)$:

$$x(t) = \left(\sum_{n=1}^I c_n(t) \right) + r(t) \quad (7.4)$$

7.2.3 Sifting

Starting from a signal $h_{(i,k)}(t)$ where $k = 1$, the extraction of the empirical mode $c_i(t)$ involves the following sifting steps. First, the extrema positions (peak and valley samples) of $h_{(i,k)}(t)$ are detected. The lower, $env_{l,(i,k)}(t)$, and upper, $env_{u,(i,k)}(t)$, envelopes of $h_{(i,k)}(t)$ are build by means of cubic spline interpolation of respectively the signal minima and maxima. Observe that the cubic spline interpolation based on the extrema positions and values is affected by the lack of information outside the signal boundaries. Therefore, boundary effect management, based on a symmetrical local reconstruction of the signal, is proposed in subsection 7.2.3.1.

The local average time series $m_{(i,k)}(t)$ is then obtained by taking the average of the lower and upper signal envelopes :

$$m_{(i,k)}(t) = \frac{env_{u,(i,k)}(t) + env_{l,(i,k)}(t)}{2} \quad (7.5)$$

Figure 7.2 illustrates the application of the k^{th} sifting iteration to a signal $h_{(i,k)}(t)$.

A test with regard to IMF properties is then applied. That test, explained in subsection 7.2.3.2, involves the number of extrema and zero-crossings as well as a local criterion assessing the local average. If the time series $h_{(i,k)}(t)$ satisfies the test, the empirical mode $c_i(t) = h_{(i,k)}(t)$ is found and sifting is stopped. Otherwise, that local average time series $m_{(i,k)}(t)$ is subtracted from $h_{(i,k)}(t)$ to build the next candidate empirical mode $h_{(i,k+1)}(t)$:

$$h_{(i,k+1)}(t) = h_{(i,k)}(t) - m_{(i,k)}(t) \quad (7.6)$$

Another iteration involving the previous steps and tests is carried out on $h_{(i,k+1)}(t)$ and so on, until the empirical mode $c_i(t) = h_{(i,k^*)}(t)$ is found after k_i^* iterations. At the end, the time series $h_{(i,1)}(t)$ may be expressed as the sum of that mode $c_i(t)$ and a provisional rest $h_{(i+1,1)}(t)$:

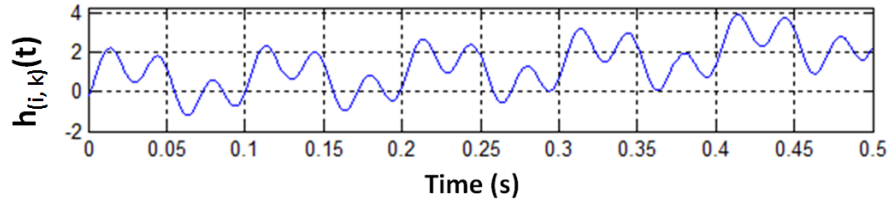
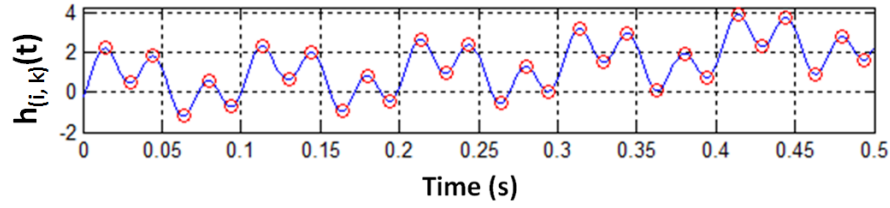
$$h_{(i,1)}(t) = h_{(i,k^*)}(t) + h_{(i+1,1)}(t) = c_i(t) + h_{(i+1,1)}(t) \quad (7.7)$$

An interesting property of the sifting is that mode $c_i(t)$ corresponds to the locally fastest fluctuations of $h_{(i,1)}(t)$. Indeed, the provisional rest $h_{(i+1,1)}(t)$ is the sum of all local average time series obtained during the sifting process giving mode $c_i(t)$. These local average time series are expected to evolve locally slower than the signal.

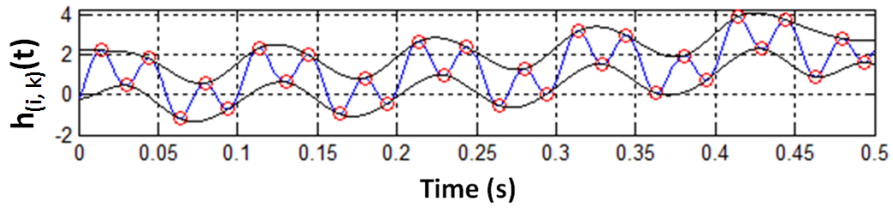
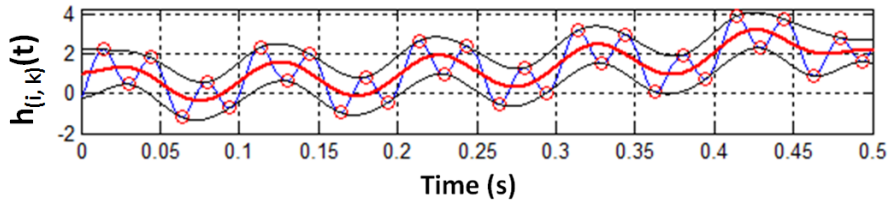
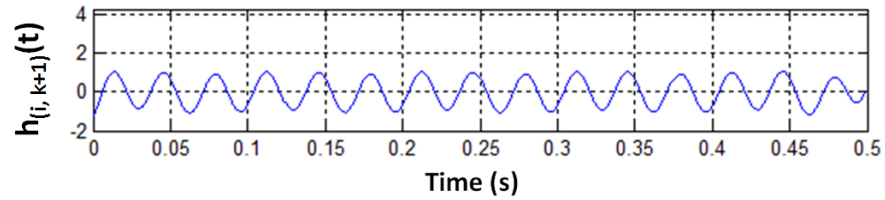
$$h_{(i+1,1)}(t) = \sum_{k=1}^{k^*-1} m_{(i,k)}(t) \quad (7.8)$$

7.2.3.1 Boundary effect management

As said previously, the local average $m_{(i,k)}(t)$ is determined on the base of the position and value of signal extrema via polynomial interpolation. With a view to estimating the interpolation polynomial coefficients, the positions of four successive peaks (or valleys) have to be known. However, at the signal boundaries, these conditions are not satisfied. Here, the boundary effects are managed by the mirror method [RFG03]. For that purpose, the signal shape is reconstructed by mirror symmetry with regard to a swivel sample.

(a) Candidate empirical mode $h_{(i,k)}(t)$ 

(b) Detection of extrema positions and values

(c) Determination of the lower, $env_{l,(i,k)}(t)$, and upper, $env_{u,(i,k)}(t)$, envelopes by polynomial interpolation of extrema(d) Computation of the local average $m_{(i,k)}(t)$ (e) Obtaining of a new candidate empirical mode $h_{(i,k+1)}(t)$ by subtracting the local average time series $m_{(i,k)}(t)$ from $h_{(i,k)}(t)$ Figure 7.2 – Application of the k^{th} sifting iteration to a signal $h_{(i,k)}(t)$

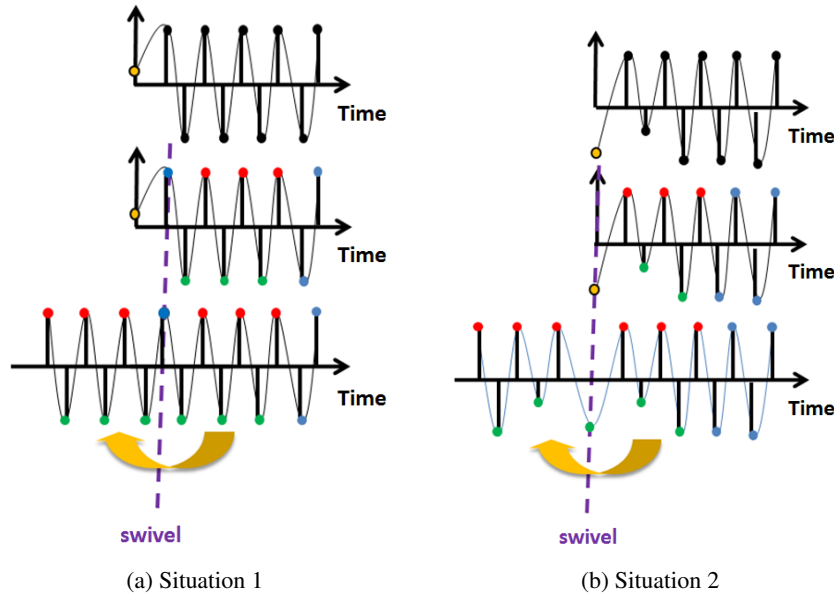


Figure 7.3 – Boundary effect management by mirror symmetry

Let us consider a time series $h_{(i,k)}(t)$. Eight possible configurations can occur (four at each boundary) according to the positions and amplitudes of the first (*resp.* last) signal sample ($h_{(i,k)}(0)$), first (*resp.* last) signal minimum ($[t_{\min(1)}, h_{(i,k)}(t_{\min(1)})]$) and first (*resp.* last) signal maximum ($[t_{\max(1)}, h_{(i,k)}(t_{\max(1)})]$).

Notice that, for the sake of simplicity, these eight different configurations can be explained on the base of the two following situations (see Figure 7.3) :

- Situation 1 : $t_{\min(1)} > t_{\max(1)}$ and $h_{(i,k)}(0) > h_{(i,k)}(t_{\min(1)})$. The swivel sample will be the first signal maximum.
- Situation 2 : $t_{\min(1)} > t_{\max(1)}$ and $h_{(i,k)}(0) < h_{(i,k)}(t_{\min(1)})$. The swivel sample will be the first signal sample.

7.2.3.2 IMF test

The sifting is stopped if function $h_{(i,k)}(t)$ satisfies simultaneously the two IMF conditions, given in section 7.2.1. The first condition, with regard to the time-evolving oscillations, is easy to check by counting the number of extrema and zero crossings but the second condition, with regard to the local zero average, is more difficult to assess in practice. Several criteria have been proposed in the literature [Hua+98] [RFG03]. In [Hua+98], the local average is assessed via the size of the standard deviation, denoted SD , computed from the two consecutive sifting iterations k and $k+1$.

$$SD = \frac{\sum_{t=0}^{T-1} m_{(i,k)}^2(t)}{\sum_{t=0}^{T-1} h_{(i,k)}^2(t)} = \frac{\sum_{t=0}^{T-1} |h_{(i,k)}(t) - h_{(i,k+1)}(t)|^2}{\sum_{t=0}^{T-1} h_{(i,k)}^2(t)} \leq \alpha \quad (7.9)$$

If SD is below a threshold α and the first condition is verified, then function $h_{(i,k)}(t)$ is a mode function. A typical limit value α for SD is proposed to be between 0.2 and 0.3. However, a drawback of this approach is that small values of $h_{(i,k)}(t)$ highly affect this criterion, so that several

additional sifting iterations have to be applied to satisfy it.

For these reasons, another criterion, proposed in [RFG03], is used. It is based on two thresholds : θ_1 and θ_2 . The goal is to guarantee globally small fluctuations in the local average time series, while taking into account locally large signal excursions. One introduces the mode amplitude $a(t)$ and the evaluation function $\sigma(t)$ defined as follows :

$$a_{(i,k)}(t) = \frac{env_{u,(i,k)}(t) - env_{l,(i,k)}(t)}{2} \quad \text{and} \quad \sigma_{(i,k)}(t) = \left| \frac{m_{(i,k)}(t)}{a_{(i,k)}(t)} \right| \quad (7.10)$$

So, the sifting is iterated until $\sigma(t) < \theta_1$ for some prescribed fraction $(1 - \zeta)$ of the total duration, while $\sigma(t) < \theta_2$ for the remaining fraction. The threshold values θ_1 and θ_2 arise from a compromise between the over-decomposition resulting from an over-iteration (small threshold values) and the relevance of the extracted mode. In [RFG03], default values have been typically chosen to be equal to $\theta_1 = 0.05$ and $\theta_2 = 10 \theta_1$ and the parameter ζ is 0.05. In this study, the parameters θ_1 and θ_2 have been chosen 10 times smaller than default values, so that :

$$\begin{cases} \theta_1 = 0.005 \\ \theta_2 = 10 \theta_1 \\ \zeta = 0.05 \end{cases} \quad (7.11)$$

As a consequence, the number of iterations during sifting as well as the number of extracted empirical modes increase, but the local average time series of each mode is closer to zero.

7.3 Extraction of the instantaneous frequencies and envelopes

The concept of instantaneous frequency and instantaneous amplitude enables the analysis of the frequency content of non-stationary mono-component signals. The idea consists in considering that the frequency and amplitude should be a function of time and have an instantaneous value. The extraction of instantaneous values is relevant if the time series are time-evolving oscillating functions with local zero mean. The empirical modes are thus possible candidate functions.

Here, each mode $c_i(t)$ is analyzed individually. The extraction of instantaneous frequency and envelope relies on an empirical AM-FM decomposition [Hua+09] the goal of which consists in expressing the mode as the product of two components as follows :

$$c_i(t) = a_i(t) \cos(\phi_i(t)) \quad (7.12)$$

The first component, $a_i(t)$, is the time-varying mode function envelope (AM component or instantaneous amplitude) and the second, $\cos(\phi_i(t))$, called the carrier, is a purely frequency modulated function with unit amplitude, where $\phi_i(t)$ designates the instantaneous phase. The instantaneous frequency $f_i(t)$ is then obtained by differentiating of the instantaneous phase $\phi_i(t)$ after phase unwrapping :

$$f_i(t) = \frac{1}{2\pi} \frac{d\phi_i(t)}{dt} \quad (7.13)$$

7.3.1 AM-FM decomposition

The AM-FM decomposition is based on an iterative process, with k the iterating variable. Starting from $s_{(i,k)}(t)$, with $s_{(i,1)}(t) = c_i(t)$, a new time series, denoted $s_{(i,k+1)}(t)$ is obtained by normalizing the time series $s_{(i,k)}(t)$ by its positive envelope, denoted $env_{(i,k)}(t)$:

$$s_{(i,k+1)}(t) = \frac{s_{(i,k)}(t)}{env_{(i,k)}(t)} \quad (7.14)$$

The normalization of $c_i(t)$ does not affect its local average. Assuming that the local average of $s_{(i,k)}(t)$ is close to zero for each time t , the envelope $env_{(i,k)}(t)$ is obtained as follows :

1. Determination of maxima positions and values of absolute value time series $|s_{(i,k)}(t)|$.
2. Obtaining the envelope $env_{(i,k)}(t)$ by cubic spline interpolation of extrema.

A test is then carried out to determine whether the new time series is the carrier. The normalization is stopped if the time series $s_{(i,k+1)}(t)$ satisfies the test, which is that its absolute sample values are smaller than unity :

$$|s_{(i,k+1)}(t)| \leq 1 \quad \forall t \quad (7.15)$$

Otherwise, a new iteration involving the preceding steps is carried out until all values of the normalized function are less or equal to unity, after k^* iterations.

After this normalization, the empirical mode $c_i(t)$ can be decomposed as the product of an oscillating time series $s_{(i,k^*+1)}(t)$ with local average close to zero and a slowly time-varying function, denoted $env_{(i,k^*)}^*(t)$, obtained on the basis of all the preceding computed envelopes.

$$c_i(t) = env_{(i,k^*)}^*(t) s_{(i,k^*+1)}(t) = \left(\prod_{j=1}^{k^*} env_{(i,j)}(t) \right) s_{(i,k^*+1)}(t) \quad (7.16)$$

The mode function $c_i(t)$ can be written as :

$$c_i(t) = a_i(t) \cos(\phi_i(t)) \quad (7.17)$$

The carrier $\cos(\phi_i(t))$ and the instantaneous envelope $a_i(t)$ are given by :

$$\begin{cases} a_i(t) = env_{(i,k^*)}^*(t) = \prod_{j=1}^{k^*} env_{(i,j)}(t) \\ \cos(\phi_i(t)) = s_{(i,k^*+1)}(t) \end{cases} \quad (7.18)$$

As an example, Figure 7.4 illustrates the decomposition of an empirical mode $c_i(t)$ obtained by simulation. The carrier as well as the envelope have been simulated on the base of a linear swept-frequency cosine signal, where the instantaneous frequencies, $f_1(t)$ and $f_2(t)$ evolve linearly from $20Hz$ to $100Hz$ and from $10Hz$ to $5Hz$ during $0.3s$:

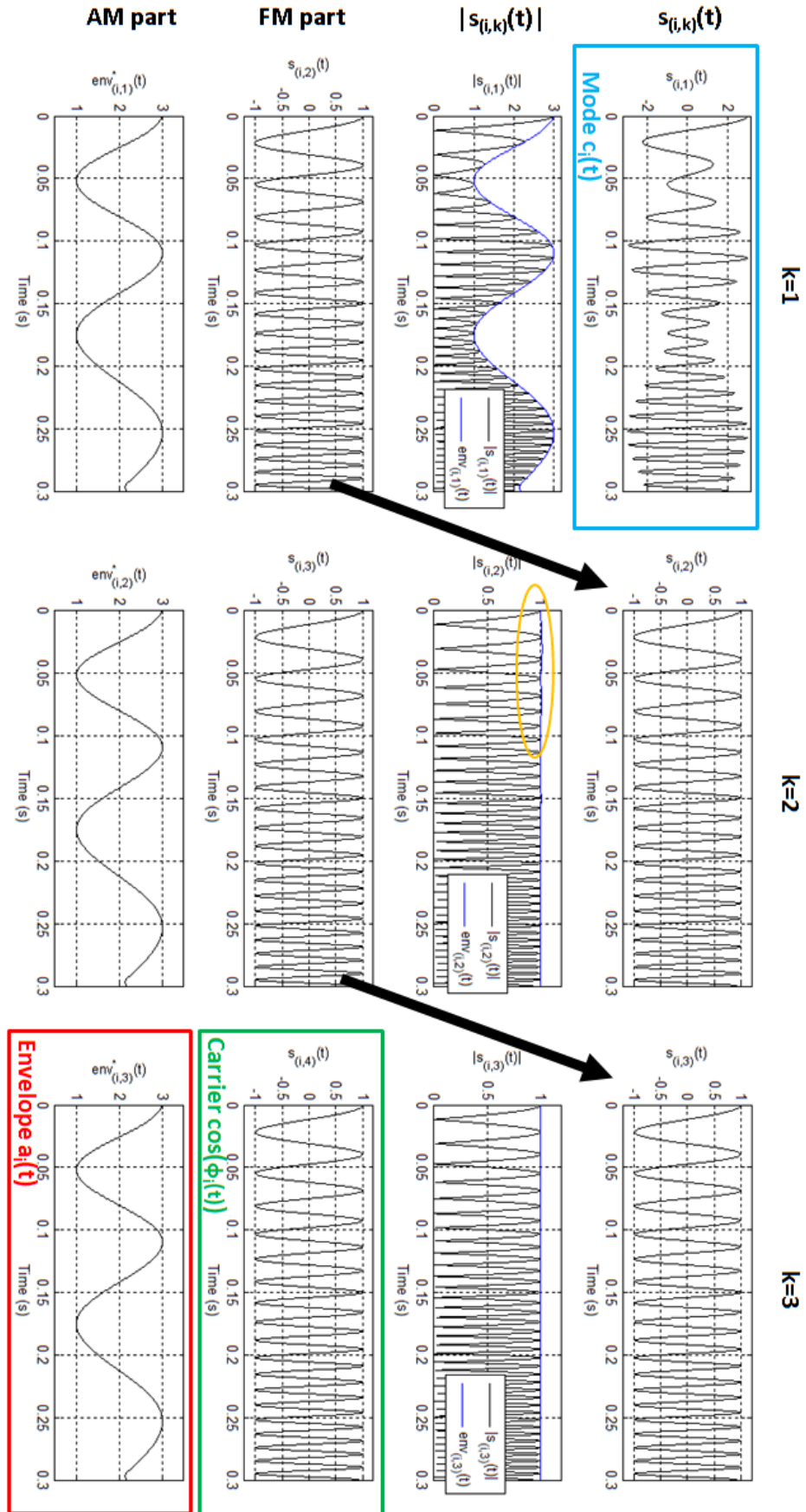


Figure 7.4 – Example of the AM-FM decomposition of a simulated empirical mode $c_i(t)$ into a carrier $\cos(\phi_i(t))$ and an envelope $a_i(t)$

$$\begin{cases} c_i(t) = (2 + \cos(\phi_1(t))) \cos(\phi_2(t)) \\ \phi_1(t) = 2\pi \int_0^t f_1(u) du = 2\pi \int_0^t \left(10 - \frac{5}{0.3}u\right) du \\ \phi_2(t) = 2\pi \int_0^t f_2(u) du = 2\pi \int_0^t \left(20 + \frac{80}{0.3}u\right) du \end{cases} \quad (7.19)$$

One observes that only a few iterations are required to decompose a mode function into instantaneous amplitude and frequency-modulated carrier.

7.3.2 Computation of the instantaneous frequency of the empirical modes

The instantaneous frequency $f_i(t)$ is obtained by numerical differentiation of the instantaneous phase $\phi_i(t)$ after phase unwrapping. The instantaneous phase $\phi_i(t)$ is first obtained by computing the inverse cosine of the carrier. Figure 7.5 illustrates the instantaneous phase of the carrier that has been determined in Figure 7.4.

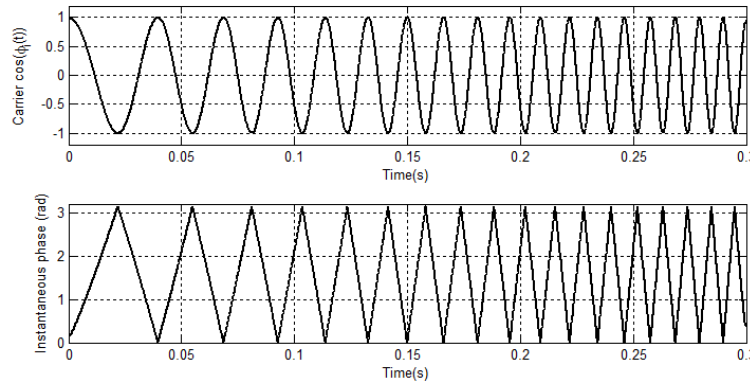


Figure 7.5 – Determination of the instantaneous phase $\phi_i(t)$ by computing the inverse cosine (example)

The results, expressed in radian, are given in the range $[0, \pi]$. Therefore, additional processing is implemented to map the phase into the four trigonometric quadrants (in the range $[0, 2\pi]$). For that, the quadrature function, $\sin(\phi_i(t))$, of the carrier is obtained. The instantaneous four-quadrant phase is then computed via :

$$\phi_i(t) = \arctan\left(\frac{\sin(\phi_i(t))}{\cos(\phi_i(t))}\right) \quad (7.20)$$

Notice that the iterated envelope normalization of the AM-FM decomposition affects slightly the shape of the carrier near its extrema. As a consequence, some instabilities occur in the instantaneous phase time series during approximatively 5 sampling steps near the extrema. To avoid error propagation, the instantaneous phase time series is smoothed by means of a 20th-order median filter and a 20th-order moving average filter (whatever the sampling frequency).

After phase unwrapping, the instantaneous frequency $f_i(t)$ is then obtained by differentiating the instantaneous phase $\phi_i(t)$.

$$f_i(t) = \frac{1}{2\pi} \frac{d\phi_i(t)}{dt} \quad (7.21)$$

As proposed in [Boa92], a q^{th} order generalized phase difference estimator is used for the numerical differentiation. The estimator order has been chosen equal to $q = 6$ and the coefficients b_k are equal to $[-\frac{1}{60}, \frac{3}{20}, -\frac{3}{4}, 0, \frac{3}{4}, -\frac{3}{20}, \frac{1}{60}]$:

$$f_i(n) = \frac{1}{2\pi} \sum_{k=-q/2}^{q/2} b_k \phi(n+k) \quad (7.22)$$

Figure 7.6 illustrates the steps involved in the determination of the instantaneous frequency time series.

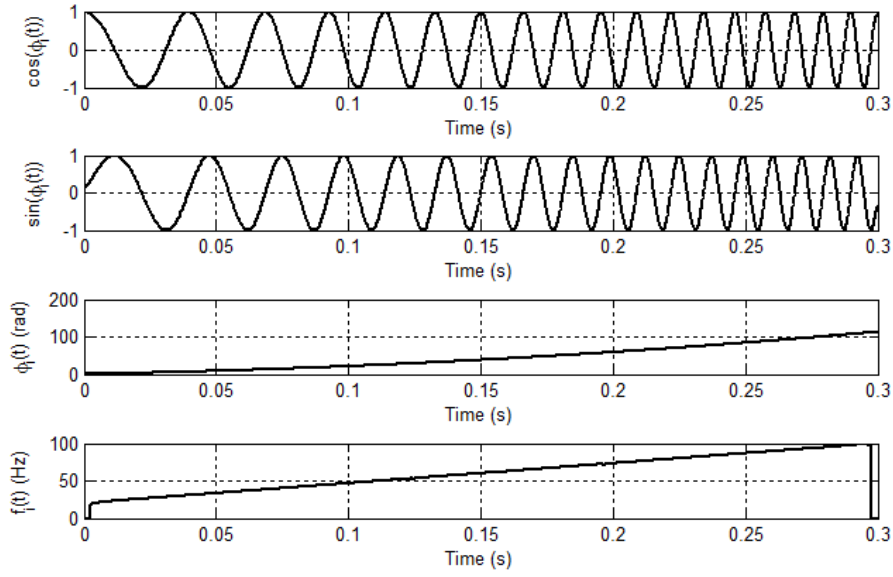


Figure 7.6 – Determination of the instantaneous frequency time series $f_i(t)$ by computing the carrier quadrature $\sin(\phi_i(t))$, computing the four-quadrant inverse tangent $\phi_i(t)$ and differentiating the unwrapped phase.

7.4 Discussion

Empirical mode decomposition enables the time-frequency analysis of data that are produced by non-linear and non-stationary processes and enables the extraction of the instantaneous frequency and envelope of the empirical modes via AM-FM decomposition. The main advantages of EMD, as opposed to conventional decomposition methods, is that it does not require an a priori selection of basis functions. Another advantage is the perfect reconstruction of the analyzed signal. However, the approach is empirical, so that precautions have to be taken with regard to obtaining and interpreting the results.

7.4.1 Preprocessing

Firstly, assuming that the sifting is based on the positions of extrema, a good temporal localization of these extrema is required. As a consequence, a desirable preprocessing step consists in up-sampling the analyzed signal.

7.4.2 Mode mixing : the empirical compromise

A major drawback of EMD, known as *mode mixing*, is frequently observed. Mode mixing occurs when a same frequency is locally shared by several successive empirical modes. As a consequence, the orthogonality between consecutive empirical modes is not guaranteed theoretically.

Mode mixing is a consequence of signal intermittency [HW08] [Hua+98] causing local perturbations in the signal to affect globally the extracted empirical modes. Therefore, serious aliasing may occur in the time-frequency distribution and individual modes may lose their physical meaning.

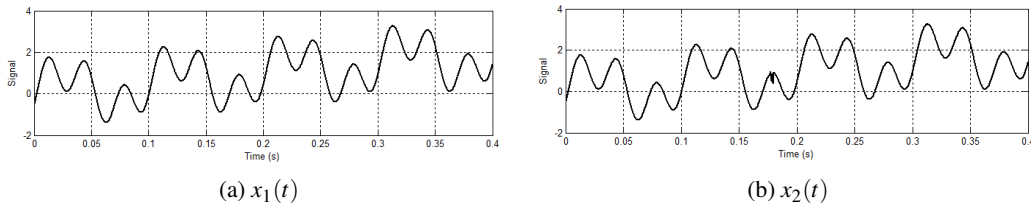


Figure 7.7 – Illustration of the mode mixing : Time series $x_1(t)$ and $x_2(t)$

As an example, let consider two signals $x_1(t)$ and $x_2(t)$. The signal $x_1(t)$ is a sum of two oscillating functions and a linear trend :

$$x_1(t) = \sin(2\pi 30t - \pi/5) + \sin(2\pi 10t) + 5t \quad (7.23)$$

Signal $x_2(t)$ is almost the same as $x_1(t)$, but a local perturbation, equal to a low-pass filtered white noise, has been added around $t \approx 0.18s$. Figure 7.7 illustrates these two time series. For each time series, EMD has been applied, decomposing $x_1(t)$ and $x_2(t)$ into a sum of 3 or 4 empirical modes and a residue (Figure 7.8).

A visual inspection of Figure 7.8a shows that signal $x_1(t)$ comprises two oscillating functions at 30Hz and 10Hz. One also observes in Figure 7.8b that the frequency component at 30Hz is missing in the first extracted mode $c_1(t)$ in the interval $[0.15s, 0.2s]$ but appears in the second and third modes. Moreover, the component at 10Hz is not easily detected. These observations suggest that mode mixing affects the physical meaning of individual modes.

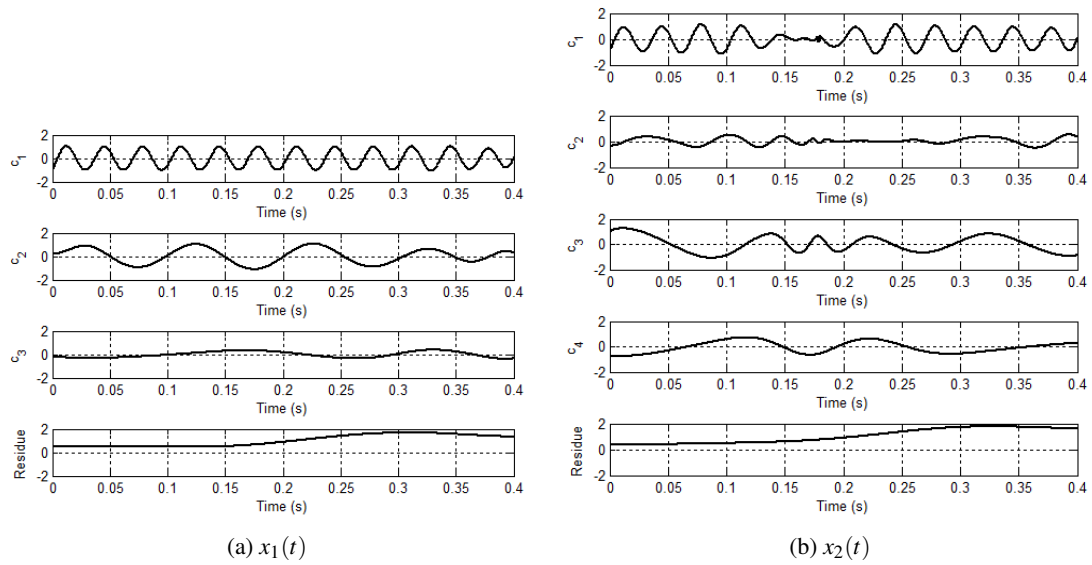
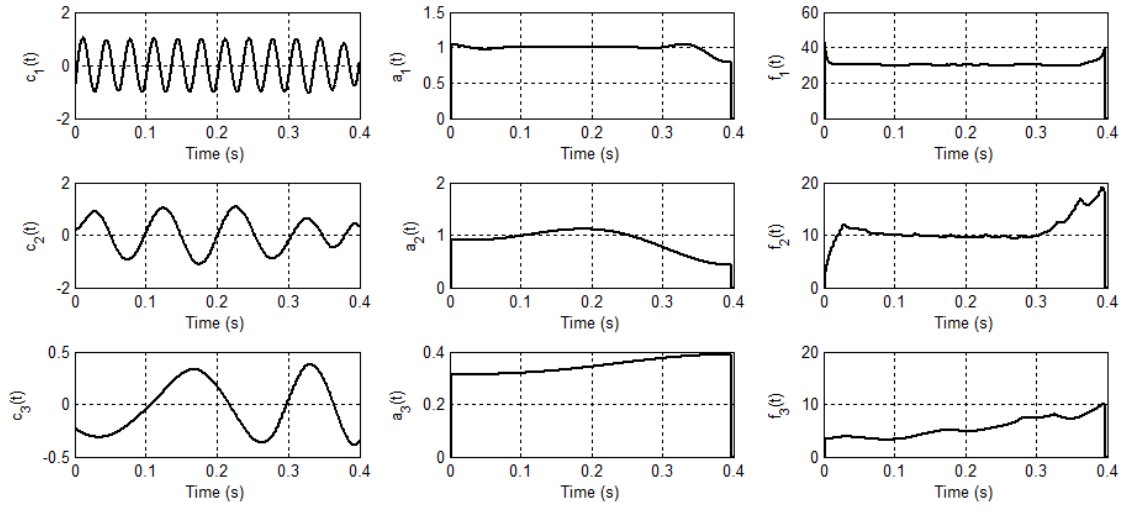
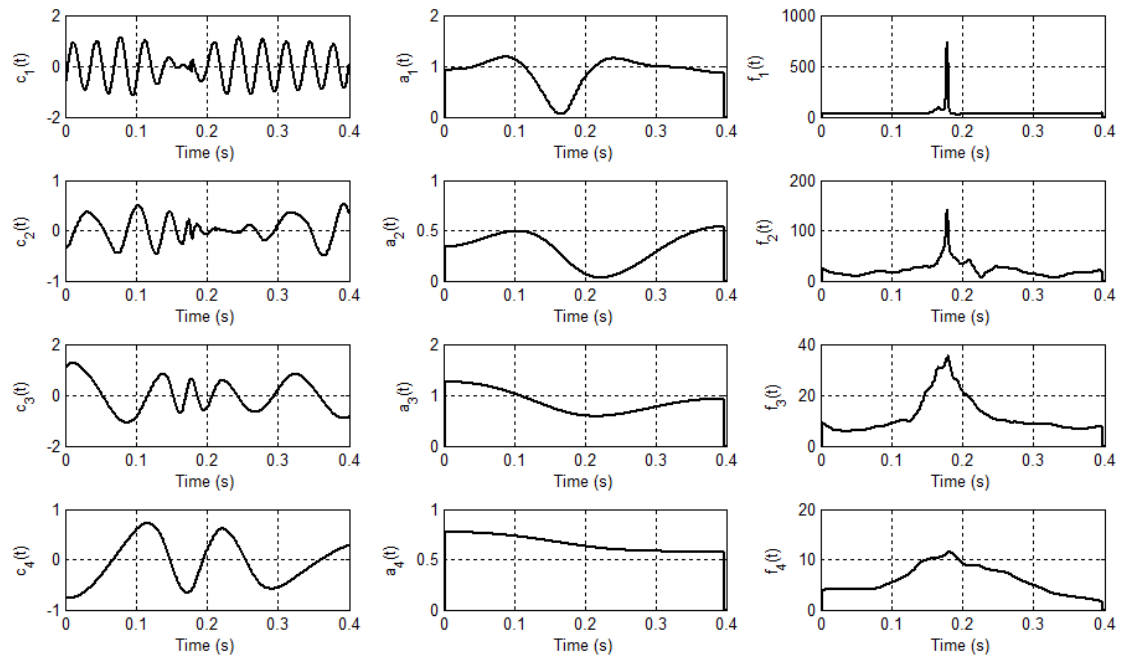


Figure 7.8 – Illustration of the mode mixing : Empirical mode decomposition of $x_1(t)$ and $x_2(t)$

The same conclusions can be drawn on the basis of the instantaneous frequency $f_i(t)$ and envelope $a_i(t)$ time series of the IMFs. One observes in Figure 7.9a that the frequencies at 10Hz and 30Hz are clearly identified for the major part of the signal $x_1(t)$, except at the end where boundary effects occur. In Figure 7.9b, one observes that the instantaneous frequency time series varies rapidly in the interval $[0.15\text{s}, 0.20\text{s}]$. There, the analysis of individual mode functions does not report relevant information with regard to frequency content. However, by considering several modes at the same time and taking into account simultaneously the instantaneous frequency and envelope time series, the components at 10Hz and 30Hz may be identified.

7.4.3 Conclusion

Empirical mode decomposition enables the analysis of time series produced by non-stationary process, but the analysis of instantaneous values may require the combination of the characteristics of several mode functions.

(a) $x_1(t)$ (b) $x_2(t)$ Figure 7.9 – Illustration of the mode mixing : Instantaneous frequency $f_i(t)$ and envelope $a_i(t)$ time series of IMF's

Key points

- Empirical mode decomposition (EMD) breaks down a signal into locally narrow-band time evolving functions with local zero mean
- EMD does not require a priori fixed basis function and enables a perfect reconstruction of the original time series by summing the extracted empirical modes
- The major drawback of EMD is known as *mode mixing* and refers to the lack of theoretical statements guaranteeing the orthogonality of empirical modes
- The instantaneous mode function envelope and phase are extracted via an empirical AM-FM decomposition. The instantaneous frequency is obtained by numerical differentiation after phase unwrapping



8. Analysis of vocal cycle length perturbations

Objectives of this chapter

- Apply and illustrate the empirical mode decomposition & AM-FM decomposition of the vocal cycle lengths time series
- Propose and illustrate a method to break-up the cycle length time series into sub time series that are respectively assigned to vocal jitter, neurological tremor, physiological tremor and a residual trend
- Propose and define cues of cycle length perturbation size
- Propose and define cue candidates of neurological tremor frequency and bandwidth

Contents

8.1	Introduction	127
8.2	Time-frequency analysis of the vocal cycle length time series	127
8.3	Categorization	131
8.4	Average vocal frequency and cycle length perturbation size	135
8.4.1	Average vocal frequency	135
8.4.2	Perturbation sizes	135
8.5	Neurological tremor frequency	136
8.5.1	Overview	136
8.5.2	Complex neurological tremor time series	136
8.5.3	Neurological tremor frequency	137
8.5.4	Neurological tremor frequency estimate	139
8.5.5	Neurological tremor frequency content analysis	141
8.6	On the choice of the weights	149
8.7	Conclusions	154

8.1 Introduction

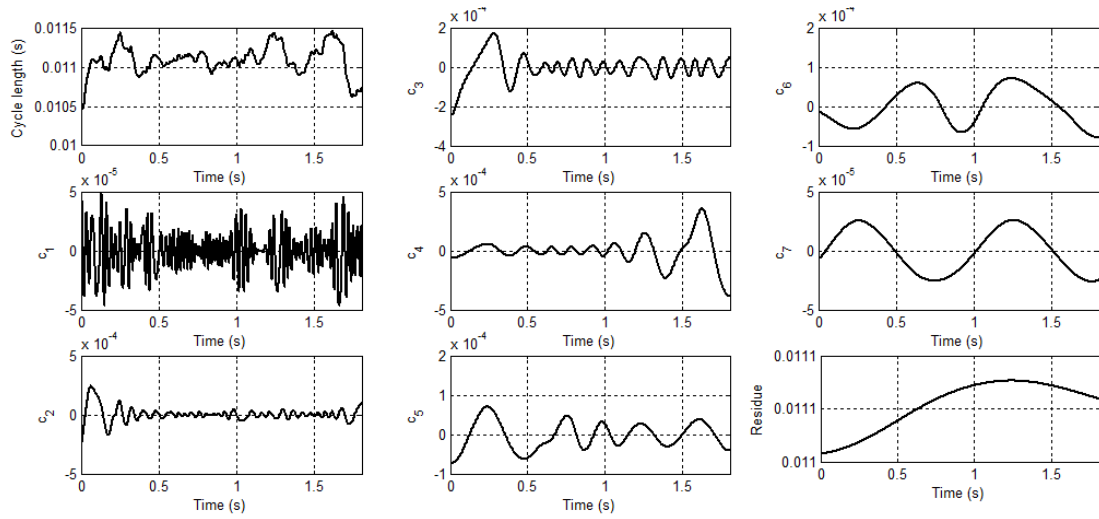
In chapter 7, Empirical Mode Decomposition (EMD) has been proposed to decompose a signal into locally narrow-band time-evolving functions with local zero mean, called *empirical modes* or IMFs. EMD is now applied to the vocal cycle length time series obtained by the methods described in chapter 5. The mode instantaneous envelopes and frequencies are then extracted via empirical AM-FM decomposition. According to their typical instantaneous frequency, these modes are then assigned to four categories and added : intonation and declination, physiological tremor, neurological tremor and vocal jitter. The so obtained time series are then further analyzed with a view to obtaining the neurological tremor size and frequency as well as jitter size.

The chapter is organized as follows : In the first section, the time-frequency analysis via EMD is applied to the vocal cycle length time series. Then, the categorization is explained. Finally, the analysis of the size of perturbations as well as the frequency content of neurological tremor is investigated. In each section, the method is illustrated on the basis of 3 sustained vowels [a] produced by the same three speakers as in chapter 5.

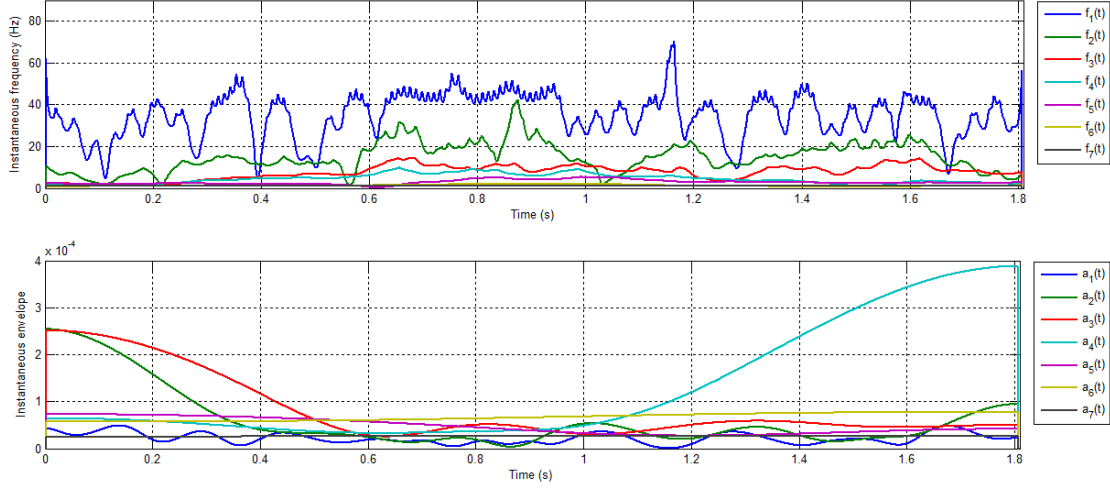
8.2 Time-frequency analysis of the vocal cycle length time series

EMD is applied to the cycle length time series. The cycle length time series has been constant-step interpolated at a sampling frequency $F_s = 8kHz$. A high sampling frequency has been chosen to guarantee a good temporal localization of extrema during EMD sifting. The parameters used for the IMF test during the sifting process are : $\theta_1 = 0.005$, $\theta_2 = 10\theta_1$ and $\zeta = 0.05$ (see section 7.2.3.2). Figures 8.1a, 8.2a and 8.3a illustrate the extracted empirical modes c_n for each time series (3 speakers). Seven empirical modes have been extracted in each case but, in general, this number may vary. One also observes in the middle column of Figure 8.2a and 8.3a the effects of mode mixing (discussed in section 7.4).

Figures 8.1b, 8.2b and 8.3b illustrate the instantaneous frequencies and envelopes that have been obtained via AM-FM decomposition. Visual inspection shows that the frequency values are comprised between 0 and $F_0/2$ (as expected). Moreover, mode mixing is clearly observed in this kind of time-frequency representation.

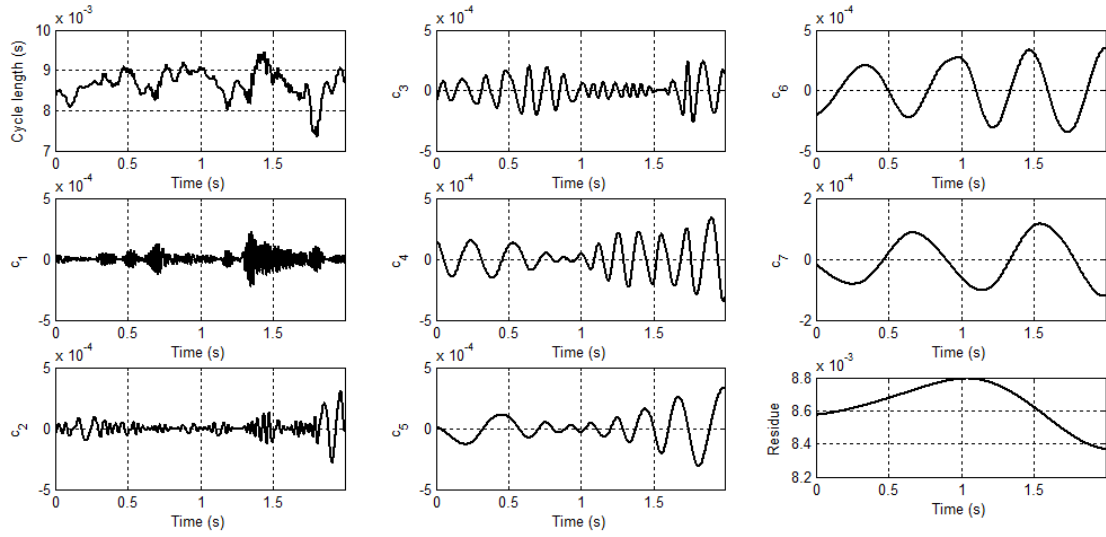


(a) Results of the EMD : cycle length time series (upper left), successive IMFs and residue

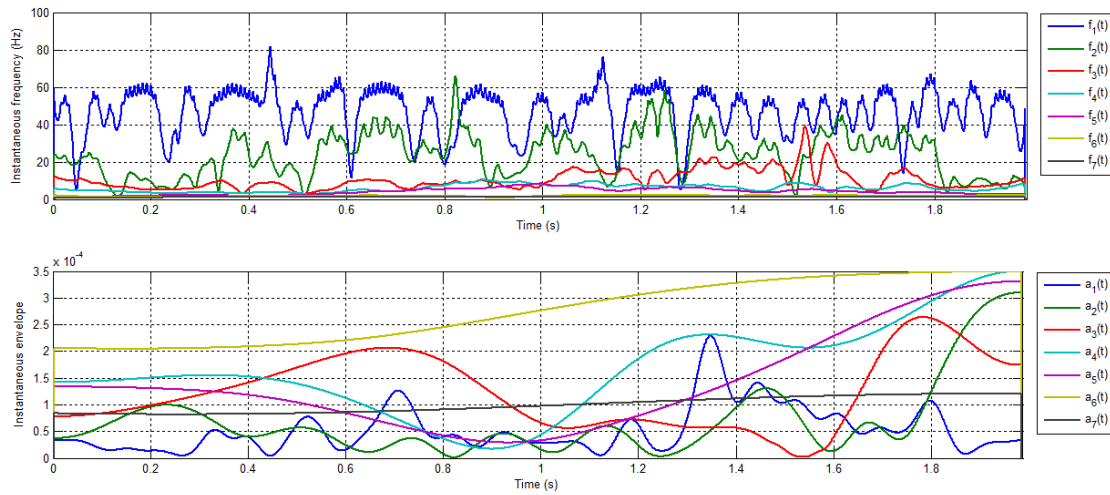


(b) Instantaneous frequency and envelope of the IMFs

Figure 8.1 – Modal voice ($F_0 \approx 88\text{Hz}$) : Empirical mode decomposition and instantaneous values of mode envelopes and frequencies

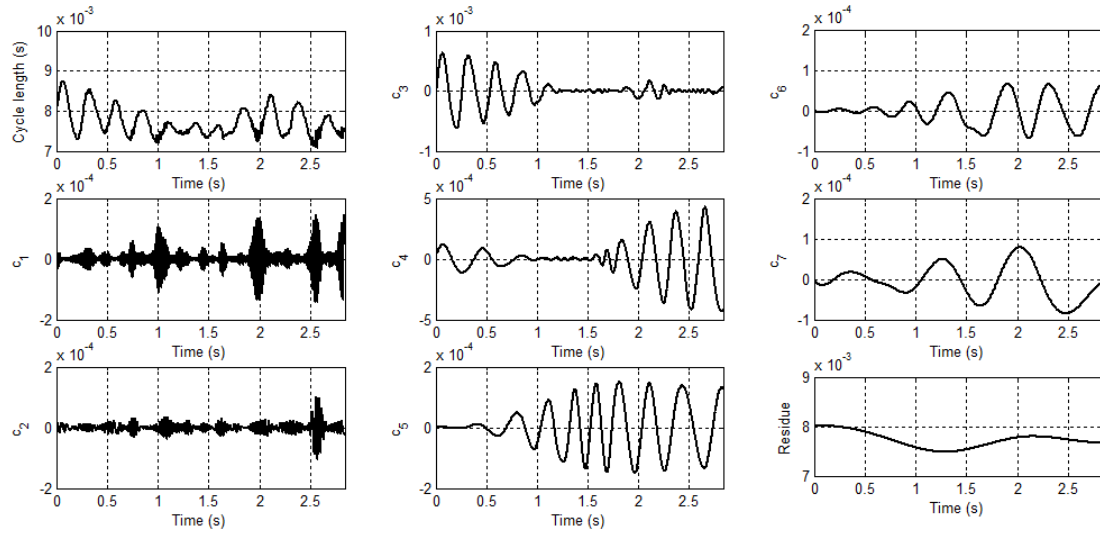


(a) Results of the EMD : cycle length time series (upper left), successive IMFs and residue

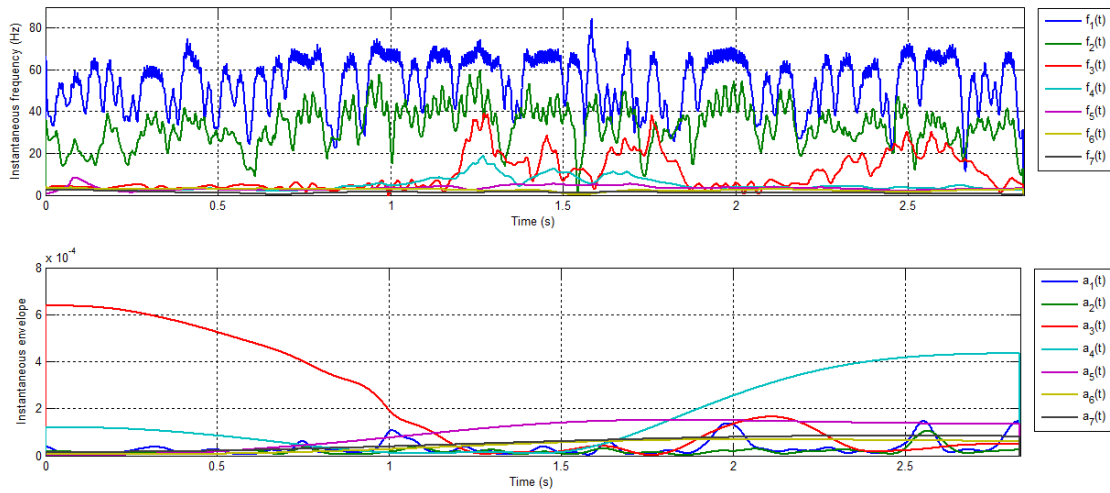


(b) Instantaneous frequency and envelope of the IMFs

Figure 8.2 – Pathological voice (speaker with Parkinson's disease, $F_0 \approx 123\text{Hz}$) : Empirical mode decomposition and instantaneous values of mode envelopes and frequencies



(a) Results of the EMD : cycle length time series (upper left), successive IMFs and residue



(b) Instantaneous frequency and envelope of the IMFs

Figure 8.3 – Pathological voice (speaker with essential tremor, $F_0 \approx 132\text{Hz}$) : Empirical mode decomposition and instantaneous values of mode envelopes and frequencies

8.3 Categorization

The next step consists in grouping and adding IMFs in four categories which are : intonation and declination, physiological tremor, neurological tremor, and vocal jitter. Individual modes $c_i(t)$, $i = 1, \dots, I$, are assigned to one of the four categories on the base of their weighted average instantaneous frequency $f_{i,aver}$ for which the weights are the instantaneous envelopes $a_i(t)$:

$$f_{i,aver} = \frac{\int_0^T f_i(t) a_i(t) dt}{\int_0^T a_i(t) dt} \quad (8.1)$$

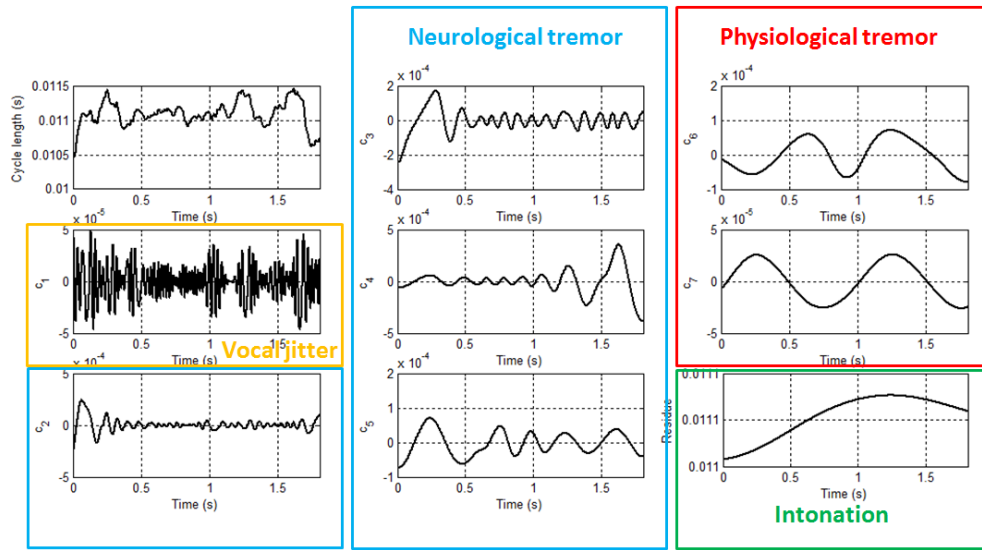
- The residue $r(t)$ of the empirical mode decomposition is assigned to intonation or declination.
- The modes $c_i(t)$, $i = 1, \dots, i_J$, with a average instantaneous frequency $f_{i,aver}$ higher than $15Hz$ are assigned to jitter.
- The modes $c_i(t)$, $i = i_J + 1, \dots, i_N$, with a average instantaneous frequency $f_{i,aver}$ comprised between $2Hz$ and $15Hz$ are assigned to neurological tremor.
- The other modes $c_i(t)$, $i = i_N + 1, \dots, I$, are assigned to physiological tremor.

A lower limit of $2Hz$ enables to decrease the effect of heart beat and breathing, and the upper limit of $15Hz$ includes the frequency interval of the great majority of tremor types [Rub+15] [MM00].

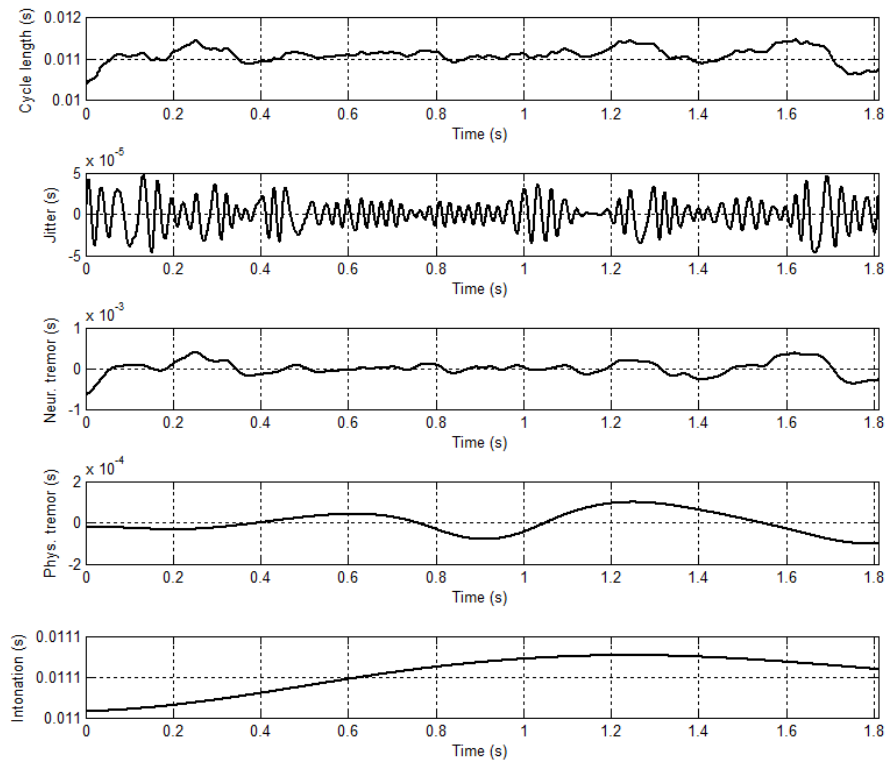
By summing modes into categories, the property of perfect reconstruction, related to EMD, is still valid, and the cycle length time series $x(t)$ may be expressed as follows :

$$\begin{aligned} x(t) &= \sum_{i=1}^I c_i(t) + r(n) \\ &= \left(\sum_{i=1}^{i_J} c_i(t) \right) + \left(\sum_{i=i_J+1}^{i_N} c_i(t) \right) + \left(\sum_{i=i_N+1}^I c_i(t) \right) + r(n) \\ &= x_{jit}(t) + x_{neur}(t) + x_{phys}(t) + x_{int}(t) \end{aligned} \quad (8.2)$$

where $x_{jit}(t)$, $x_{neur}(t)$, $x_{phys}(t)$ and $x_{int}(t)$ designate the time series assigned to the four categories. Figures 8.4, 8.5 and 8.6 illustrate the IMF clustering and the resulting slow and fast cycle length perturbations for a fragment of a vowel [a] sustained by the same 3 speakers.

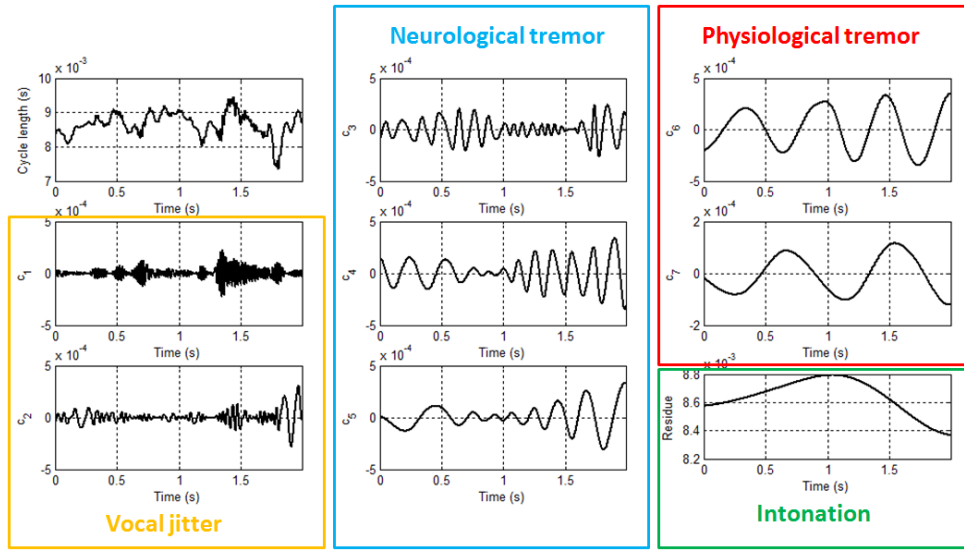


(a) Empirical mode clustering

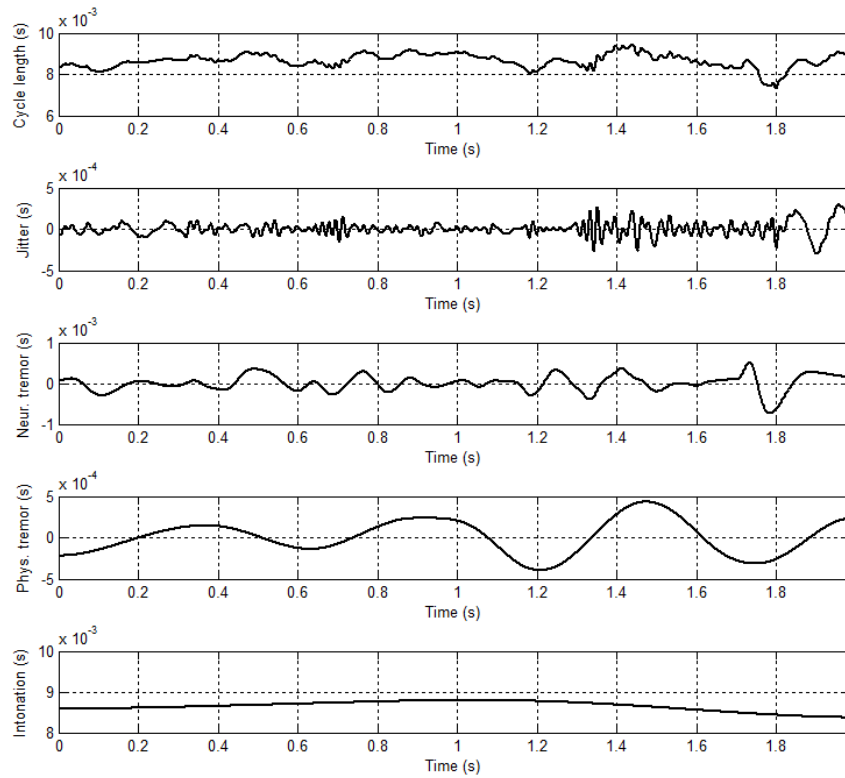


(b) Vocal cycle length perturbation time series

Figure 8.4 – Modal voice : Categorization

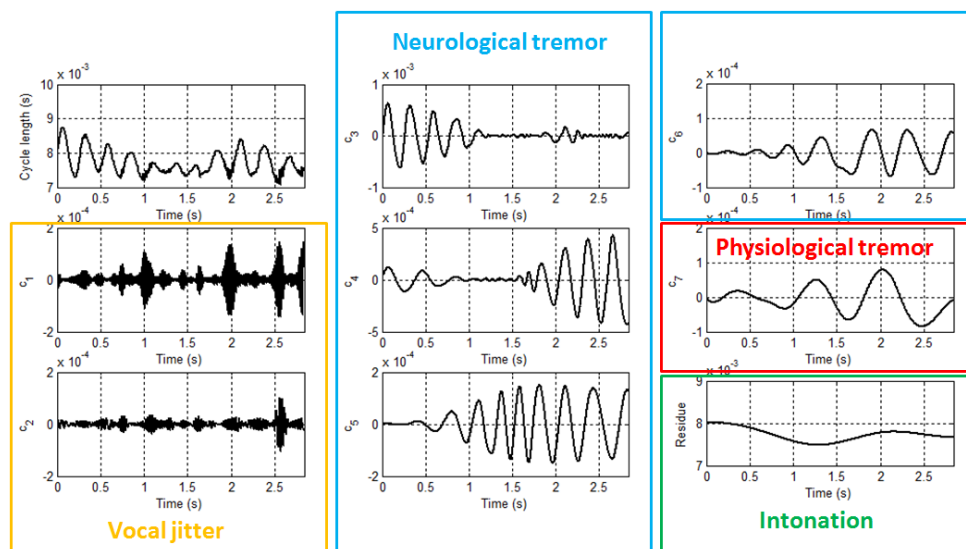


(a) Empirical mode clustering

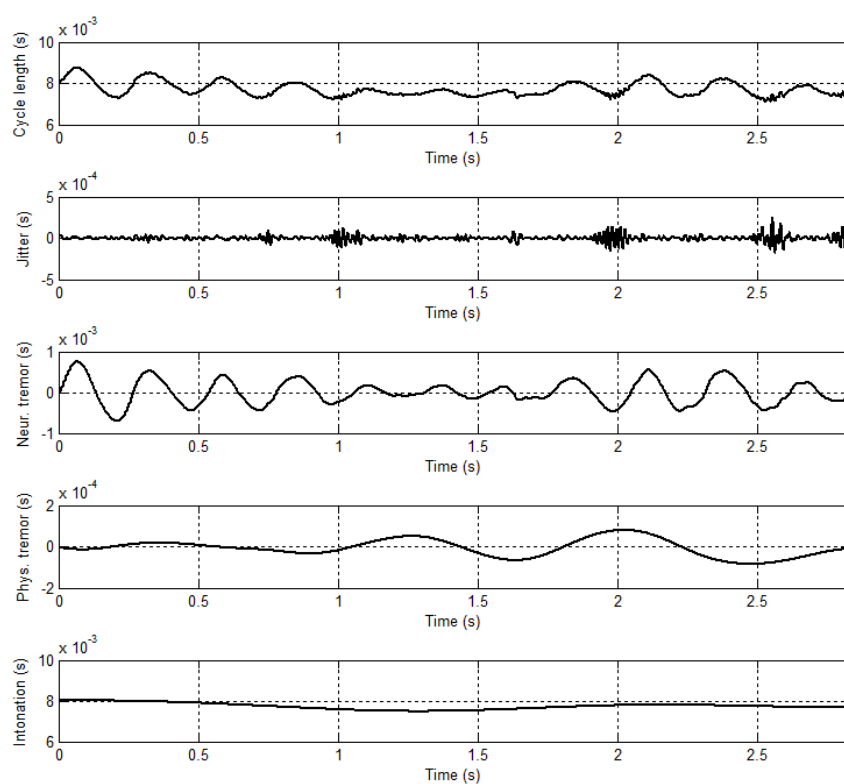


(b) Vocal cycle length perturbation time series

Figure 8.5 – Pathological voice (speaker with Parkinson's disease) : Categorization



(a) Empirical mode clustering



(b) Vocal cycle length perturbation time series

Figure 8.6 – Pathological voice (speaker with essential tremor) : Categorization

8.4 Average vocal frequency and cycle length perturbation size

8.4.1 Average vocal frequency

The average vocal frequency F_0 is computed via the inverse of the average of the intonation time series, T = length of analysis interval.

$$F_0 = \frac{1}{x_{int}(t)} = \frac{1}{\frac{1}{T} \int_0^T x_{int}(t) dt} = \frac{1}{T_0} \quad (8.3)$$

8.4.2 Perturbation sizes

The size of the vocal cycle length perturbations is assessed via the standard deviation of their corresponding time series. Let us introduce the following notations :

σ_{jit} :	Vocal jitter size
σ_{neur} :	Neurological tremor depth
σ_{phys} :	Physiological tremor depth
σ_{tre} :	Total tremor depth
σ_{pert} :	Total perturbation size

Vocal jitter size σ_{jit} , neurological and physiological tremor depths, σ_{neur} and σ_{phys} , are related to the standard deviation of the jitter time series, the physiological and neurological tremor time series, divided by the average T_0 of the intonation time series.

$$\left\{ \begin{array}{l} \sigma_{jit} = \frac{1}{T_0} \sqrt{\frac{1}{T} \int_0^T \left(x_{jit}(t) - \overline{x_{jit}(t)} \right)^2 dt} \\ \sigma_{neur} = \frac{1}{T_0} \sqrt{\frac{1}{T} \int_0^T \left(x_{neur}(t) - \overline{x_{neur}(t)} \right)^2 dt} \\ \sigma_{phys} = \frac{1}{T_0} \sqrt{\frac{1}{T} \int_0^T \left(x_{phys}(t) - \overline{x_{phys}(t)} \right)^2 dt} \end{array} \right. \quad (8.4)$$

Total tremor size σ_{tre} and total perturbation size σ_{pert} are related to the sum of the physiological and neurological tremor time series or the sum of the jitter and tremor time series (divided by T_0).

$$\left\{ \begin{array}{l} \sigma_{tre} = \frac{1}{T_0} \sqrt{\frac{1}{T} \int_0^T \left(x_{phys}(t) + x_{neur}(t) - \overline{x_{phys}(t) + x_{neur}(t)} \right)^2 dt} \\ \sigma_{pert} = \frac{1}{T_0} \sqrt{\frac{1}{T} \int_0^T \left(x_{phys}(t) + x_{neur}(t) + x_{jit}(t) - \overline{x_{phys}(t) + x_{neur}(t) + x_{jit}(t)} \right)^2 dt} \end{array} \right. \quad (8.5)$$

Hereafter are reported the vocal cycle perturbation sizes for the 3 previous speakers. One observes that modal voice is characterized by feeble perturbation sizes. For the two other voices, the size of vocal cycle length perturbations are larger. Speakers with Parkinson's disease or essential tremor have in these examples the same total perturbation sizes but their origins differ. Moreover, one observes that the size of the physiological tremor is significantly larger for the speaker with Parkinson's disease. That may suggest that the vocal tremor of this speaker has globally a larger frequency bandwidth (or that tremor affects breathing).

	$\sigma_{jit}(\%)$	$\sigma_{neur}(\%)$	$\sigma_{phys}(\%)$	$\sigma_{tre}(\%)$	$\sigma_{pert}(\%)$
Modal	0.15	1.56	0.48	1.63	1.63
Parkinson's disease	0.94	2.28	2.37	3.51	3.67
Essential tremor	0.45	3.80	0.53	3.83	3.86

8.5 Neurological tremor frequency

8.5.1 Overview

The neurological tremor frequency is obtained via the instantaneous frequencies, phases and envelopes of the empirical modes in the neurological tremor category. Different neurological tremor frequency cues are proposed. The first tremor frequency cue is obtained via a weighted instantaneous average of the mode frequencies followed by a weighted temporal average. Two other neurological tremor cues are related to the center of gravity of a weighted instantaneous frequency probability density estimated by means of a Gaussian kernel.

8.5.2 Complex neurological tremor time series

Only empirical modes in the neurological tremor category are considered. The neurological tremor time series $x_{neur}(t)$ is given by the sum of these modes $c_i(t)$.

$$x_{neur}(t) = \sum_{i=i_J+1}^{i_N} c_i(t) \quad (8.6)$$

As explained in chapter 6, each time series $c_i(t)$ may describe the trajectory of a point P , attached to the circumference of a wheel in non-harmonic rotating motion. In the complex domain, this trajectory is described on the basis of a time-varying vector, $z_i(t)$, characterized by the instantaneous phase $\phi_i(t)$ and envelope $a_i(t)$ of the mode :

$$c_i(t) = a_i(t) \cos(\phi_i(t)) = \text{Re} \left\{ a_i(t) e^{j\phi_i(t)} \right\} = \text{Re} \{ z_i(t) \} \quad (8.7)$$

As a consequence, relation (8.6) may be rewritten in the complex domain to express the complex neurological tremor time series $z_{neur}(t)$ as the trajectory of concatenated wheels :

$$\begin{aligned}
 x_{neur}(t) &= \sum_{i=i_J+1}^{i_N} a_i(t) \cos(\phi_i(t)) = \sum_{i=i_J+1}^{i_N} \text{Re} \left\{ a_i(t) e^{j\phi_i(t)} \right\} \\
 &= \text{Re} \left\{ \sum_{i=i_J+1}^{i_N} a_i(t) e^{j\phi_i(t)} \right\} = \text{Re} \left\{ \sum_{i=i_J+1}^{i_N} z_i(t) \right\} \\
 \Rightarrow z_{neur}(t) &= \sum_{i=i_J+1}^{i_N} z_i(t) = |z_{neur}(t)| e^{j\phi_{neur}(t)}
 \end{aligned} \quad (8.8)$$

As an example, Figure 8.7 illustrates 4 complex empirical modes ($z_A(t)$, $z_B(t)$, $z_C(t)$ and $z_D(t)$) in the complex plane at instant t as well as their complex sum $z_{neur}(t)$. One observes that complex mode $z_i(t)$ orientations and sizes differ widely so that they contribute differently to the sum $z_{neur}(t)$.

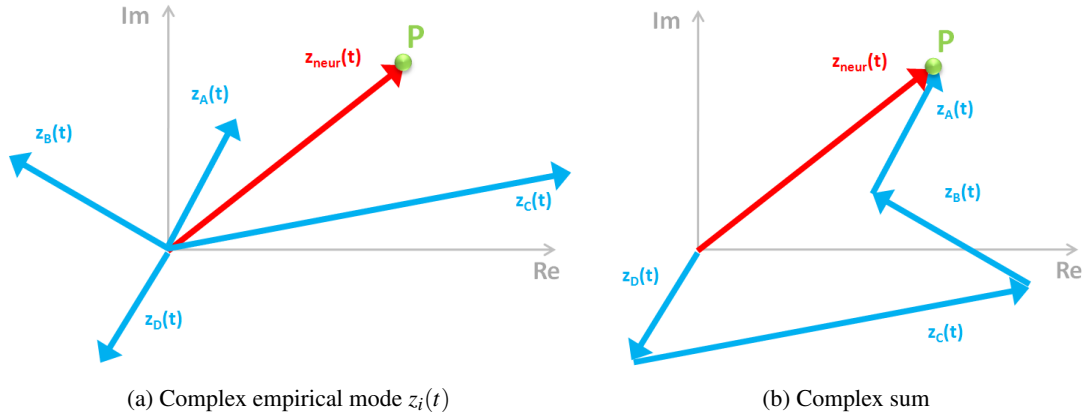


Figure 8.7 – Complex empirical mode sum

8.5.3 Neurological tremor frequency

The complex neurological tremor $z_{neur}(t)$ is given by the sum of complex empirical modes $z_i(t)$. Here, the derivative of $z_{neur}(t)$ is investigated to obtain the neurological tremor frequency $f_{neur}(t)$ as a function of empirical mode instantaneous frequencies $f_i(t)$.

$$\begin{aligned}
 z_{neur}(t) &= |z_{neur}(t)| e^{j\phi_{neur}(t)} = \sum_{i=i_J+1}^{i_N} z_i(t) = \sum_{i=i_J+1}^{i_N} a_i(t) e^{j\phi_i(t)} \\
 \Rightarrow \frac{dz_{neur}(t)}{dt} &= \frac{d|z_{neur}(t)|}{dt} e^{j\phi_{neur}(t)} + |z_{neur}(t)| e^{j\phi_{neur}(t)} j \frac{d\phi_{neur}(t)}{dt} \\
 &= \sum_{i=i_J+1}^{i_N} \frac{da_i(t)}{dt} e^{j\phi_i(t)} + a_i e^{j\phi_i(t)} j \frac{d\phi_i(t)}{dt}
 \end{aligned} \tag{8.9}$$

The derivative of the neurological tremor phase function $\phi_{neur}(t)$ is the neurological tremor frequency $f_{neur}(t)$:

$$\begin{aligned}
 \frac{d\phi_{neur}(t)}{dt} &= \frac{\sum_{i=i_J+1}^{i_N} \left(\frac{da_i(t)}{dt} e^{j\phi_i(t)} + a_i e^{j(\phi_i(t) + \frac{\pi}{2})} \frac{d\phi_i(t)}{dt} \right) - \frac{d|z_{neur}(t)|}{dt} e^{j\phi_{neur}(t)}}{|z_{neur}(t)| e^{j(\phi_{neur}(t) + \frac{\pi}{2})}} \\
 &= \frac{\sum_{i=i_J+1}^{i_N} \left(\frac{da_i(t)}{dt} e^{j(\phi_i(t) - \phi_{neur}(t) - \frac{\pi}{2})} + a_i e^{j(\phi_i(t) - \phi_{neur}(t))} \frac{d\phi_i(t)}{dt} \right) - \frac{d|z_{neur}(t)|}{dt} e^{-j\frac{\pi}{2}}}{|z_{neur}(t)|} \\
 \Rightarrow f_{neur}(t) &= \frac{1}{2\pi} \frac{\sum_{i=i_J+1}^{i_N} \left(\frac{da_i(t)}{dt} e^{j(\phi_i(t) - \phi_{neur}(t) - \frac{\pi}{2})} + 2\pi f_i(t) a_i e^{j(\phi_i(t) - \phi_{neur}(t))} \right) - \frac{d|z_{neur}(t)|}{dt} e^{-j\frac{\pi}{2}}}{|z_{neur}(t)|}
 \end{aligned} \tag{8.10}$$

Notice that the neurological tremor frequency f_{neur} and envelope $|z_{neur}(t)|$ are real scalars so that relation (8.10) may be rewritten as follows :

$$f_{neur}(t) = \frac{1}{2\pi} \frac{\sum_{i=i_J+1}^{i_N} \frac{da_i(t)}{dt} \sin(\phi_i(t) - \phi_{neur}(t))}{|z_{neur}(t)|} + \frac{\sum_{i=i_J+1}^{i_N} f_i(t) a_i \cos(\phi_i(t) - \phi_{neur}(t))}{|z_{neur}(t)|} \tag{8.11}$$

Assuming that the empirical mode envelopes $a_i(t)$ fluctuate more slowly than the carrier, the terms $\frac{da_i(t)}{dt}$, related to the derivation of empirical mode envelopes, may be disregarded.

$$f_{neur}(t) \approx \frac{\sum_{i=i_J+1}^{i_N} f_i(t) a_i \cos(\phi_i(t) - \phi_{neur}(t))}{|z_{neur}(t)|} \quad (8.12)$$

Moreover, by zeroing the neurological tremor phase ϕ_{neur} via the multiplication by $e^{-j\phi_{neur}(t)}$, $z_{neur}(t) e^{-j\phi_{neur}(t)}$ is aligned with the real axis and corresponds to the neurological tremor envelope $|z_{neur}(t)|$. The instantaneous envelope $|z_{neur}(t)|$ may thus be expressed as the sum of terms that take into account the mode envelopes $a_i(t)$, phase $\phi_i(t)$ and the neurological tremor phase $\phi_{neur}(t)$:

$$\begin{aligned} |z_{neur}(t)| &= \text{Re} \{ |z_{neur}(t)| \} = \text{Re} \{ z_{neur}(t) e^{-j\phi_{neur}(t)} \} \\ &= \text{Re} \left\{ \left(\sum_{i=i_J+1}^{i_N} a_i(t) e^{j\phi_i(t)} \right) e^{-j\phi_{neur}(t)} \right\} \\ &= \text{Re} \left\{ \sum_{i=i_J+1}^{i_N} a_i(t) e^{j(\phi_i(t) - \phi_{neur}(t))} \right\} \\ &= \sum_{i=i_J+1}^{i_N} \text{Re} \left\{ a_i(t) e^{j(\phi_i(t) - \phi_{neur}(t))} \right\} = \sum_{i=i_J+1}^{i_N} a_i(t) \cos(\phi_i(t) - \phi_{neur}(t)) \end{aligned} \quad (8.13)$$

One defines $w_i(t)$ as follows :

$$w_i(t) = a_i(t) \cos(\phi_i(t) - \phi_{neur}(t)) \quad (8.14)$$

By associating relations (8.13) and (8.14), relation (8.12) is given by :

$$f_{neur}(t) \approx \frac{\sum_{i=i_J+1}^{i_N} w_i(t) f_i(t)}{\sum_{i=i_J+1}^{i_N} w_i(t)} \quad (8.15)$$

As a consequence, the neurological tremor frequency $f_{neur}(t)$ is given by the weighted sum of instantaneous empirical mode frequencies $f_i(t)$. The weights, $w_i(t)$ are obtained by taking the projections of complex modes $z_i(t)$ on the complex sum $z_{neur}(t)$. Figure 8.8 illustrates the computation of the weights $w_C(t)$ and $w_D(t)$ which are related to empirical modes $c_C(t)$ and $c_D(t)$ at instant t . One observes that empirical modes that are quasi-aligned with the complex mode sum contribute most, positively or negatively. Also, the contribution of complex empirical modes that are in anti-phase is negligible.

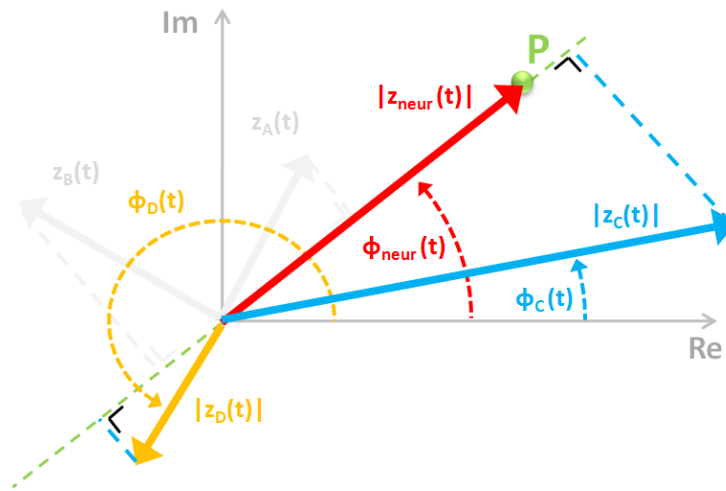


Figure 8.8 – Neurological tremor frequency : Weight computation

8.5.4 Neurological tremor frequency estimate

Previous developments have shown that the neurological tremor frequency $f_{neur}(t)$ may be estimated on the basis of a weighted sum of individual mode frequencies $f_i(t)$. An estimate of the neurological tremor frequency, denoted $\hat{f}_{neur}(t)$, is computed as follows :

$$\hat{f}_{neur}(t) = \frac{\sum_{i=i_J+1}^{i_N} w_i(t) f_i(t)}{\sum_{i=i_J+1}^{i_N} w_i(t)} \quad (8.16)$$

Figure 8.9 illustrates the estimated neurological tremor frequency time series obtained for a fragment of vowel [a] sustained by the 3 previous speakers. One observes that frequency $\hat{f}_{neur}(t)$ can be locally negative.

One reason is the occasional negativity of weights $w_i(t)$, which is the consequence of particular phase relations between the complex sum of the modes and individual modes. Another reason is the negativity of instantaneous mode frequencies $f_i(t)$. The instantaneous frequencies of individual modes are positive given their definition, but negativity may appear because of the iteration involved in the empirical AM-FM decomposition that may turn a mode inflection point into a pair of a local maximum and minimum. A local non-negative minimum or non-positive maximum causes the instantaneous frequency to become locally negative. The estimated neurological tremor frequency time series is therefore smoothed by means of a moving average filter of length $0.2s$.

This smoothed estimated frequency time series is summarized by its temporal average $\hat{f}_{\mu,neur}$ and standard deviation $\delta\hat{f}_{neur}$ weighted sample-by-sample by module $|z_{neur}(t)|$, which is the neurological tremor envelope. The weights are the neurological tremor envelope. Note that these cues refer to the average and variability over time of the neurological tremor frequency. Hereafter are reported the neurological tremor frequency and variability for the same 3 speakers.

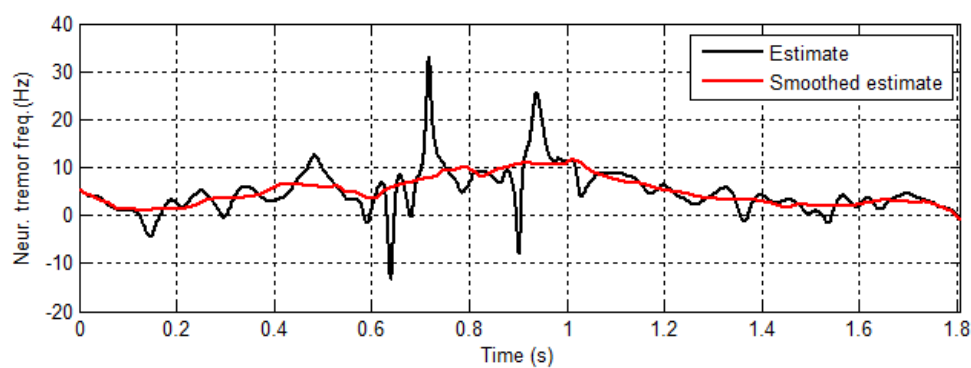
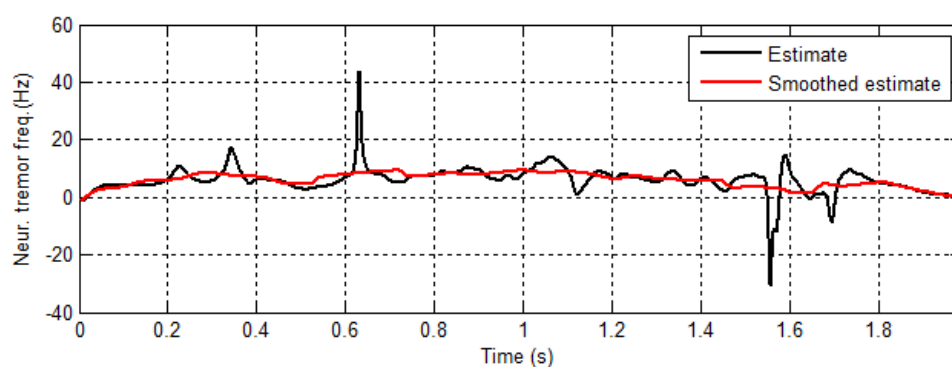
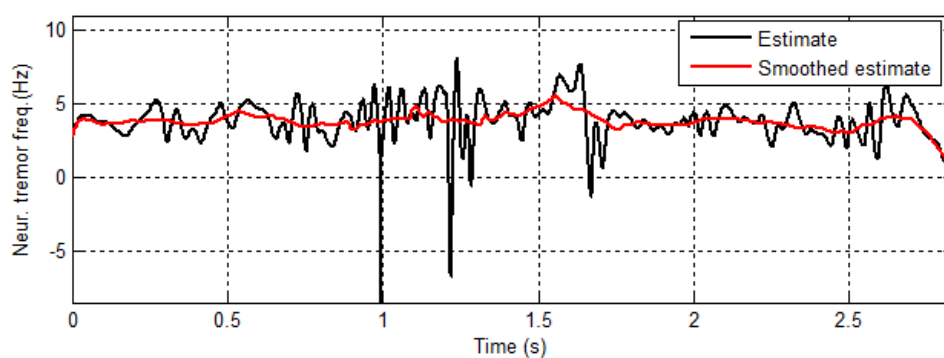
(a) Modal voice ($F_0 \approx 88\text{Hz}$)(b) Pathological voice (speaker with Parkinson's disease, $F_0 \approx 123\text{Hz}$)(c) Pathological voice (speaker with essential tremor, $F_0 \approx 132\text{Hz}$)

Figure 8.9 – Estimation of the neurological tremor frequency

	$\hat{f}_{\mu,neur}(Hz)$	$\delta \hat{f}_{neur}(Hz)$
Modal	3.52	2.18
Parkinson's disease	5.49	2.12
Essential tremor	3.73	0.42

8.5.5 Neurological tremor frequency content analysis

8.5.5.1 Univariate Kernel density estimation

As an alternative, the neurological tremor frequency content is investigated via a weighted instantaneous frequency probability density obtained by means of a Gaussian kernel. In statistics, kernel density estimation [WJ95] is a data smoothing problem where inferences about the population are made, based on a finite data samples that are independently and identically distributed. A weighted density estimation involves weights that are assigned to these variables.

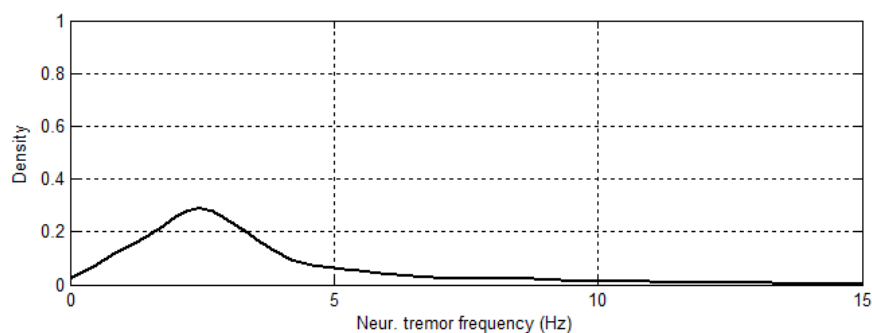
Frequency content analysis is carried out over the whole duration of the speech signal. One considers that the instantaneous frequencies $f_i(t)$, $t = 0 \dots T$ are independent and that a weight $w_i(t)$ is assigned to each. The global density estimate, \hat{F} , is obtained by summing Gaussian kernels, denoted K , that are centred on each frequency $f_i(t)$:

$$\hat{F}(f;h) = \frac{\int_0^T \sum_{i=i_J+1}^{i_N} w_i(t) \frac{1}{h} K\left(\frac{f-f_i(t)}{h}\right) dt}{\int_0^T \sum_{i=i_J+1}^{i_N} w_i(t) dt} \quad (8.17)$$

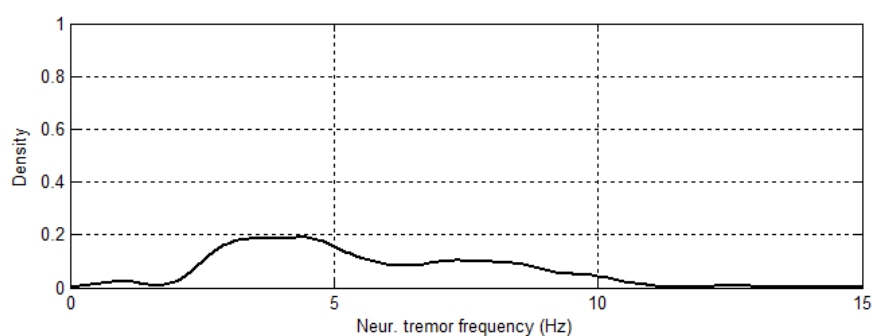
The kernel bandwidth, h , has a strong influence on the density estimate. If h is small, spurious fine structure becomes visible, while when h is too large, all details, spurious or not, are smeared out. Here, h has been chosen on the basis of the optimality criterion, proposed in [SG86], where $\hat{\sigma}$ is the standard deviation of the samples and n is their number :

$$h = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{\frac{1}{5}} \quad (8.18)$$

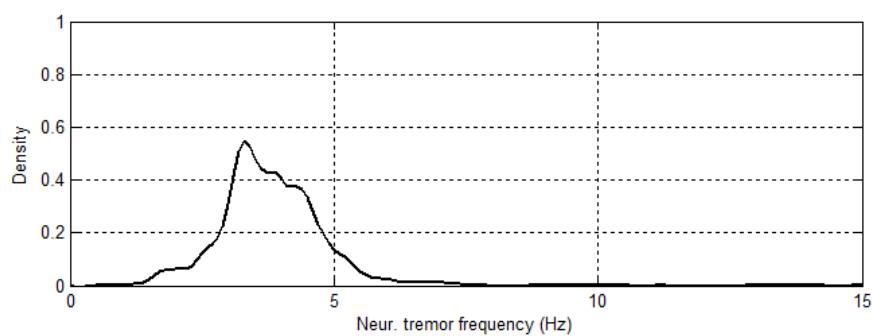
Figure 8.10 illustrates the neurological tremor frequency density estimated for a fragment of vowel [a] sustained by the 3 previous speakers. The shape of the density differs between speakers. The integral over the entire frequency range is equal to one. Therefore, one observes that the speaker with essential tremor is characterized by a narrow-band density and a prominent peak (Figure 8.10c), while the density of the the patient with Parkinson's disease suggests that the neurological tremor frequency is wide-band.



(a) Modal voice ($F_0 \approx 88\text{Hz}$)



(b) Pathological voice (speaker with Parkinson's disease, $F_0 \approx 123\text{Hz}$)



(c) Pathological voice (speaker with essential tremor, $F_0 \approx 132\text{Hz}$)

Figure 8.10 – Neurological tremor frequency density estimation

8.5.5.2 Center of gravity

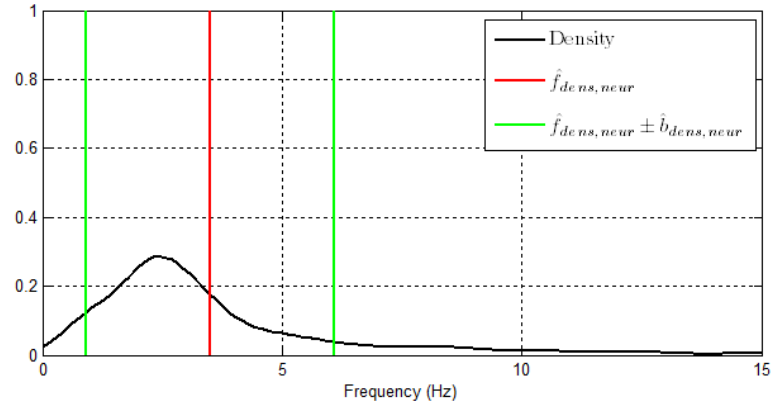
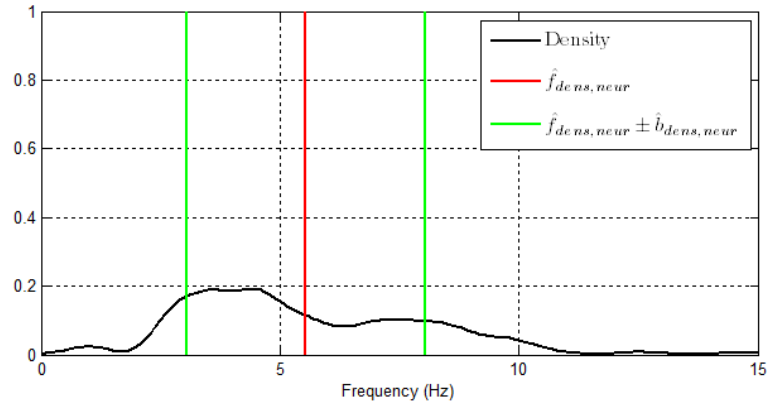
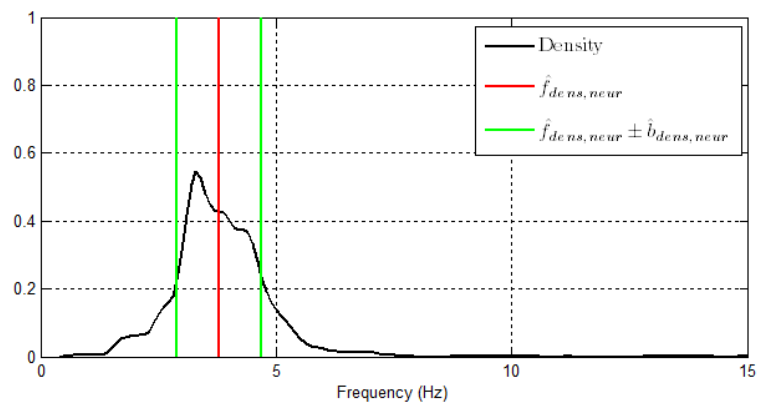
Here, the estimated neurological tremor frequency density is summarized by means of the abscissa of the center of gravity $\hat{f}_{dens,neur}$ and its standard deviation $\hat{b}_{dens,neur}$ in the frequency interval $[0, 15Hz]$.

$$\hat{f}_{dens,neur} = \frac{\int_0^{15Hz} \hat{F}(f;h) f df}{\int_0^{15Hz} \hat{F}(f;h) df} \quad (8.19)$$

$$\hat{b}_{dens,neur} = \sqrt{\frac{\int_0^{15Hz} (f - \hat{f}_{dens,neur})^2 \hat{F}(f;h) df}{\int_0^{15Hz} \hat{F}(f;h) df}} \quad (8.20)$$

As an example, Figure 8.11 illustrates the center of gravity as well as the standard deviation for the neurological tremor frequency densities obtained for the same 3 speakers. Hereafter are reported neurological frequency $\hat{f}_{dens,neur}$ and its associated bandwidth $\hat{b}_{dens,neur}$:

	$\hat{f}_{dens,neur}(Hz)$	$\hat{b}_{dens,neur}(Hz)$
Modal	3.50	2.58
Parkinson's disease	5.54	2.48
Essential tremor	3.78	0.89

(a) Modal voice ($F_0 \approx 88\text{Hz}$)(b) Pathological voice (speaker with Parkinson's disease, $F_0 \approx 123\text{Hz}$)(c) Pathological voice (speaker with essential tremor, $F_0 \approx 132\text{Hz}$)Figure 8.11 – Center of gravity $\hat{f}_{dens,neur}$ and standard deviation $\hat{b}_{dens,neur}$ of the neurological tremor frequency density

8.5.5.3 Scalar quantization

Here, the neurological tremor frequency density is summarized by means of a prominent central frequency and a bandwidth obtained via scalar quantization [Say12]. For that, the density in the range $[0, 15Hz]$ is subdivided into L arbitrary frequency intervals. The lower and upper interval boundaries are denoted $b_{l,j}(k)$ and $b_{u,j}(k)$ where $k = 1 \dots L$ and j is the iteration index. In each interval k , the positions of the center of gravity $f_{G,j}(k)$ are computed.

$$f_{G,j}(k) = \frac{\int_{b_{l,j}(k)}^{b_{u,j}(k)} \hat{F}(f;h) f df}{\int_{b_{l,j}(k)}^{b_{u,j}(k)} \hat{F}(f;h) df} \quad (8.21)$$

The global distortion D_j is computed by summing distortions $d_j(k)$ obtained for all frequency intervals :

$$D_j = \sum_{k=1}^L d_j(k) = \sum_{k=1}^L \int_{b_{l,j}(k)}^{b_{u,j}(k)} \left(f - f_{G,j}(k) \right)^2 \hat{F}(f;h) df \quad (8.22)$$

L new frequency intervals are then determined. The new interval boundaries are the averages of consecutive frequencies $f_{G,j}$.

$$\begin{cases} b_{l,j+1}(k) &= \frac{f_{G,j}(k-1) + f_{G,j}(k)}{2}; \quad k = 2 \dots L \\ b_{u,j+1}(k) &= b_{l,j+1}(k+1); \quad k = 1 \dots L-1 \end{cases} \quad (8.23)$$

The leftmost and rightmost boundaries are fixed to $0Hz$ and $15Hz$.

$$\begin{cases} b_{l,j+1}(1) &= 0Hz \\ b_{u,j+1}(L) &= 15Hz \end{cases} \quad (8.24)$$

This procedure is iterated several times so that at each iteration, the global distortion decreases. As a consequence, the iterative process converges to a local minimum, which depends on the initial frequency interval subdivision. In this study, the number of iterations has been fixed to $J = 30$ and the number of intervals L has been chosen equal to 15. Moreover, the initial frequency intervals uniformly subdivide the frequency range.

Finally, the density is summarized via L local centers of gravity $f_{G,J}(k)$, which are not uniformly distributed on the frequency axis. Indeed, the width of intervals in the vicinity of prominent peaks is smaller than elsewhere. The weight (or mass) of the corresponding interval is :

$$\hat{F}_{G,J}(k) = \int_{b_{l,J}(k)}^{b_{u,J}(k)} \hat{F}(f;h) df \quad (8.25)$$

The prominent neurological tremor frequency is obtained by selecting a subset of intervals $k = k_1 \dots k_2$ for which the mass is larger than the average density. Adjacent intervals are grouped together and a prominent neurological tremor frequency is computed via the center of gravity of this subset.

$$\hat{f}_{quant,neur} = \frac{\sum_{k=k_1}^{k_2} f_{G,J}(k) \hat{F}_{G,J}(k)}{\sum_{k=k_1}^{k_2} \hat{F}_{G,J}(k)} \quad (8.26)$$

If more than one relevant group of adjacent intervals is detected, only the group with the highest local average density is considered.

The neurological tremor bandwidth $\hat{b}_{quant,neur}$ is defined as the width of the retained group of adjacent intervals.

$$\hat{b}_{quant,neur} = b_{u,J}(k_2) - b_{l,J}(k_1) \quad (8.27)$$

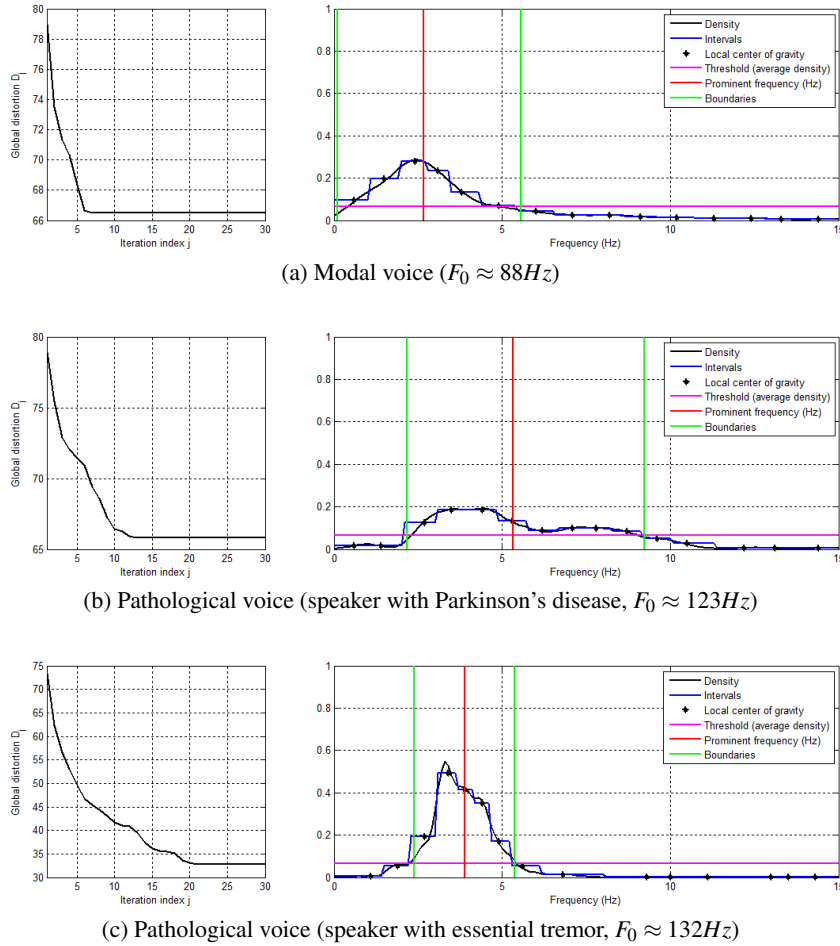


Figure 8.12 – Scalar quantization of the neurological tremor frequency density

As an example, Figure 8.12 illustrates the application of scalar quantization to the neurological tremor frequency densities obtained for the 3 previous speakers. Hereafter are reported the prominent neurological frequencies and their associated bandwidths :

	$\hat{f}_{quant,neur}(Hz)$	$\hat{b}_{quant,neur}(Hz)$
Modal	2.65	5.50
Parkinson's disease	5.32	7.10
Essential tremor	3.89	3.00

8.5.5.4 The time-frequency analysis of the neurological tremor frequency

Neurological tremor frequency density estimation may be frame-based to enable the time-varying analysis of the neurological tremor frequency content. Here, a sliding analysis window of length 0.1s is used. This window is placed at the beginning of the neurological tremor time series and moves to the right with overlap. For each position of the sliding window, several empirical mode instantaneous frequencies $f_i(t)$ and envelopes $a_i(t)$ are available. A weighted probability density estimate of the frequency content of that interval is estimated by means of a Gaussian kernel, as previously. The final time-frequency representation is obtained via the concatenation of the density estimates. The trajectories of the estimated neurological tremor frequency and its bandwidth are tracked via scalar quantization. Figure 8.13 illustrates these time-frequency representations for the same 3 speakers. The time-frequency representations have been coded according to a grayscale level where black and white refer respectively to the the lowest and highest density values. The trajectories of the neurological tremor frequency (in red) as well as the boundaries (in green) are also represented. Visual inspection of these representations suggests that the neurological tremor modulation frequency is time-varying.

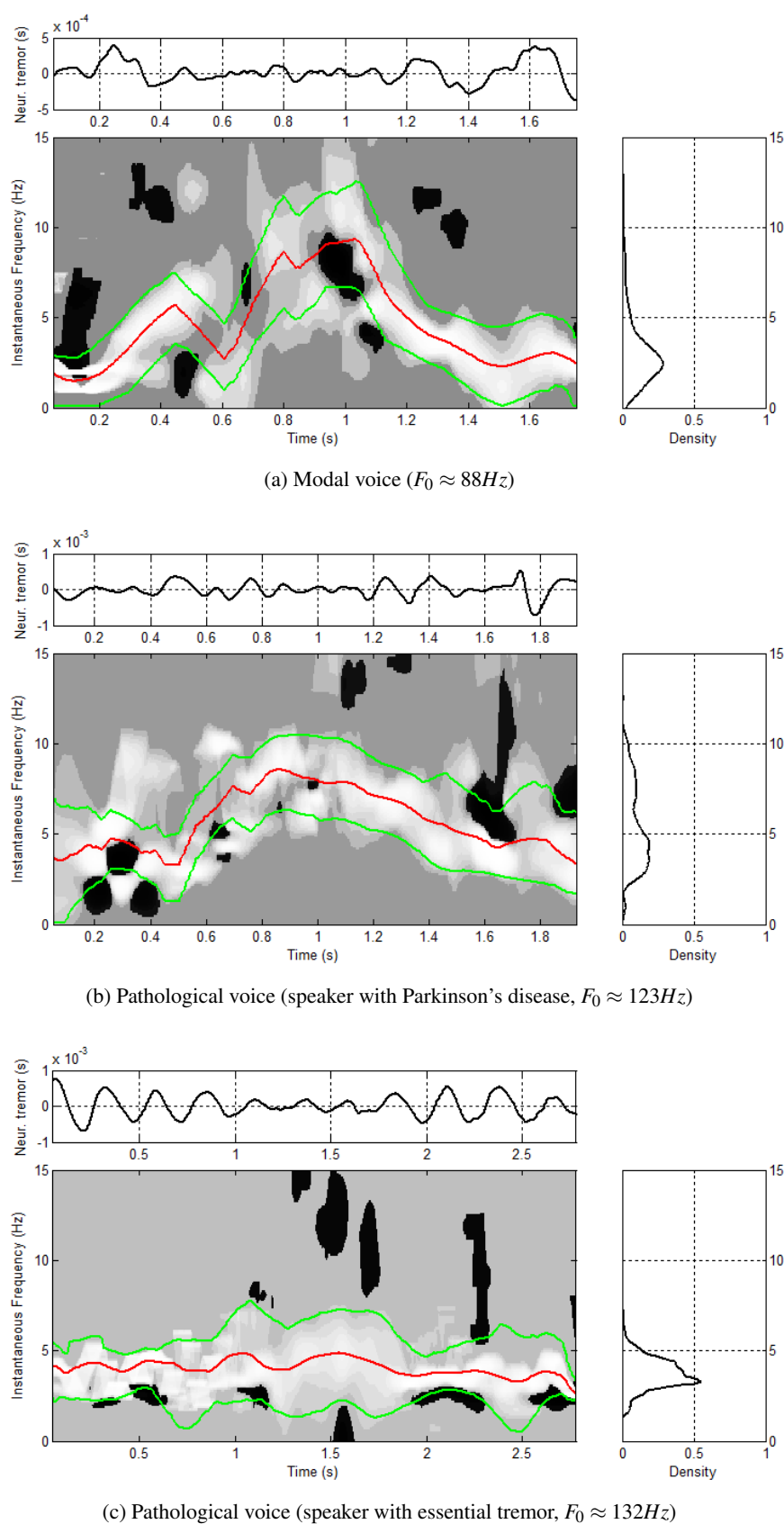


Figure 8.13 – Time-varying representation of the prominent neurological tremor frequency and bandwidth obtained via scalar quantization

8.6 On the choice of the weights

Here, the choice of the weight formulation used to compute the neurological tremor frequency estimates and/or its frequency densities is discussed and validated. As a reminder, the weight formulation has been obtained from the theoretical development expressing the neurological tremor frequency as a function of the weighted individual instantaneous mode frequencies (see section 8.5.3), and the weights were defined as the projections of complex modes $z_i(t)$ on the complex sum $z_{neur}(t)$ (see relation (8.14)).

Two signals have been considered. The first, $x_1(t)$, is a sinusoidal time series oscillating at 7Hz. The second, $x_2(t)$, is obtained by perturbing locally the sinusoidal time series. Figure 8.14 illustrates these two time series.

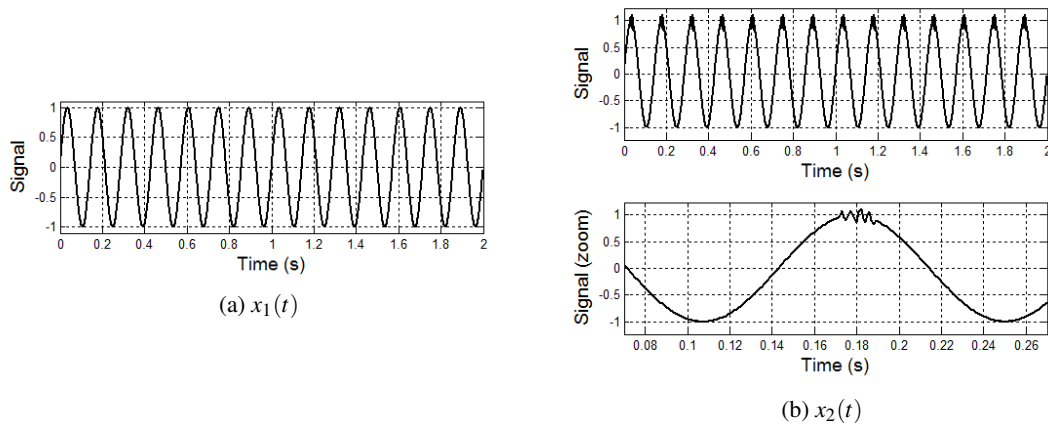


Figure 8.14 – Time series $x_1(t)$ and $x_2(t)$

$$\begin{aligned}
 x_1(t) &= \sin(2\pi f_0 t) & t \in [0, 2s] \\
 x_2(t) &= x_1(t) + 10 \sum_n \delta(t - \frac{n}{f_0}) \otimes p(t)
 \end{aligned} \tag{8.28}$$

where \otimes designates the convolution, $\delta(t)$ is an impulse and $p(t)$ is a perturbation defined as follows :

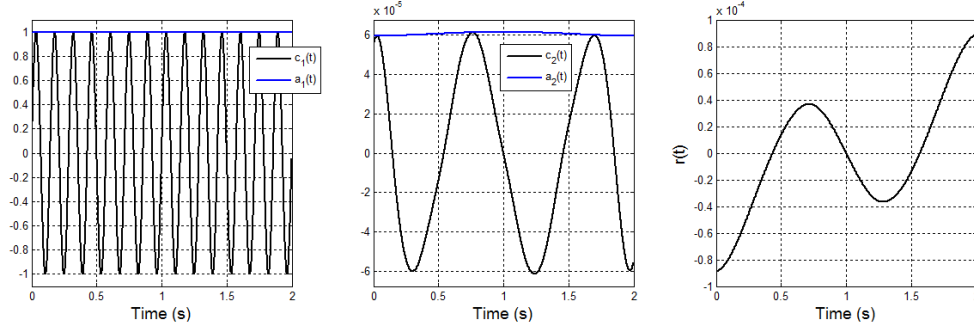
$$p(t) = \begin{cases} f_0 t - (0.2 + 0.03m) & t \in \left[\frac{0.2 + 0.03m}{f_0}, \frac{0.215 + 0.03m}{f_0} \right] \\ -f_0 t + (0.215 + 0.03m) & t \in \left[\frac{0.215 + 0.03m}{f_0}, \frac{0.23 + 0.03m}{f_0} \right] \end{cases} \tag{8.29}$$

$$m = 0, 1, 2, 3$$

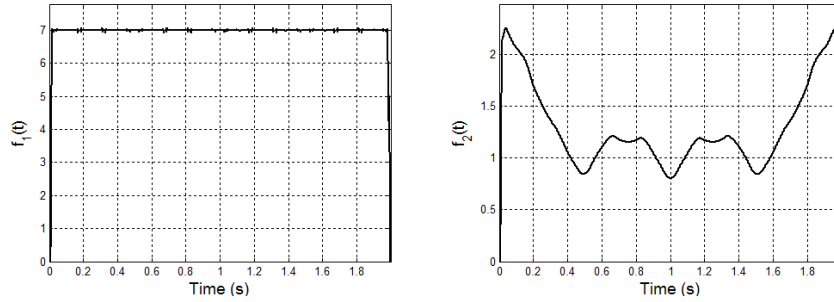
For each signal, the empirical mode decomposition as well as the extraction of instantaneous values via AM-FM decomposition have been carried out. In this section, no categorization is carried out. As a consequence, all empirical modes (except the residue) are considered for further processing.

Figure 8.15 shows the extracted empirical modes $c_i(t)$ as well as the instantaneous mode frequencies $f_i(t)$ and envelopes $a_i(t)$ of $x_1(t)$. One observes that the sinusoidal time series, that

satisfies theoretically the mode properties, has been decomposed into 2 empirical modes and a residue. However, the amplitudes of the second mode and the residue are very feeble compared to the first empirical mode.



(a) Empirical modes $c_i(t)$ and instantaneous mode envelopes $a_i(t)$



(b) Instantaneous mode frequencies $f_i(t)$

Figure 8.15 – EMD & AM-FM decomposition applied to $x_1(t)$

Three instantaneous frequency probability densities have been estimated. These three densities, illustrated in Figure 8.16 differ by the weights assigned to each instantaneous frequency. In Figure 8.16a, all instantaneous frequencies $f_i(t)$ are equally weighted. In Figure 8.16b, the weights are the instantaneous envelopes $a_i(t)$. Finally, in Figure 8.16c, the proposed projection-based weights are considered. Additionally, the frequency probability densities have been estimated for two kernel width values h : the first small and arbitrary chosen kernel width $h_1 = 0.1$ enables the visualization of the fine structure of the distribution, and the second kernel width $h_2 = 0.63$ has been determined by means of the optimality criterion defined in relation (8.18).

Visual inspection shows that the frequency densities obtained on the basis of the instantaneous mode envelopes $a_i(t)$ or the projections of the complex modes $z_i(t)$ on the complex sum $z_{resultant}$ gives identical distributions. On the other hand, the density estimated by means of equally weighted instantaneous frequencies may be discarded because additional frequency components that are related to small and spurious extracted modes are reported.

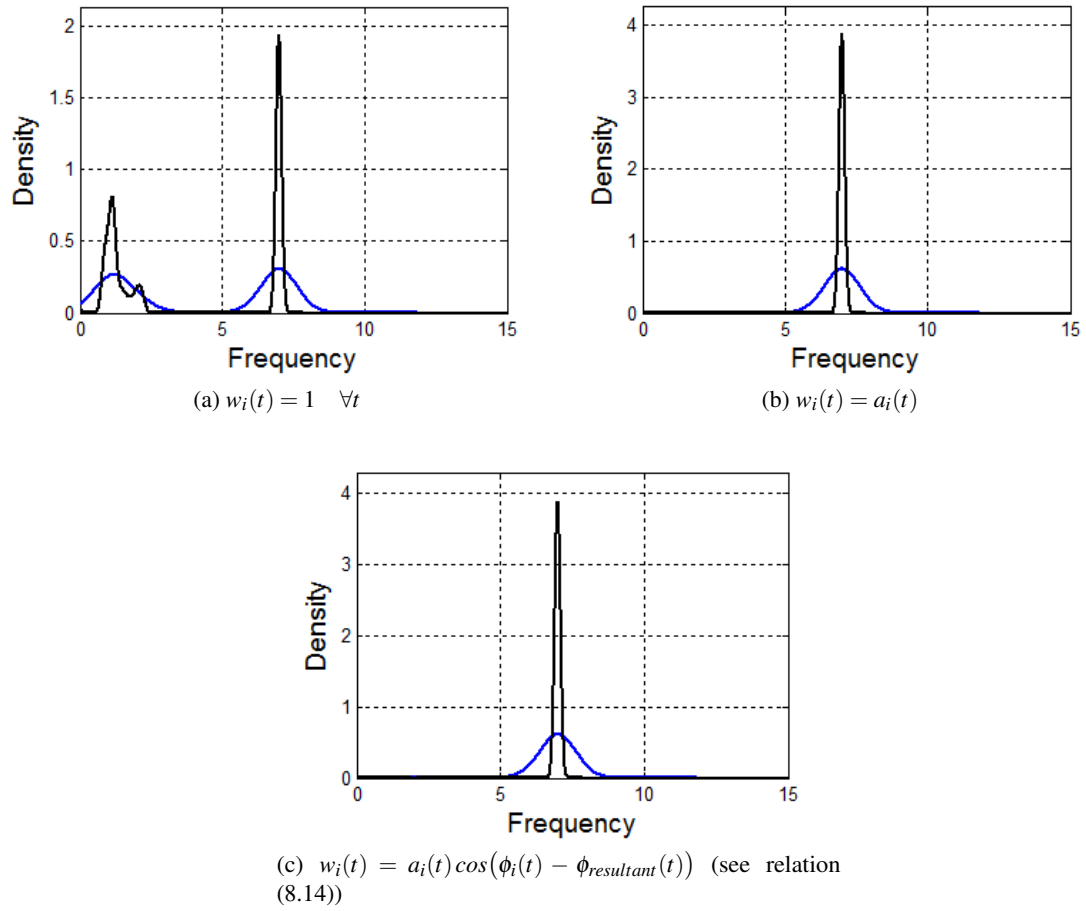


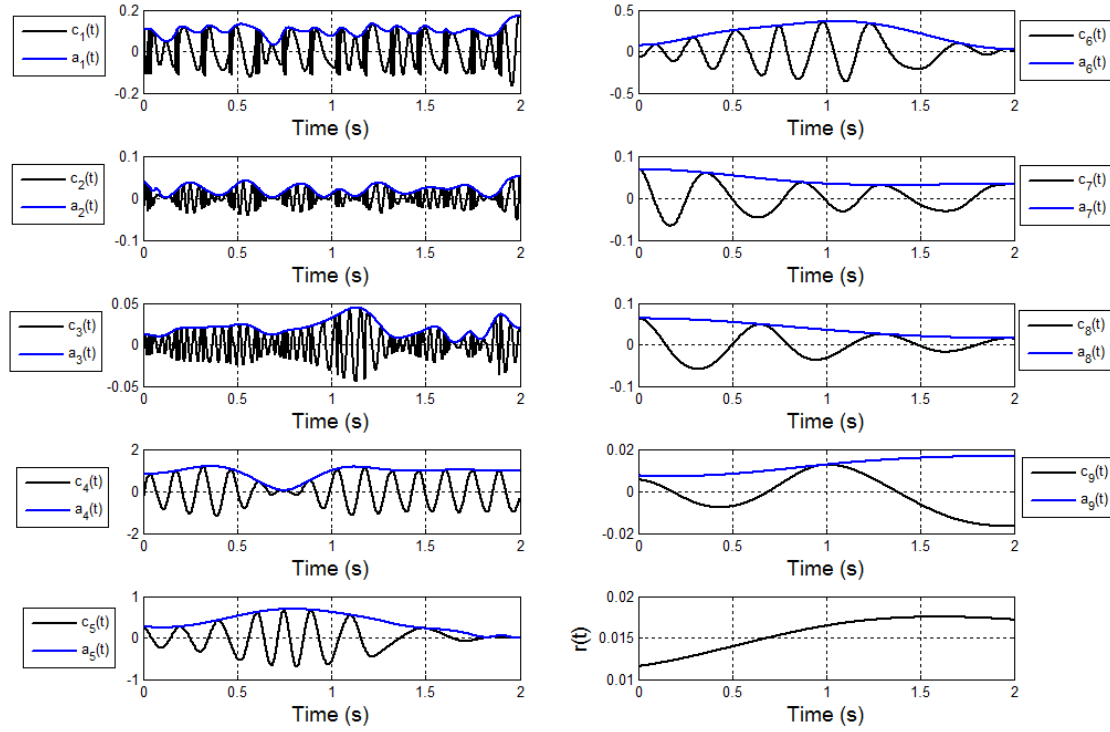
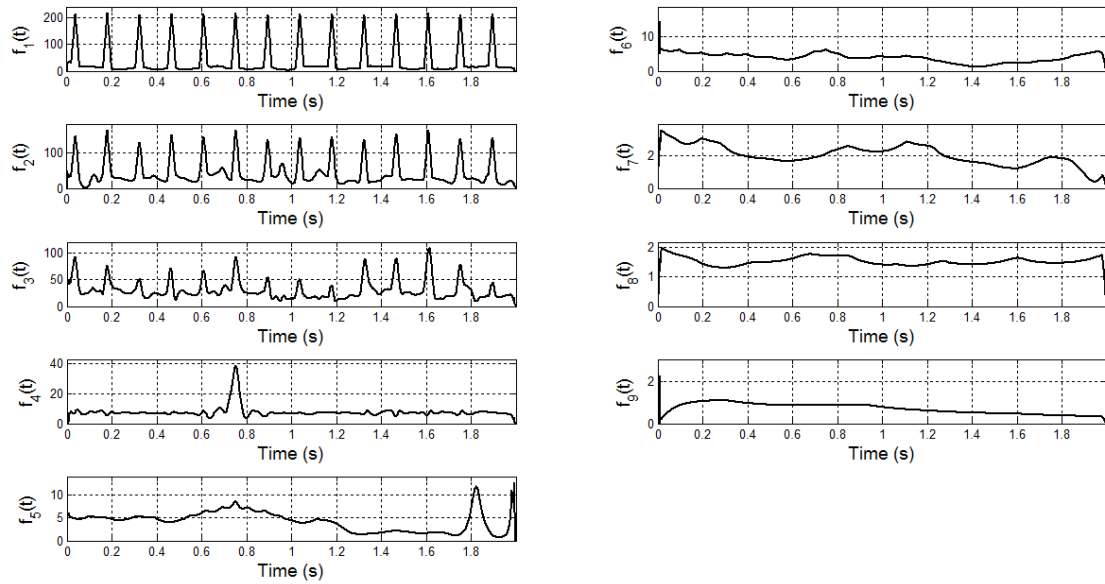
Figure 8.16 – Instantaneous frequency probability density estimates of $x_1(t)$ obtained for three different weight formulations, and two kernel widths $h_1 = 0.1$ (black) and $h_2 = 0.63$ (blue).

Figure 8.17 illustrates the results of the empirical mode decomposition and the AM-FM decomposition applied to the second time series $x_2(t)$. Nine empirical modes have been extracted. One observes that the decomposition is affected by mode mixing. More exactly, the main effects of the mode mixing are observed for the 4th and 5th empirical modes in which an oscillatory component at $\approx 7Hz$ is locally present. Also, careful analysis of Figure 8.17b reveals that large frequency components appear locally in two or several adjacent modes that together do not contribute much to the original time series $x_2(t)$ because they are out of phase. For instance, a $\approx 1.6Hz$ component is shared by empirical mode $c_7(t)$ and $c_8(t)$ in the interval $[0.5s, 0.7s]$.

Figure 8.18 illustrates two weighted frequency probability density estimates obtained by means of envelope-based or projection-based weighting of the instantaneous frequencies $f_i(t)$, and for two kernel widths $h_1 = 0.1$ (arbitrarily chosen) and $h_2 = 0.75$ (optimality criterion).

At present, the instantaneous frequency probability density estimates differ as opposed to the previous example. Indeed, one observes that the projection-based weighted probability density estimate (Figure 8.18b) reports only one prominent peak centred around $7Hz$. On the other hand, the envelope-based weighted probability density estimate (Figure 8.18a) is more dispersed and the presence of additional spurious low-frequency components is clearly observed.

To sum up, these two experiments report that a weighting of the instantaneous mode frequency is required for the estimation of the neurological tremor frequency and/or its frequency probability density. Moreover, the choice of the weight formulation based on complex mode projection is desirable so that only in-phase frequency components contribute to the final estimate.

(a) Empirical modes $c_i(t)$ and instantaneous mode envelopes $a_i(t)$ (b) Instantaneous mode frequencies $f_i(t)$ Figure 8.17 – EMD & AM-FM decomposition applied to $x_2(t)$

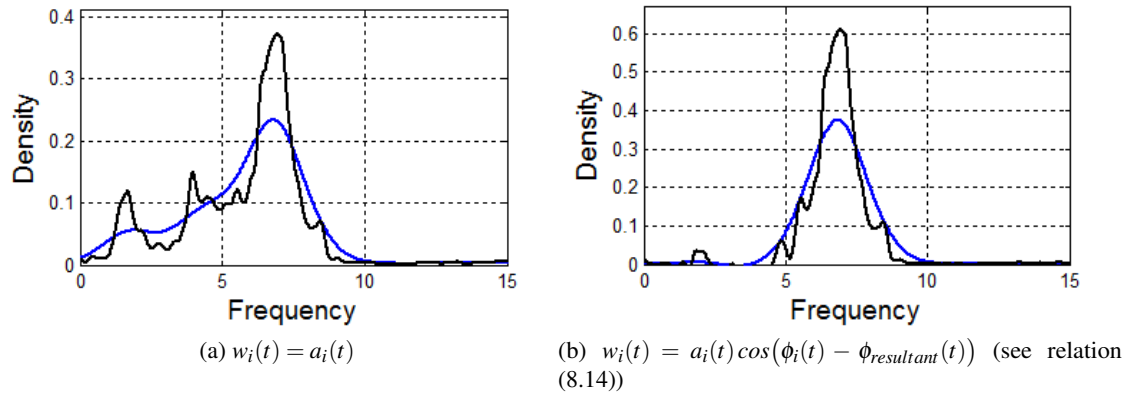


Figure 8.18 – Instantaneous frequency probability density estimates of $x_2(t)$ obtained for two different weight formulations, and two kernel widths $h_1 = 0.1$ (black) and $h_2 = 0.75$ (blue).

8.7 Conclusions

In this chapter, the cycle length time series has been decomposed into sub time series that are respectively assigned to vocal jitter, neurological tremor, physiological tremor and a residual trend, which is due to intonation and declination. The break-up is based on empirical modes that are assigned to one of the four categories and the cycle length sub time series are obtained by summing the modes assigned to a same category.

The sizes of vocal jitter and of physiological and neurological tremor depths have been estimated via the standard deviations of the corresponding sub time series divided by the average cycle length. Ideally, the tremor frequencies could be estimated via weighted averages of the instantaneous frequencies of the empirical modes that belong to the neurological or physiological tremor categories. However, one observes that due to mode mixing large frequency components may appear in two adjacent modes that together do not contribute much to the corresponding sub time series because they are out of phase. Computing arithmetic averages of mode frequencies would therefore assign undue weight to frequencies that exist in individual modes, but which do not contribute to the corresponding tremor category. The solution that has been explored here consists in estimating the instantaneous phase and amplitude of each empirical mode via empirical AM-FM decomposition. These then enable a complex mode to be assigned to each empirical mode and the weighted average of the instantaneous mode frequencies is defined in the complex plane so that only in-phase frequency components contribute to the final estimate. Three neurological tremor frequency and bandwidth cues have been proposed.

The proposed vocal cycle perturbation size, frequency and bandwidth cues will be validated in Chapter 9.

Key points

- The cycle length time series has been decomposed into temporal sub-series that are respectively assigned to vocal jitter, neurological tremor, physiological tremor and a residual trend, which is due to intonation and declination
- The sizes of vocal jitter and of physiological and neurological tremor depths have been estimated via the standard deviations of the corresponding temporal sub-series divided by the average cycle length
- The neurological tremor frequency and bandwidth cues have been obtained via a weighted instantaneous average of the mode frequencies and/or summarized via a weighted average instantaneous frequency probability density, involving complex empirical modes



Validation and Results

9	Validation	159
9.1	Introduction	
9.2	The synthesizer of disordered voices	
9.3	Tracking of cycle lengths	
9.4	Perturbation analysis	
9.5	Conclusions	
10	Parkinson and control speakers	185
10.1	Introduction	
10.2	Corpora	
10.3	Vocal cues	
10.4	Analysis of the corpus from Bochum University Clinic	
10.5	Analysis of the corpus from Pays d'Aix Hospital	
10.6	Comparison between corpora	
10.7	Discussion and conclusion	
11	Conclusions & perspectives	213
11.1	Objectives and motivations	
11.2	Key results	
11.3	Improvements & perspectives	
	Bibliography	219
	Index	225



9. Validation

Objectives of this chapter

- Validate *Saliency-based Cycle Length Tracking*
- Compare the performance of SCLT with a well-known frame-based tracking method
- Validate vocal cycle length perturbation analysis

Contents

9.1	Introduction	161
9.2	The synthesizer of disordered voices	161
9.2.1	Glottal source phase function perturbation model	161
9.2.2	Glottal airflow model	163
9.2.3	Propagation through the vocal tract	164
9.3	Tracking of cycle lengths	164
9.3.1	Overview	164
9.3.2	Corpus	164
9.3.3	Influence of the perturbation sizes	165
9.3.4	Influence of background noise	168
9.3.5	Comparison with Praat Software	172
9.4	Perturbation analysis	177
9.4.1	Corpus	177
9.4.2	Method	177
9.4.3	Results	178
9.5	Conclusions	183

9.1 Introduction

The extraction of vocal cycle lengths via salience analysis and dynamic programming as well as the analysis of cycle length perturbations have been validated by means of simulations. The simulated stimuli have been generated via a synthesizer of disordered voices.

Notice that a clear difference is made between parameters and cues. A *parameter* is related to the synthesizer and designates the action of a user on the characteristics of the generated synthetic stimulus. On the other hand, the term *cue* refers to the a posteriori computation of a quantity.

This chapter is organized as follows. In the first part, the synthesizer of disordered voices is presented. In the second part, the cycle length tracking is validated by means of synthetic voiced speech sounds and the robustness with regard to background noise is assessed. Then, the perturbation method analysis is validated by means of synthetic cycle length sequences and assessed via multiple linear regression models.

9.2 The synthesizer of disordered voices

Synthetic stimuli that are used for validation have been generated by a synthesizer of disordered voices. The latter produces sounds that mimic the vocal quality of speakers suffering from a voice or speech pathology or dysfunction. This synthesizer is composed of three major parts that are respectively :

- the glottal source phase function perturbation model. The glottal source phase function, denoted ϕ_0 , is build and perturbed involving the average vocal frequency, its trend (intonation/declination) and the size of vocal perturbations (jitter, neurological tremor, physiological tremor).
- the glottal airflow model consists in simulating the vocal chord vibration dynamics to determine the glottal airflow rate at the entrance of the vocal tract.
- the waveform propagation through the vocal tract is then simulated via formant-based filtering.

9.2.1 Glottal source phase function perturbation model

Figure 9.1 illustrates the scheme of the glottal phase perturbation. The time series that are related to vocal jitter, neurological tremor, physiological tremor are computed individually via white noise filtering. The average vocal frequency as well as its trend are also computed. The perturbation and trend time series are then grouped to build the glottal source phase function $\phi_0(t)$.

9.2.1.1 Intonation and declination

The average vocal frequency and its trend (intonation/declination) are specified via constant or slowly varying time series.

9.2.1.2 Physiological and neurological tremor

Physiological and neurological tremor perturbation time series are generated via white noise filtering by means of a second-order resonator with a gain equal to 1. The parameters are the tremor frequency and the tremor bandwidth. Figures 9.2a and 9.2b illustrate several frequency responses (in modulus) obtained for different tremor frequencies and bandwidths.

Tremor gain is controlled by means of the tremor size parameter which plays the role of a constant gain multiplier. Notice that, via this procedure, the size of the tremor perturbation is influenced by the tremor size parameter but also by the tremor bandwidth. The influence of the tremor frequency on tremor size is feeble.

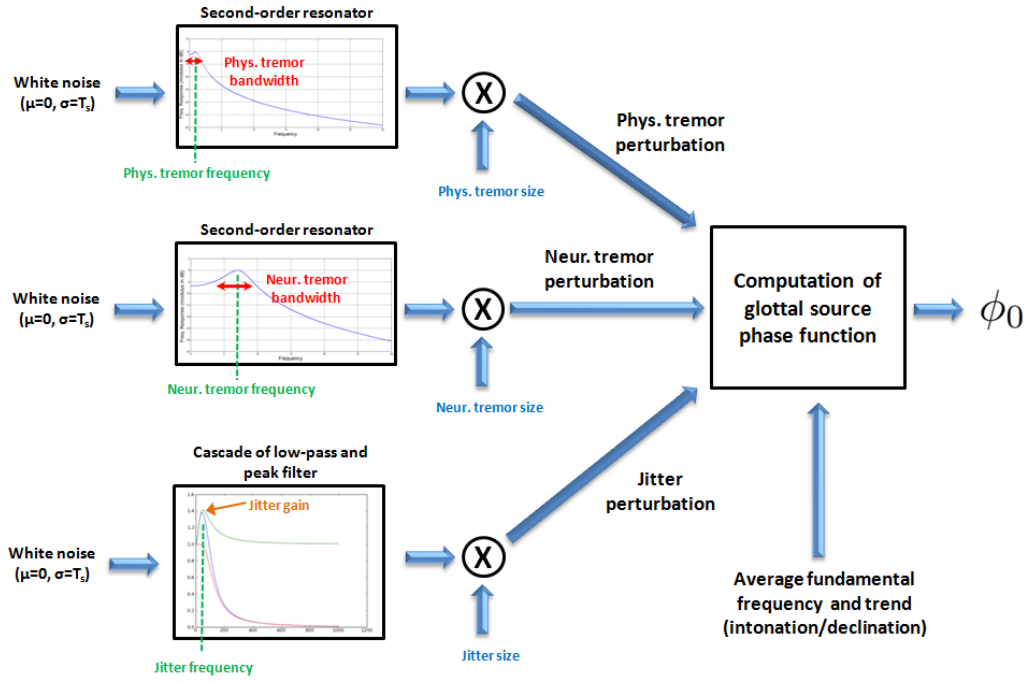


Figure 9.1 – Synthesizer of disordered voices : Phase function perturbation

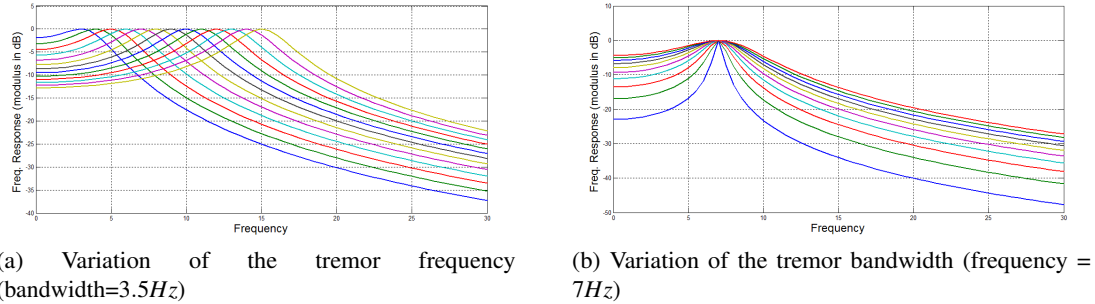


Figure 9.2 – Synthesizer of disordered voices : Frequency response of tremor filter

9.2.1.3 Jitter

The jitter perturbation time series is generated via white noise low-pass filtering. Vocal frequency jitter is related to muscle tension jitter during phonation. Figure 9.3a illustrates the amplitude spectrum of an electromyographic (EMG) signal of the CT muscle. One observes that the major part of energy is located below 400Hz and that a prominent frequency component appears in the vicinity of 50Hz . A filter (illustrated in Figure 9.3) is obtained via the cascade of a low-pass filter (with cut-off frequency = 100Hz) and a peak filter that can boost spectral components (specified via the jitter frequency and jitter boost parameters). A constant gain multiplier (i.e. jitter size parameter) is then used to fix the size of the frequency jitter.

9.2.1.4 Glottal source phase

The glottal source phase function $\phi_0(t)$ involves the average vocal frequency, its trend, and simulated perturbations. Assuming that the instantaneous glottal phase function derivative is the instantaneous vocal frequency, we obtain the following relation :

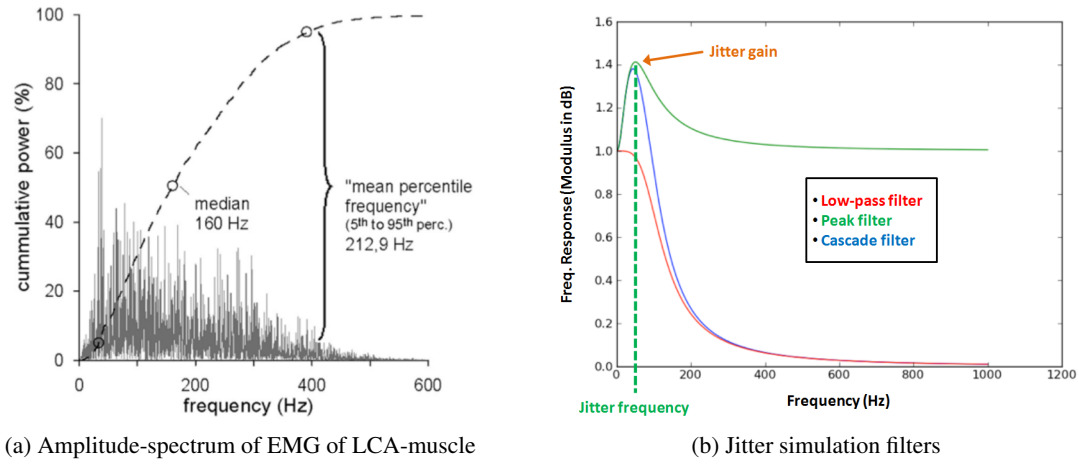


Figure 9.3 – Synthesizer of disordered voices : Jitter model

$$d\phi_0(t) = 2\pi f_{0,trend}(t) dt + p(t) \quad (9.1)$$

where $f_{0,trend}(t)$ and $p(t)$ are the trend and the total phase perturbation time series (sum of physiological, $p_{phys}(t)$, neurological, $p_{neur}(t)$, and jitter, $p_{jit}(t)$, perturbation components).

$$p(t) = p_{phys}(t) + p_{neur}(t) + p_{jit}(t) \quad (9.2)$$

Assuming that the sampling frequency is equal to $F_s = 200kHz$, the phase function is then computed for each sampling instant n .

$$\begin{cases} p[n] &= p_{phys}[n] + p_{neur}[n] + p_{jit}[n] \\ \Delta\phi_0[n] &= 2\pi f_{0,trend}[n] \Delta t + p[n] \\ \phi_0[n] &= \phi_0[n-1] + \Delta\phi_0[n] \end{cases} \quad (9.3)$$

9.2.2 Glottal airflow model

Figure 9.4 illustrates a cross-section of the glottis. The vocal cords are approximated by two oscillators moving out of phase. The system is assumed symmetrical. The motion of the vocal folds is controlled via the virtual glottal hemi-widths d_{entr} and d_{exit} . The instantaneous vibration frequency of the vocal cords is controlled by the glottal source phase function $\phi_0(t)$.

$$\begin{cases} d_{entr} &= \xi_{0,entr} + \xi_{1,entr} \sin(\phi_0) \\ d_{exit} &= \xi_{0,exit} + \xi_{1,exit} \sin(\phi_0 - \Phi) \end{cases} \quad (9.4)$$

Symbols ξ designate the vibration amplitudes or abduction amplitudes respectively. Phase delay Φ mimics the propagation delay between entrance and exit (i.e. converging/diverging glottis).

The glottal cross-section is thus a trapeze with bases denoted w_{entr} and w_{exit} . Assuming that the glottis area at the entrance and the exit must be positive or equal to 0 when the vocal cords enter in contact, widths w_{entr} and w_{exit} are expressed by means of the the virtual glottal hemi-widths as follows :

$$\begin{cases} w_{entr} &= \max\{0, 2d_{entr}\} \\ w_{exit} &= \max\{0, 2d_{exit}\} \end{cases} \quad (9.5)$$

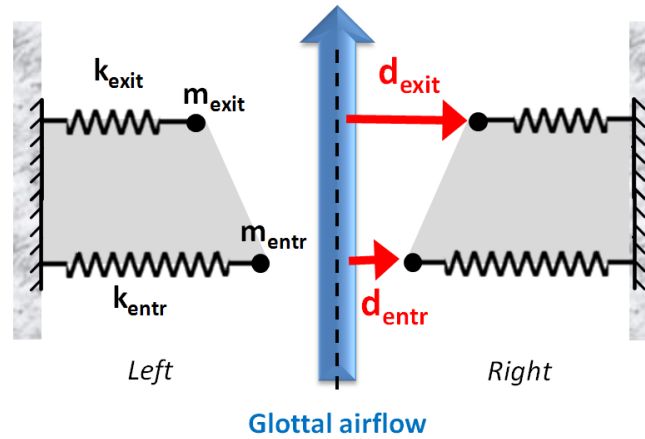


Figure 9.4 – Synthesizer of disordered voices : Glottis model (coronal cross-section)

The glottal airflow increases with glottis area A_g . Here, one assumes that the glottis area A_g is given by :

$$A_g = L_g \min \{w_{entr}, w_{exit}\} \quad (9.6)$$

where L_g is the glottal length.

In the synthesizer, the glottal airflow rate is then computed via the Rothenberg model [BA83].

9.2.3 Propagation through the vocal tract

The propagation of the glottal source excitation through the vocal tract is modelled by means of formant synthesizer [Kla80]. The latter involves several resonators, connected in series, that simulate the transfer function of the vocal tract. The frequency of each formant is controlled individually. The bandwidths are fixed via a statistical model involving the formant frequencies.

9.3 Tracking of cycle lengths

9.3.1 Overview

The cycle length tracking has been validated on the basis of synthetic stimuli. The goal of the cycle length tracking consists in selecting a subset of speech signal peaks that characterize the same glottal events. For that, the speech signal peaks are characterized by their positions and their saliences, and the selection is based on dynamic programming.

9.3.2 Corpus

The synthetic stimuli have been generated separately with increasing jitter or neurological tremor size, and for different values of F_0 (100Hz and 300Hz). For a fixed set of parameters, the synthesis is performed 10 times. Table 9.1 reports the parameters and their values/ranges of 4 experiments. The other parameters of the synthesizer have been kept constant.

In Experiment 1 and 3, only jitter parameters are modified. The jitter size parameter varies in the range $[1, 2, \dots, 10]$, which covers a very large set of perturbation sizes. The jitter boost has been fixed to 0dB. As a consequence, no narrow frequency band is favoured, which yields a cycle length perturbation time series that can be assimilated to low-pass filtered white noise.

In Experiment 2 and 4, cycle lengths are perturbed by low-frequency modulations only. The default

neurological tremor frequency and bandwidth are fixed respectively to $7Hz$ and $3.5Hz$ and the neurological tremor size parameter varies in the range $[1, 2, \dots, 20]$.

		EXPERIMENTS			
		1	2	3	4
Jitter	Size	$[1, 2, \dots, 10]$	0	$[1, 2, \dots, 10]$	0
	Frequency (Hz)	(*)	(**)	(*)	(**)
	Bandwidth (Hz)	100	(**)	100	(**)
	Gain (dB)	0	(**)	0	(**)
Neur. tremor	Size	0	$[1, 2, \dots, 20]$	0	$[1, 2, \dots, 20]$
	Frequency (Hz)	(**)	7	(**)	7
	Bandwidth (Hz)	(**)	3.5	(**)	3.5
F_0	Mean (Hz)	100	100	300	300

(*) : no influence, jitter boost = $0dB$

(**) : no influence, perturbation size = 0

Table 9.1 – Experimental conditions for the tracking validation

9.3.3 Influence of the perturbation sizes

9.3.3.1 Experiment conditions

For each synthetic voiced speech sound, the cycle length sequence $[d_1, d_2, d_3, \dots, d_I]$ (with I the number of cycles) is obtained separately from the synthesizer. A constant-step interpolated reference cycle length time series $x_{ref}(t)$ is then obtained by reconstructing the temporal axis as described in section 5.5.

The cycle length tracking, explained in Chapters 4 and 5, is applied to obtain the cycle length time series $x_{sclt}(t)$. The tracking parameters are :

Vocal frequency range	:	$[60Hz, 400Hz]$
Maximal local length perturbation	:	$\alpha = 35\%$
Second-order perturbation weight	:	$\gamma_1 = 1$
Salience weight	:	$\gamma_2 = 2$
Admissible preceding triplet selection	:	$\gamma_3 = 60\%$
Triplet sequence length weight	:	$\gamma_4 = 3$

The synthetic and tracked cycle length time series are then inter-correlated. The delay is fixed so that the inter-correlation coefficient is maximum. Cycle length time series $x_{sclt}(t)$ and $x_{ref}(t)$ are then analyzed within a same interval to compute their cycle length averages ($T_{0,sclt}$ and $T_{0,ref}$), their average fundamental frequencies ($F_{0,sclt} = 1/T_{0,sclt}$ and $F_{0,ref} = 1/T_{0,ref}$) and their coefficients of variation (CV_{sclt} and CV_{ref}).

For each set of parameters, the synthesizer generates 10 realizations, and thus 10 cue values. These are pooled and an ensemble average is computed for each set of parameters.

9.3.3.2 Results

Figure 9.5 shows the ensemble average vocal frequency $F_{0,sclt}$ (upper figure), the ensemble average coefficient of variation CV_{sclt} (middle figure) as well as the ensemble average correlation coefficient between the extracted and the reference cycle length time series (bottom figure). In the upper and middle figures, the red dotted line reports to the ensemble average reference cue ($F_{0,ref}$ and CV_{ref}).

One observes that, for Experiments 2 and 4, the ensemble average $F_{0,sclt}$ and CV_{sclt} cues are very close to the reference cue values. This agreement is confirmed by very high correlation coefficients

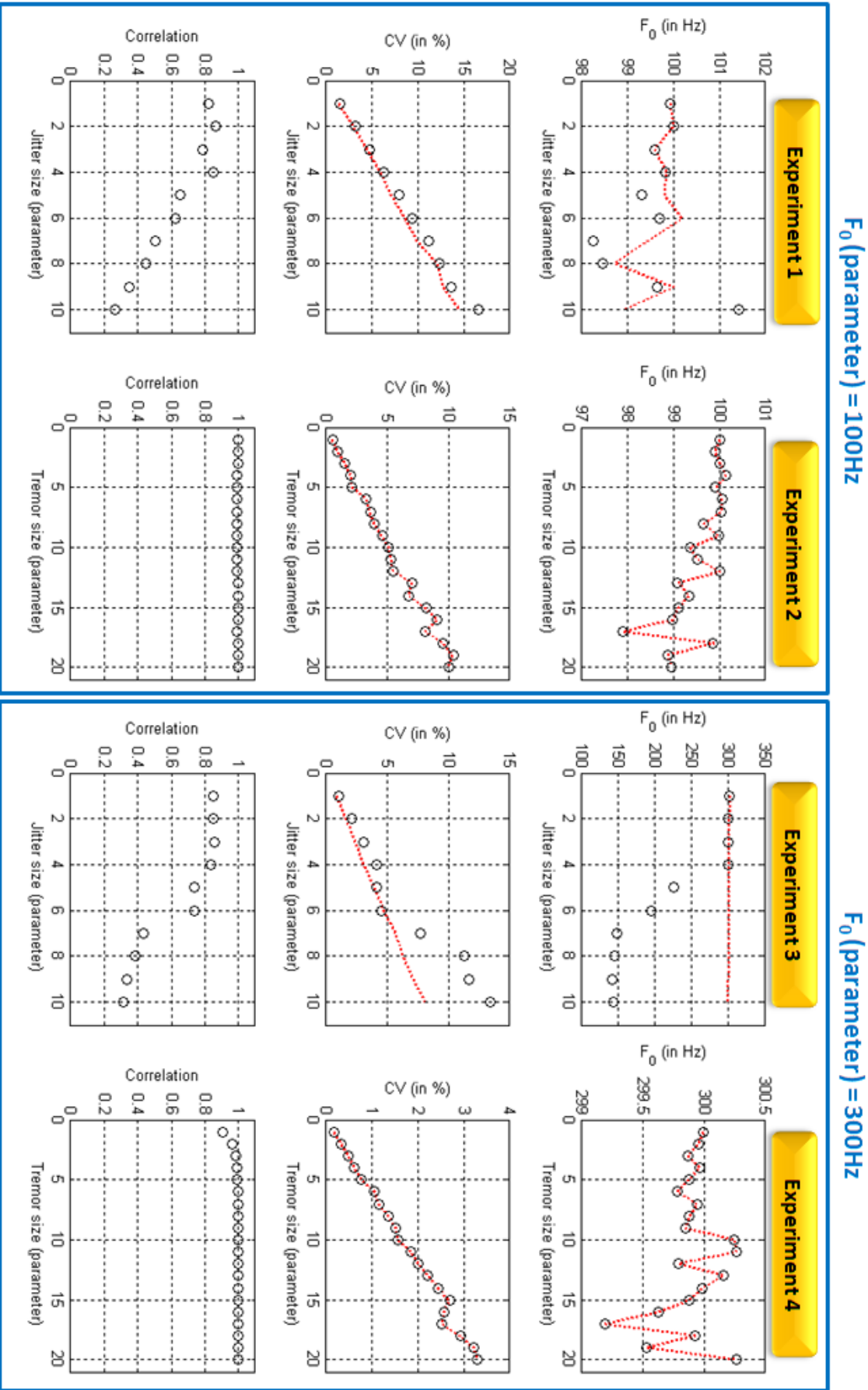


Figure 9.5 – Influence of perturbation size : Results of cycle length tracking

between the extracted and the reference time series, especially for increasing tremor size values. For small tremor size values, the correlation coefficients are slightly smaller. A possible explanation is that the tracking is based on peak selection. A high precision with regard to the peak position is thus required. For that, the speech signal has been upsampled to a sampling frequency equal to 200kHz and the peak positions have been refined. However, for small CV_{ref} values (i.e. small low-frequency modulation depths), these variations in amplitude are not detectable. Figure 9.6 illustrates two synthetic stimuli ($F_{0,ref} \approx 100\text{Hz}$ or 300Hz) characterized by high neurological tremor sizes (i.e. neurological tremor size parameter = 20). Red dots indicate the peaks that are selected by the tracking method.

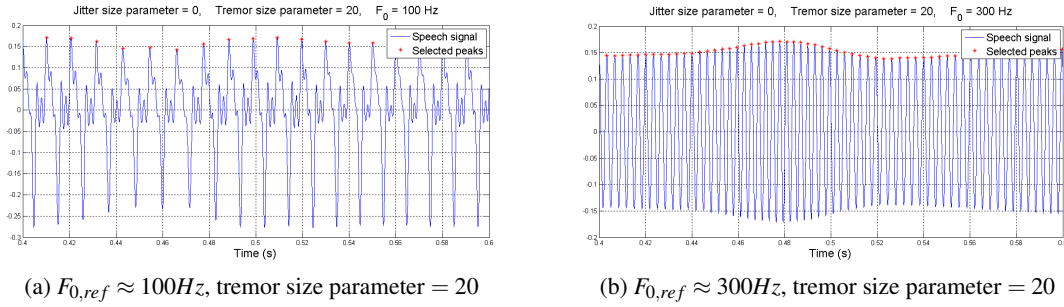


Figure 9.6 – Influence of the low-frequency perturbation size : Illustration of Experiments 2 and 4

In Experiments 1 and 3, only jitter size evolves. One observes in Experiment 1 that the correlation coefficients are smaller compared to Experiment 2 and 4 and decrease starting from a parameter value equal to 5 (i.e. a $CV_{ref} \approx 7\%$). Considering that the glottal source phase function is perturbed by low-pass filtered white noise, an explanation is that the temporal peak-based tracking selects a subset of speech cycle peaks via a dynamic programming approach that rests on a pseudo-regularity criterion (second order difference of successive cycle length candidates). For large and fast cycle length perturbations, other peaks in the vicinity of the relevant speech cycle peaks may be preferred to obtain a less perturbed pseudo-regular cycle length time series. However, one observes that, globally, the perturbation size cues CV_{sclt} are slightly higher than the reference cues CV_{ref} . Figure 9.7 illustrates two synthetic stimuli ($F_{0,ref} \approx 100\text{Hz}$ or 300Hz) characterized by large jitter (i.e. jitter size parameter = 10) and reports the peaks that are selected by the tracking method.

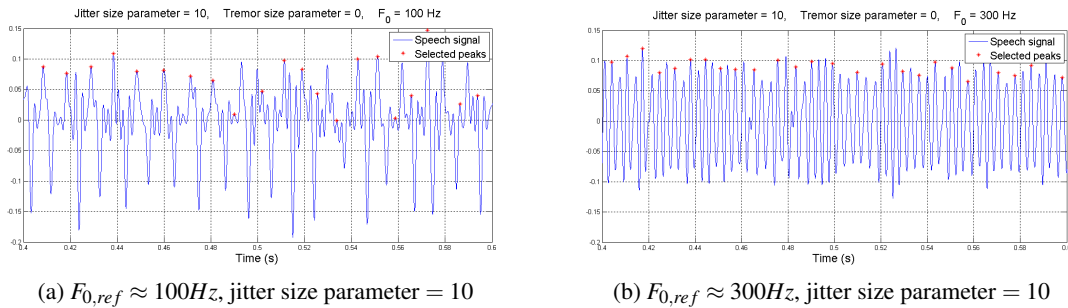


Figure 9.7 – Illustration of Experiments 1 and 3

Finally, in Experiment 3, one observes that the accuracy of the cycle length tracking method drops starting from a size equal to 5 (i.e. $CV_{ref} \approx 4\%$). Indeed, the ensemble average $F_{0,sclt}$ decreases monotonically from 300Hz to 150Hz and then remains constant for larger perturbations.

Moreover, the evolution of F_0 with jitter sizes indicates that the average extracted cycle lengths evolves from $1/300\text{Hz}$ to $1/150\text{Hz}$ when the jitter size parameter increases. These results suggest that for large perturbations the tracking favours sequences of twice the cycle length. Notice that the dynamic programming algorithm, via its state variable L and parameter γ_4 , enables tracking cycle length time series characterized by large cycle-to-cycle length perturbations, but, beyond a threshold, the tracking method prefers less perturbed cycle length sequences. Another explanation is that large excursions around the average $F_{0,ref}$ lead to proportionally larger excursions of the $F_{0,ref}$ harmonics. Assuming that the frequency response of the vocal tract is kept unchanged during synthesis, the perturbation of the vocal frequency generates perturbations of the cycle amplitudes. These variations of the speech cycle amplitudes (called *shimmer*) therefore influence the salience assigned to speech cycle peaks, which may bias the peak selection.

In conclusion, the results indicate that the cycle length tracking is able to deliver reliable cycle length sequences. Especially, in presence of low-frequency modulations of the cycle lengths, the tracker is able to deal with very large perturbations ($\approx 10\%$). In case of high-frequency perturbations, the tracker is able to extract relevant cycle length perturbations up to $\approx 4\%$ over the whole range of expected fundamental frequencies.

9.3.4 Influence of background noise

9.3.4.1 Experiment conditions

The proposed cycle length tracking is assigned to the *cycle-synchronous event analysis* category. Such techniques enable an accurate analysis of time-evolving F_0 speech sounds, but are directly affected by additive noise. Here, the reliability of the tracking is evaluated by adding increasing amount of white noise to the previous synthesized speech sounds. Let $n_w(t)$ be a normally distributed white noise (with mean 0 and standard deviation 1) and snr the signal-to-noise ratio, the synthetic speech sounds $y(t)$ have been modified as follows :

$$y_{noisy}(t) = y(t) + a_w n_w(t) \quad \text{where : } a_w = \frac{\sigma_y}{10^{(snr/20)}} \quad (9.7)$$

In this experiment, the snr values evolves from 30dB to 5dB . For example, Figures 9.8a and 9.8b illustrate two synthetic sounds (clean ($snr = \infty$) and noisy ($snr = 5\text{dB}$)).

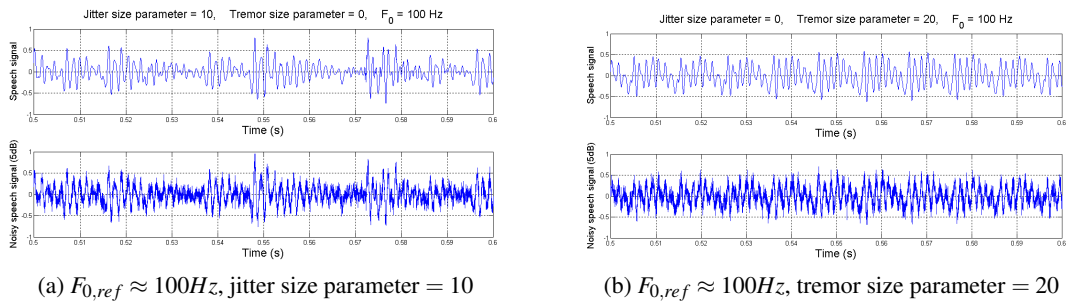


Figure 9.8 – Illustration of noise addition

9.3.4.2 Results

Figure 9.9 illustrates, for several snr values, the ensemble average $F_{0,sclt}$ (upper figure) and CV_{sclt} (middle figure) cues as well as the ensemble average correlation coefficient between the extracted and the synthetic reference cycle length time series (bottom figure). In the upper and middle figures, the red dotted line reports the ensemble average reference cue ($F_{0,ref}$ and CV_{ref}).

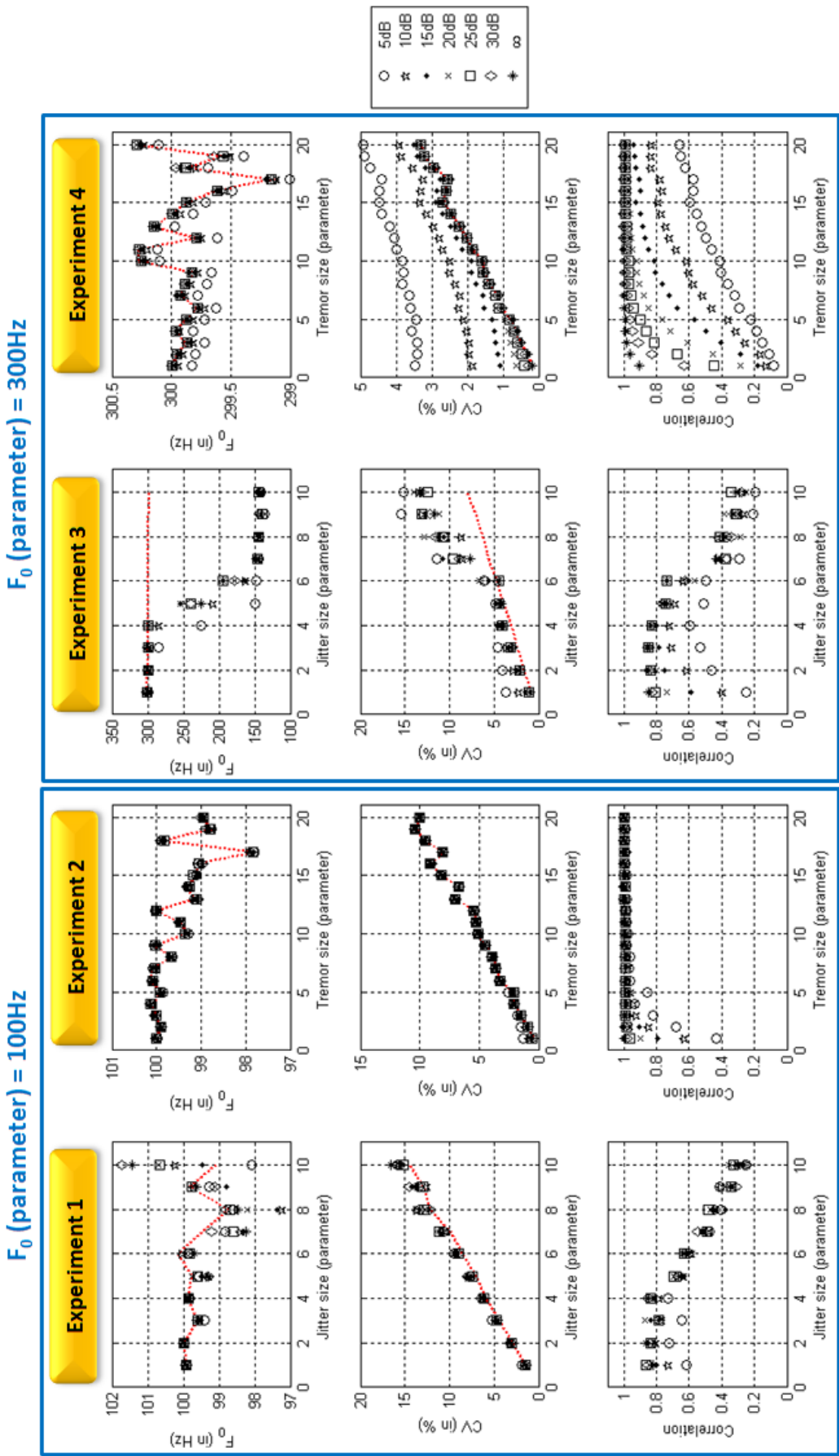
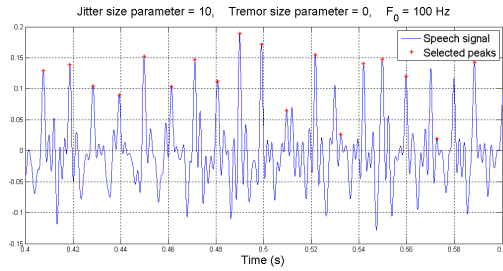
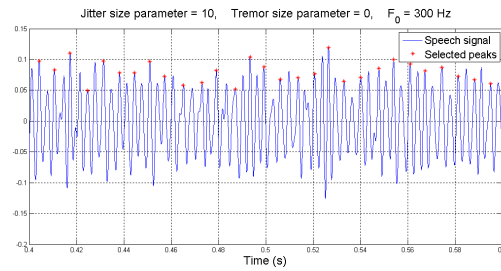


Figure 9.9 – Cycle length tracking with additive noise. The *snr* values evolve from 30dB to 5dB

One observes the influence of background noise on the tracking results. For Experiments 1 and 3 (i.e. variations of the jitter size), the correlation coefficients decrease with increasing noise levels, which suggests the selection of spurious peaks in the vicinity of the relevant speech cycle peaks. Because of the noise, the cycle lengths are locally perturbed and the tracked perturbation sizes increase. Notice that this effect is mainly visible for Experiment 3 ($F_{0,ref} \approx 300\text{Hz}$) for which the re-affiliation of selected peaks induces higher relative perturbations because of the shorter cycle lengths. Figure 9.10 illustrates two synthetic stimuli ($F_{0,ref} \approx 100\text{Hz}$ or 300Hz) characterized by large jitter (i.e. jitter size = 10) and reports the peaks that are selected by the tracking for $snr=5\text{dB}$. The Figure displays synthetic speech fragments after band-pass filtering (60Hz - 500Hz).



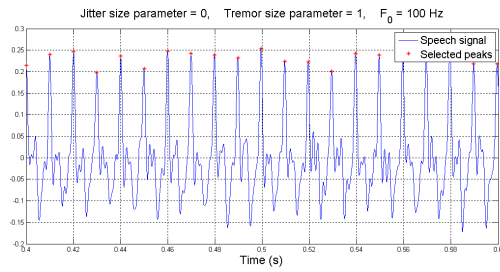
(a) $F_{0,ref} \approx 100\text{Hz}$, jitter size parameter = 10, $snr=5\text{dB}$



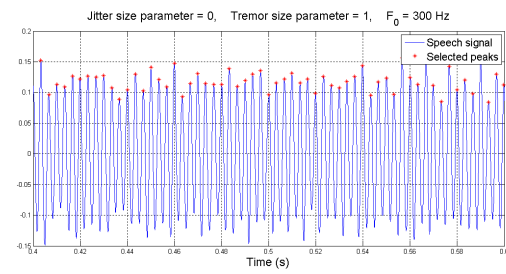
(b) $F_{0,ref} \approx 300\text{Hz}$, jitter size parameter = 10, $snr=5\text{dB}$

Figure 9.10 – Illustration of Experiments 1 and 3 with additive noise, after band-pass filtering

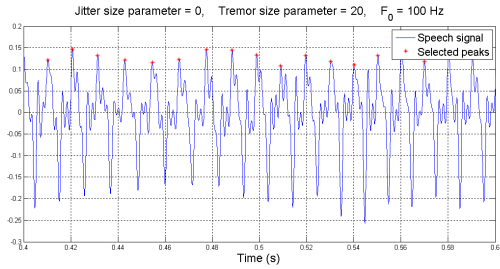
The influence of the background noise is also present for Experiments 2 and 4 (i.e. variations of the neurological tremor size). Especially in Experiment 4, feeble correlation coefficients and high CV_{sclt} values for snr values lower than 15dB are observed due to the combination of small reference perturbation sizes and short cycle lengths. Figure 9.11 illustrates four stimuli ($F_{0,ref} \approx 100\text{Hz}$ or 300Hz) synthesized with several neurological tremor sizes and reports the tracked peaks for $snr=5\text{dB}$.



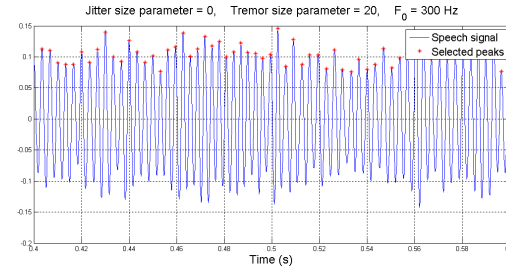
(a) $F_{0,ref} \approx 100\text{Hz}$, tremor size parameter = 1, $snr = 5\text{dB}$



(b) $F_{0,ref} \approx 300\text{Hz}$, tremor size parameter = 1, $snr = 5\text{dB}$



(c) $F_{0,ref} \approx 100\text{Hz}$, tremor size parameter = 20, $snr = 5\text{dB}$



(d) $F_{0,ref} \approx 300\text{Hz}$, tremor size parameter = 20, $snr = 5\text{dB}$

Figure 9.11 – Illustration of Experiments 2 and 4 with additive noise, after band-pass filtering

9.3.5 Comparison with Praat Software

9.3.5.1 Experimental conditions

The proposed tracking is now compared to the method used in a well-known speech analysis software, called *Praat* [BW01]. In this software, two short-term analysis methods, based on the autocorrelation or the cross-correlation, are proposed for extracting the pitch contour. Praat developers suggest the use of the cross-correlation settings to obtain the most reliable cycle length values. The algorithm (called *Pitch (cc)*) performs an acoustic periodicity detection on the basis of a forward cross-correlation analysis [Tal95].

The *Praat* algorithm is frame-based and delivers an F_0 estimate for each frame position. Based on that estimate, the cycle length is determined and several cues are computed, which are the standard deviation $\sigma_{0,praat}$ and average $T_{0,praat}$ of the raw period sequence $[p_1, p_2, \dots, p_L]$ (i.e. non-constant step interpolated sequence). Default parameter values have been specified, except for the vocal frequency range that has been set to $[60Hz, 400Hz]$ to enable the comparison with our *salience-based cycle length tracking* (SCLT) method.

The short-term analysis method has been applied to the previously synthesized corpus (variation of jitter and neurological tremor size parameters, 10 realizations for each set of parameters) with increasing *snr* values. For each stimulus, the average vocal frequency $F_{0,praat}^*$ and the coefficient of variation CV_{praat}^* of the period sequence is computed :

$$CV_{praat}^* = \frac{\sigma_{0,praat}}{T_{0,praat}} \quad (9.8)$$

As a consequence, reference cues, CV_{ref}^* and $F_{0,ref}^*$, related to the synthesizer and cues CV_{sclt}^* and $F_{0,sclt}^*$ obtained via the proposed cycle length tracking have also been computed on the basis of the corresponding non-constant step interpolated cycle length sequences.

For each set of parameters, the ten cue values have been pooled and an ensemble average computed.

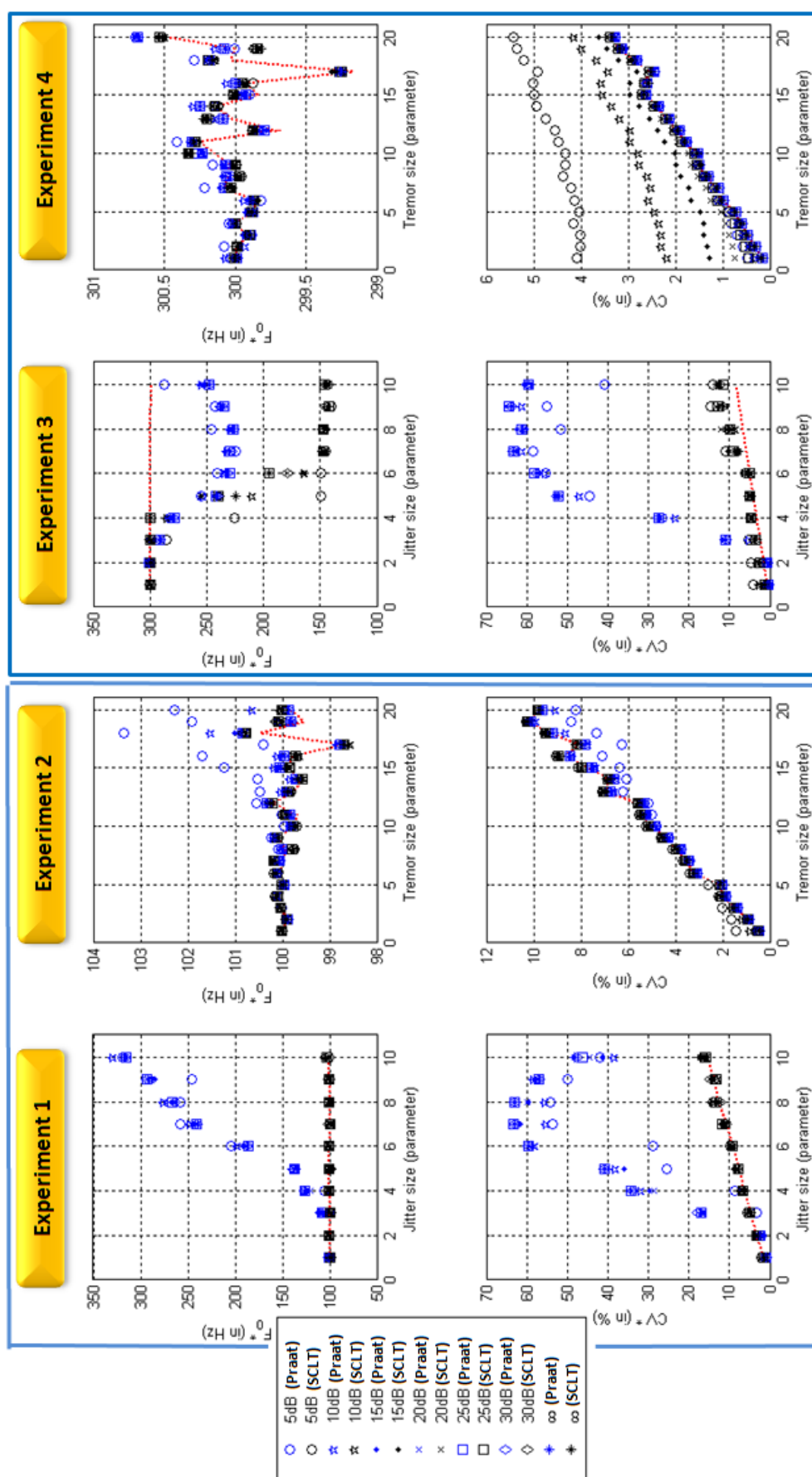
9.3.5.2 Results

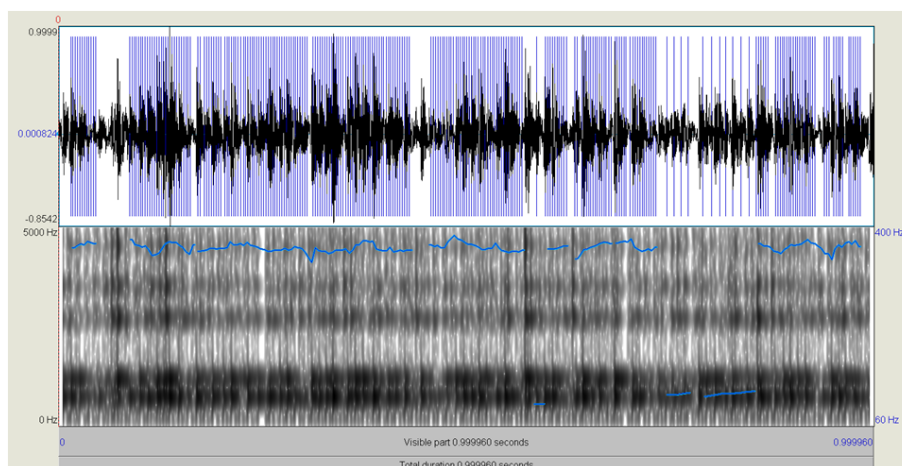
Comparison with unconstrained *Praat* pitch extraction method

Figure 9.12 illustrates, for various *snr* values, the ensemble average vocal frequencies cues $F_{0,sclt}^*$ and $F_{0,praat}^*$ (upper figure), and ensemble average coefficient of variation cues, CV_{sclt}^* and CV_{praat}^* (bottom figure). The red dotted line refers to the ensemble average reference cue ($F_{0,ref}^*$ and CV_{ref}^*). The cues related to the SCLT tracking and *Praat* tracking are displayed respectively in black and blue.

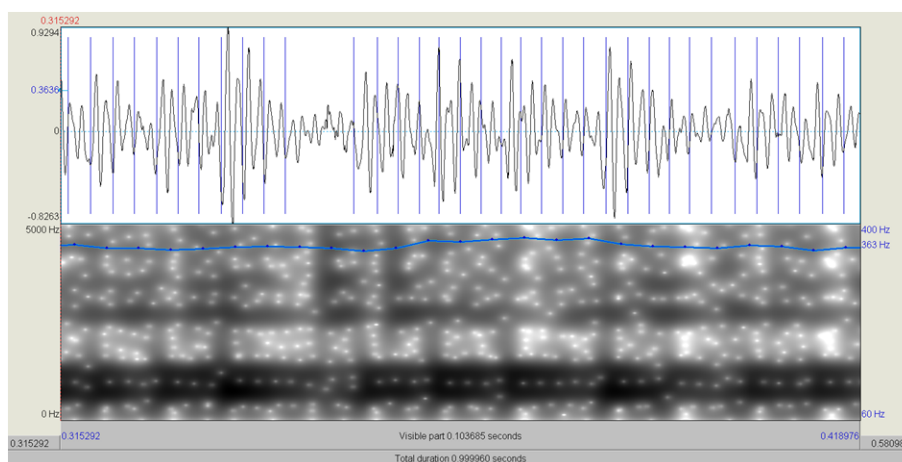
For Experiment 1 and 3, one observes that the short term analysis method, implemented in *Praat*, is unable to track reliably cycle lengths in presence of high fast cycle-to-cycle length perturbations (jitter size parameter higher than 2, corresponding to a reference coefficient of variation $CV_{ref}^* \geq 3\%$). However, for low-frequency perturbations of the cycle lengths (Experiments 2 and 4), one observes a good agreement between the ensemble average vocal frequency $F_{0,praat}^*$ and $F_{0,ref}^*$, except for low *snr* levels and large neurological tremor sizes.

For instance, Figure 9.13 reports two screen shots of the *Praat* user interface. Each Figure illustrates the analyzed speech sound in the temporal domain (upper part) and its spectrogram (bottom part). The cycle length tracking results are reported in the temporal domain by means of blue markers that delimit the cycle boundaries. In this example, the test stimulus is a synthetic voiced speech sound with a jitter size parameter = 10 and a F_0 parameter = 100Hz (*snr*=5dB). One

Figure 9.12 – Comparison with unconstrained Praat method. *SCLT* refers to *salience-based cycle length tracking*



(a) Total duration



(b) Zoom

Figure 9.13 – Praat user interface (unconstrained method), synthetic test stimulus

observes in Figure 9.13a that vocal cycles are missed and that the extracted cycle lengths differ widely in time. Figure 9.13b reports a fragment of the stimulus. In Figure 9.13b, the pitch estimate is $\approx 363\text{Hz}$. Therefore, pitch markers have tracked intra-cycle oscillations. This occurs in the presence of rapid F_0 changes and is a consequence of the lack of reliability of short-term analyses that rely on the cycle length regularity assumption.

The problems with *Praat* in the presence of high cycle length perturbations may be solved by assuming that the vocal frequency range is known a priori. Therefore, the previous experimental conditions have been modified to facilitate the cycle length tracking by *Praat*. For that, the vocal frequency range in *Praat* has been set to respectively $[70\text{Hz}, 130\text{Hz}]$ and $[250\text{Hz}, 350\text{Hz}]$ in agreement with the vocal frequency synthesis parameter (i.e. 100Hz for Experiments 1 and 2, and 300Hz for Experiments 3 and 4). The vocal frequency ranges of the SCLT method have been kept unchanged.

Comparison with constrained *Praat* pitch extraction method

Figure 9.15 illustrates, for several *snr* values, the ensemble average vocal frequencies cues, $F_{0,sclt}^*$ and $F_{0,praat}^*$, and ensemble average coefficients of variation, CV_{sclt}^* and CV_{praat}^* .

At present, a good agreement between $F_{0,ref}$ and $F_{0,praat}$ is obtained. The robustness of the *Praat* short-term analysis with regard to the background noise has improved. Indeed, frame-based analysis methods are less affected by noise or signal degradation because of their reliance on the signal auto-correlation or cross-correlation [Hes83].

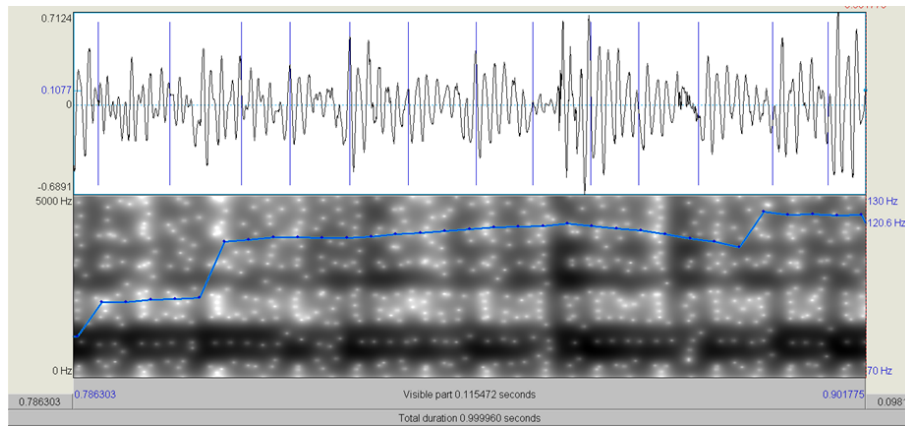
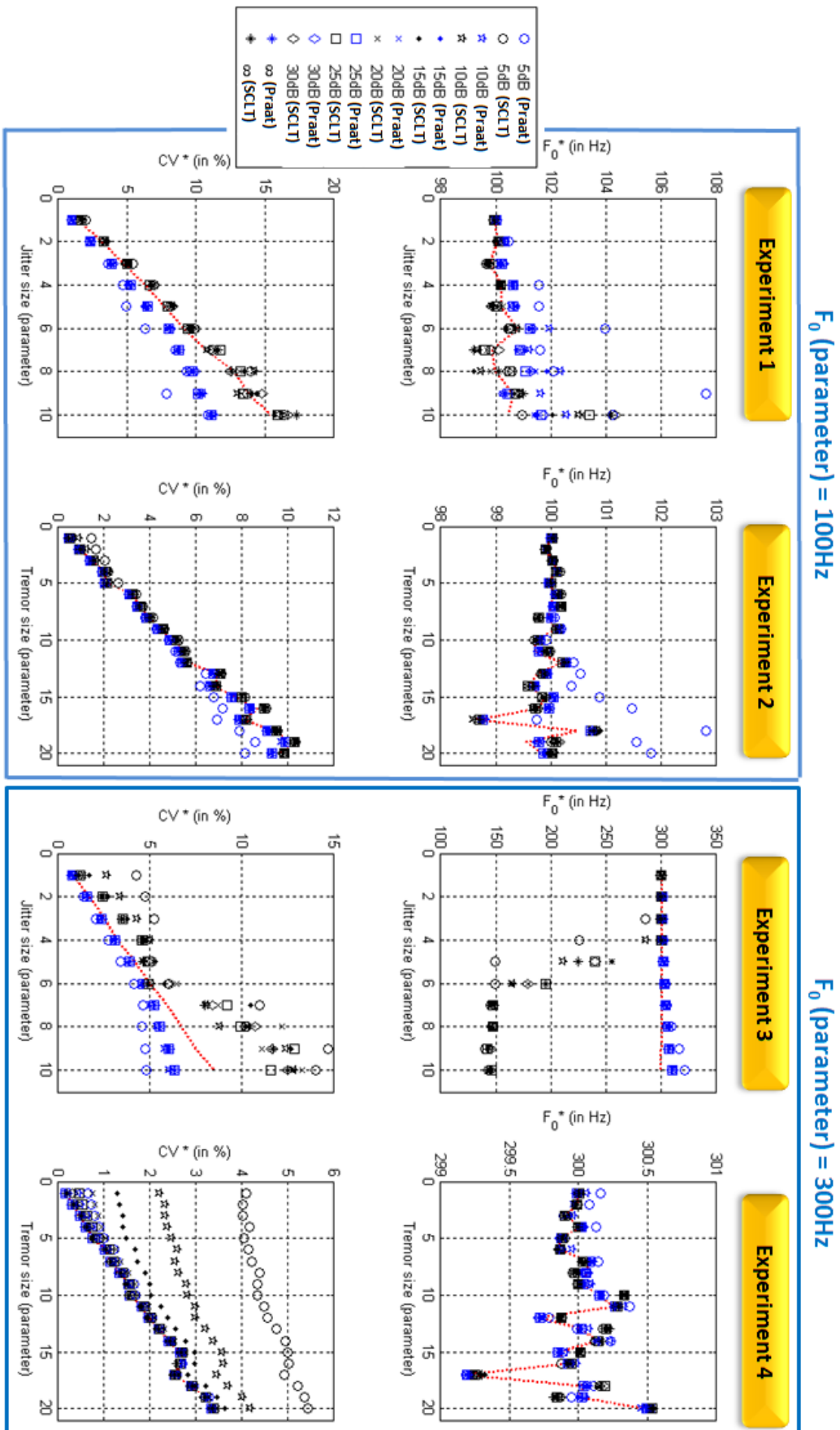


Figure 9.14 – Praat user interface (constrained method), test stimulus

Nevertheless, one observes that, in many cases, perturbation size cues $CV_{0,praat}$ are lower than the reference cues CV_{ref} and are unable to deal with large jitter (see Experiment 1 and 3). Conversely, CV_{sclt} cues are generally slightly larger than the reference cue values. As an example, Figure 9.14 shows a screen shot of the *Praat* user interface during the analysis of a synthetic stimulus with a jitter size parameter = 10 and a F_0 parameter = 100Hz ($snr=5\text{dB}$). One observes that the blue pulse markers that delimit glottal cycles are located at different instants within each cycle and, therefore, do not report exactly the same glottal events. As a consequence, the cycle length tracking method implemented in *Praat* obtains less perturbed cycle sequences, which explains the smaller computed perturbation size cues.

Figure 9.15 – Comparison with constrained Praat method. *SCLT* refers to *salience-based cycle length tracking*

9.4 Perturbation analysis

9.4.1 Corpus

The cycle length perturbation analysis method has been validated by means of synthetic stimuli generated by the same synthesizer of disordered voices (section 9.2). Here, only the cycle length sequences have been synthesized. The synthesis parameter ranges are listed in Table 9.2. These perturbation size ranges have been reduced compared to the previous simulations because the goal, here, is not to test the limitations of the cycle tracking. Notice that the physiological tremor has not been simulated. For each set of parameters, only one synthetic cycle length sequence is generated, i.e. a total of $11 \cdot 10 \cdot 13 \cdot 10 = 14300$ sequences.

	Symbol	Parameter	Range	# values
Jitter	$X_{\sigma,jit}$	Size	$[0, 0.1, 0.2, 0.3, \dots, 0.9, 1]$	11
		Frequency (Hz)	(*)	/
		Bandwidth (Hz)	100	1
		Gain (dB)	0	1
Neur. tremor	$X_{\sigma,neur}$	Size	$[1, 2, 3, \dots, 9, 10]$	10
	$X_{f,neur}$	Frequency (Hz)	$[3, 4, 5, \dots, 14, 15]$	13
	$X_{bw,neur}$	Bandwidth (Hz)	$[0.5, 1, 1.5, \dots, 4.5, 5]$	10
F_0		Mean (Hz)	100	1

(*) : no influence, jitter boost = 0dB

Table 9.2 – Experimental conditions for the perturbation analysis validation

9.4.2 Method

For each cycle length sequence, the temporal axis has been reconstructed to obtain a constant-step interpolated time series $x_{ref}(t)$, followed by empirical mode decomposition, perturbation categorization and vocal cue computation.

The vocal cues that have been computed in the framework of this validation are listed in Table 9.3. They report the sizes of jitter, neurological tremor and physiological tremor, the neurological tremor frequency and bandwidth described in chapter 8.

The validation consists in reporting the relation between synthesizer parameters and cues. For that, a multiple linear regression model is fitted to the data. The predictors are the synthesizer parameters and the responses are individual vocal cues. Notice that, to enable coefficient value comparison, the predictors as well as the response are z-normalized by subtracting the variable mean and dividing by the standard deviation.

Let $\{\mathcal{X}\}$ be an array of P predictors (i.e. the normalized synthesizer parameter values) and $\{\mathcal{C}\}$ an array of the P corresponding coefficients. Each cue is thus expressed via a linear multi-dimensional relation as follows :

$$cue = \{\mathcal{C}\} \cdot \{\mathcal{X}\}^T \quad (9.9)$$

The reliability of the cues is then assessed by means of the values of the linear model coefficients and the percentage of variability in the response (R^2) explained by the model. The computed coefficients are reported by means of spider plots, which comprises $2P$ axes : P axes for positive coefficient values and P axes for negative values.

Jitter size	σ_{jit}	Standard deviation of the jitter time series $x_{jit}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Neurological tremor depth	σ_{neur}	Standard deviation of the neurological tremor time series $x_{neur}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Physiological tremor depth	σ_{phys}	Standard deviation of the physiological tremor time series $x_{phys}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Neurological tremor frequency	$\hat{f}_{\mu,neur}$	Average of the estimated neurological tremor frequency time series $\hat{f}_{neur}(t)$ weighted by the neurological tremor envelope.	Section 8.5.4
	$\hat{f}_{dens,neur}$	Abscissa of the center of gravity of the estimated neurological tremor frequency density $\hat{F}(f;h)$ in the frequency interval $[0, 15Hz]$.	Eq. (8.19)
	$\hat{f}_{quant,neur}$	Abscissa of the center of gravity of the estimated neurological tremor frequency density $\hat{F}(f;h)$ in the frequency interval selected via scalar quantization.	Eq. (8.26)
Neurological tremor bandwidth	$\delta \hat{f}_{neur}$	Standard deviation of the estimated neurological tremor frequency time series $\hat{f}_{neur}(t)$ weighted by the neurological tremor envelope.	Section 8.5.4
	$\hat{b}_{dens,neur}$	Standard deviation of the estimated neurological tremor frequency density $\hat{F}(f;h)$ in the frequency interval $[0, 15Hz]$.	Eq. (8.20)
	$\hat{b}_{quant,neur}$	Width of the frequency interval retained after scalar quantization.	Eq. (8.27)

Table 9.3 – Summary of vocal cues considered for perturbation analysis validation

9.4.3 Results

9.4.3.1 Perturbation sizes

Perturbations σ_{jit} , σ_{neur} and σ_{phys} have been predicted by multiple linear regression models. The predictors are the jitter size synthesizer parameter $X_{\sigma,jit}$ and the neurological tremor size $X_{\sigma,neur}$, frequency $X_{f,neur}$ and bandwidth $X_{bw,neur}$ parameters. The models are thus characterized by the following linear relations.

$$\begin{cases} \sigma_{jit} &= \{\mathcal{C}_1\} \cdot \{\mathcal{X}\}^T \\ \sigma_{neur} &= \{\mathcal{C}_2\} \cdot \{\mathcal{X}\}^T \\ \sigma_{phys} &= \{\mathcal{C}_3\} \cdot \{\mathcal{X}\}^T \end{cases} \quad (9.10)$$

where :

$$\begin{cases} \{\mathcal{C}\} &= \{C_{\sigma,jit}, C_{\sigma,neur}, C_{f,neur}, C_{bw,neur}\} \\ \{\mathcal{X}\} &= \{X_{\sigma,jit}, X_{\sigma,neur}, X_{f,neur}, X_{bw,neur}\} \end{cases} \quad (9.11)$$

Figure 9.16 reports the normalized coefficients, as well as the R^2 values, for the 3 perturbation cues. One observes that the neurological tremor depth model explains $\approx 82\%$ of the variance. Moreover, neurological tremor depth σ_{neur} is mainly influenced by the neurological tremor size $X_{\sigma,neur}$ and bandwidth $X_{bw,neur}$ parameters. The influences of $X_{f,neur}$ and $X_{\sigma,jit}$ are feeble.

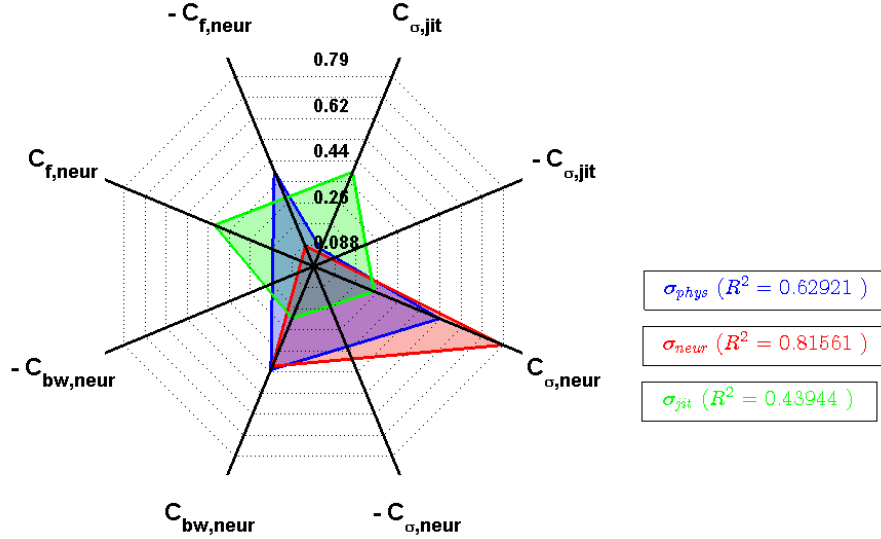


Figure 9.16 – Linear regression model coefficients of the perturbation size cues

The physiological tremor depth model explains $\approx 63\%$ of the variance. Once again, the coefficients $C_{\sigma,neur}$ and $C_{bw,neur}$ are high and positive. One observes also a high and negative coefficient $C_{f,neur}$. This suggests that the frequency interval $< 2Hz$ is less affected by high-frequency neurological tremor components.

Finally, the jitter size model explains $\approx 44\%$ of the variance. Here, the four parameters contribute positively to jitter size, especially $X_{jit,size}$ and $X_{neur,freq}$. So, the jitter size σ_{jit} increases with the jitter size parameter and increasing neurological tremor frequency parameters transfer a fraction of the neurological tremor energy to the frequency interval $> 15Hz$.

The observation of significant influences of $X_{\sigma,neur}$ and $X_{bw,neur}$ suggest an alternative multiple linear regression model. This is explained by the fact that, during synthesis, the size of the neurological tremor phase perturbation is influenced by the neurological tremor depth and bandwidth parameters (see section 9.2.1.2). So, here, a fifth predictor variable is added, equal to the product of the tremor depth and tremor bandwidth parameters.

$$\begin{cases} \sigma_{jit} &= \{\mathcal{C}_1^*\} \cdot \{\mathcal{X}^*\}^T \\ \sigma_{neur} &= \{\mathcal{C}_2^*\} \cdot \{\mathcal{X}^*\}^T \\ \sigma_{phys} &= \{\mathcal{C}_3^*\} \cdot \{\mathcal{X}^*\}^T \end{cases} \quad (9.12)$$

where :

$$\begin{aligned} \{\mathcal{C}^*\} &= \{C_{\sigma,jit}, C_{\sigma,neur}, C_{f,neur}, C_{bw,neur}, C_{bw,neur} * \sigma_{neur}\} \\ \{\mathcal{X}^*\} &= \{X_{\sigma,jit}, X_{\sigma,neur}, X_{f,neur}, X_{bw,neur}, X_{bw,neur} * X_{\sigma,neur}\} \end{aligned} \quad (9.13)$$

Figure 9.17 reports the 5 coefficients, as well as the R^2 values, for the 3 perturbation size cues. Globally, the percentages of explained variance increase. Moreover, at present, the interaction between the neurological tremor size and bandwidth parameters has the greatest influence on

σ_{neur} and σ_{phys} . Previous observations with regard to the transfer of neurological tremor energy to adjacent frequency intervals are confirmed.

Notice that the influence of jitter size on σ_{neur} is always smaller than the influence of neurological tremor depth (or interaction between depth and bandwidth) on σ_{jit} . This is expected because the jitter energy is spread out on a wide frequency band and, therefore, only a small fraction of the jitter energy affects the frequency interval of neurological tremor ($\approx [2Hz, 15Hz]$).

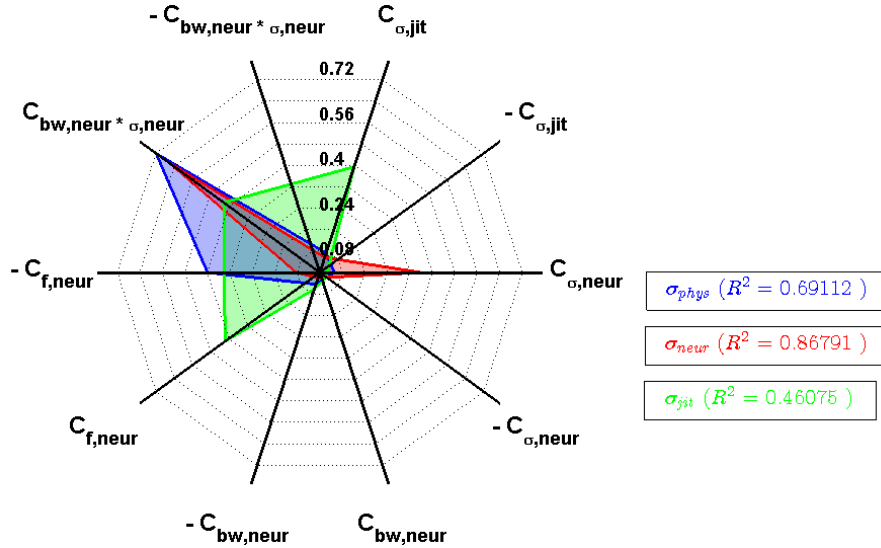


Figure 9.17 – Linear regression model coefficients for the perturbation size cues (with interaction)

9.4.3.2 Neurological tremor frequency

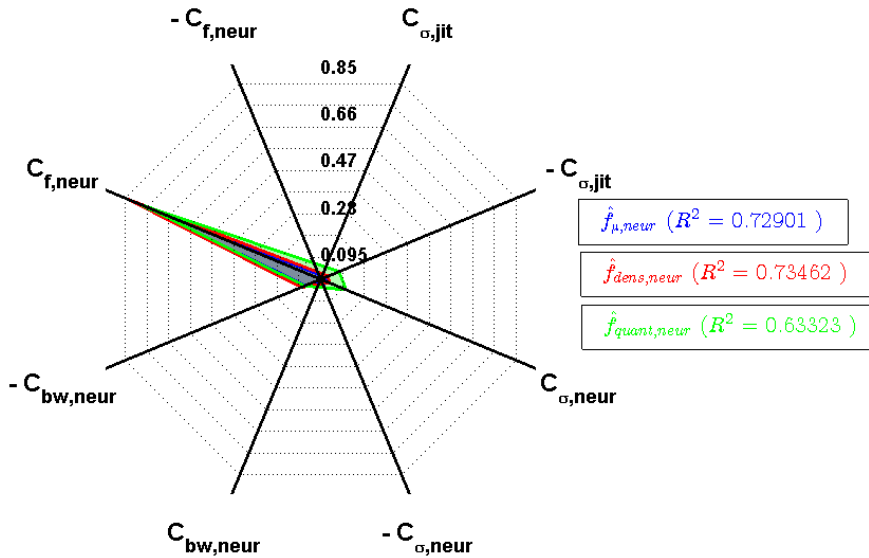


Figure 9.18 – Linear regression model coefficients for the neurological tremor frequency cues

Figure 9.18 reports the coefficients, as well as the R^2 values, for the 3 neurological tremor frequency cues ($\hat{f}_{\mu,neur}$, $\hat{f}_{dens,neur}$ and $\hat{f}_{quant,neur}$).

$$\begin{cases} \hat{f}_{\mu,neur} &= \{\mathcal{C}_4\} \cdot \{\mathcal{X}\}^T \\ \hat{f}_{dens,neur} &= \{\mathcal{C}_5\} \cdot \{\mathcal{X}\}^T \\ \hat{f}_{quant,neur} &= \{\mathcal{C}_6\} \cdot \{\mathcal{X}\}^T \end{cases} \quad (9.14)$$

where :

$$\begin{cases} \{\mathcal{C}\} &= \{C_{\sigma,jit}, C_{\sigma,neur}, C_{f,neur}, C_{bw,neur}\} \\ \{\mathcal{X}\} &= \{X_{\sigma,jit}, X_{\sigma,neur}, X_{f,neur}, X_{bw,neur}\} \end{cases} \quad (9.15)$$

The predictors $\{\mathcal{X}\}$ are the jitter size parameter $X_{\sigma,jit}$ and the neurological tremor depth $X_{\sigma,neur}$, frequency $X_{f,neur}$ and bandwidth $X_{bw,neur}$ parameters.

One observes that the 3 neurological tremor frequency cues are mainly influenced by the neurological tremor frequency $X_{f,neur}$. However, the R^2 values are different. These differences may be explained by inspecting Figure 9.19 that reports two scattergrams. These scattergrams compare the $\hat{f}_{dens,neur}$ cue to respectively $\hat{f}_{\mu,neur}$ and $\hat{f}_{quant,neur}$. Black and blue dots refer to the stimuli characterized respectively by a small or large neurological tremor depth σ_{neur} . The threshold is median σ_{neur} . The red line is the bisector.

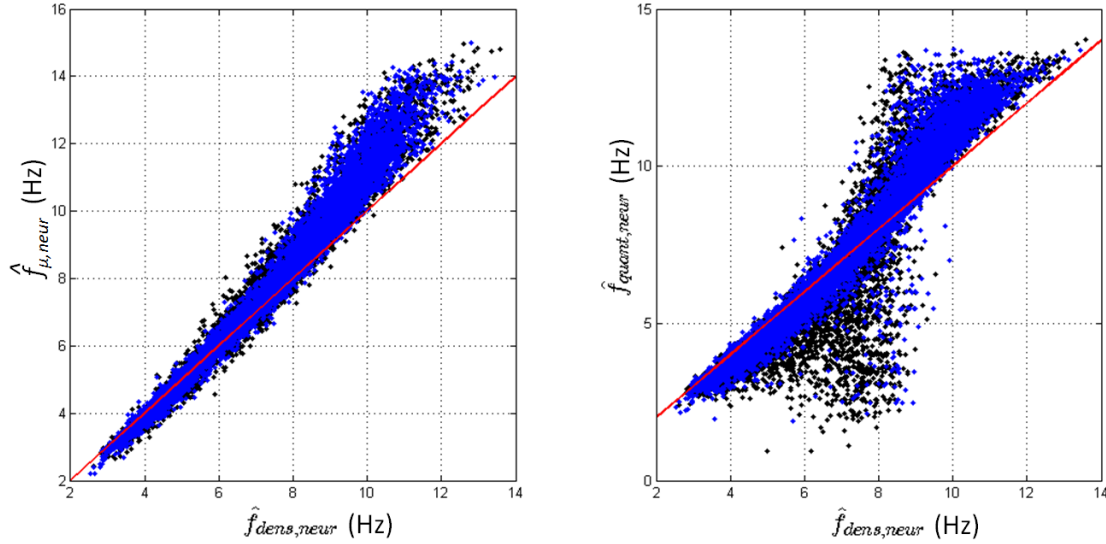


Figure 9.19 – Comparison of the neurological tremor frequency cue candidates

On the one hand, one observes a good agreement between $\hat{f}_{dens,neur}$ and $\hat{f}_{\mu,neur}$ cues. This is expected because, as a first approximation, these two cues have been computed on the basis of the instantaneous mode frequencies and envelopes by integrating successively, but in reverse order, along the temporal and frequency axis. Moreover, these two cues appear to be unaffected by the neurological tremor depth. On the other hand, one observes that the $\hat{f}_{quant,neur}$ is directly affected by the neurological tremor depth so that the selection of a relevant frequency interval by scalar quantization is harder when the neurological tremor depth is small compared to the jitter size.

9.4.3.3 Neurological tremor bandwidth

Figure 9.20 reports the 4 coefficients, as well as the R^2 values, for the 3 neurological tremor bandwidth cues. ($\delta \hat{f}_{neur}$, $\hat{b}_{dens,neur}$ and $\hat{b}_{quant,neur}$).

$$\begin{cases} \delta \hat{f}_{neur} &= \{\mathcal{C}_7\} \cdot \{\mathcal{X}\}^T \\ \hat{b}_{dens,neur} &= \{\mathcal{C}_8\} \cdot \{\mathcal{X}\}^T \\ \hat{b}_{quant,neur} &= \{\mathcal{C}_9\} \cdot \{\mathcal{X}\}^T \end{cases} \quad (9.16)$$

where :

$$\begin{aligned} \{\mathcal{C}\} &= \{C_{\sigma,jit}, C_{\sigma,neur}, C_{f,neur}, C_{bw,neur}\} \\ \{\mathcal{X}\} &= \{X_{\sigma,jit}, X_{\sigma,neur}, X_{f,neur}, X_{bw,neur}\} \end{aligned} \quad (9.17)$$

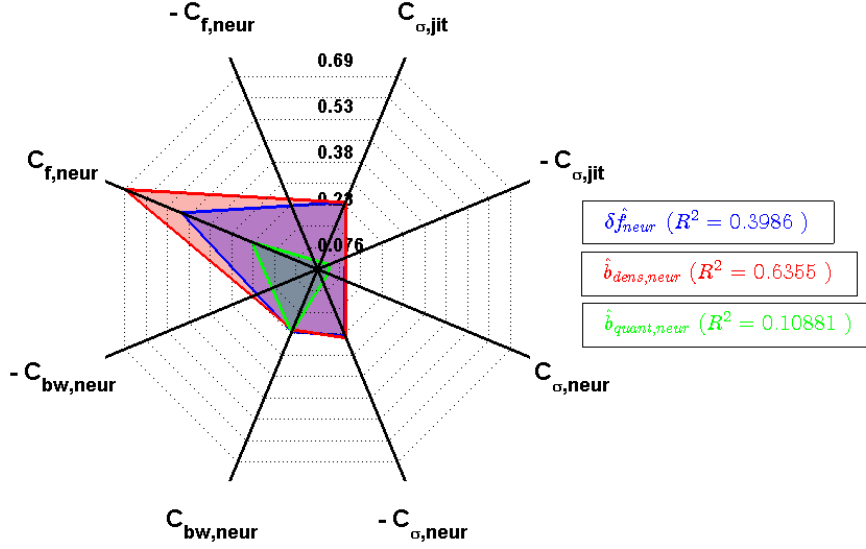


Figure 9.20 – Linear regression model coefficients for the neurological tremor bandwidth cues

First, the $\hat{b}_{quant,neur}$ cue, which reports only the width of the frequency interval selected by scalar quantization, can be rejected ($R^2 \approx 10\%$). The neurological tremor cues obtained by scalar quantization, $\hat{f}_{quant,neur}$ and $\hat{b}_{quant,neur}$, are therefore discarded in the following. For the two other cues, one observes that they are mainly influenced by the neurological tremor frequency parameter $X_{f,neur}$ rather than by the neurological tremor bandwidth parameter $X_{bw,neur}$. A first explanation is that the $\hat{b}_{dens,neur}$ cue is computed within a fixed frequency interval $[0Hz, 15Hz]$. Boundary effects may bias the computation of the bandwidth. Another explanation is related to the second-order resonator used for the synthesis of the neurological tremor phase perturbation (see section 9.2.1.2). Figure 9.21 illustrates several frequency responses (modulus) of the second-order resonator obtained for different tremor bandwidth and frequency parameters. For each set of parameters, one computes a bandwidth cue via the same mathematical expression than $\hat{b}_{dens,neur}$ (see Equation 8.20) by computing into the frequency interval $[0, 15Hz]$ the abscissa of the center of gravity f_g and, then, computing the weighted standard deviation with regard to f_g .

One observes that the weighted standard deviation $\hat{b}_{dens,neur}$ increases with increasing $X_{bw,neur}$ but also with increasing $X_{f,neur}$ parameter values, as previously reported by the cue. One observes also a positive contribution of the jitter size parameter $X_{\sigma,jit}$ and a negative contribution of the neurological tremor depth parameter $X_{\sigma,neur}$. In other terms, the computation of relevant bandwidths require a relative neurological tremor depth that is large compared to the other perturbation sizes.

Due to the higher R^2 value (see Figures 9.18 and 9.20), only the neurological frequency cues $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ obtained via the frequency probability density are retained in the next chapter for the analysis of natural signals.

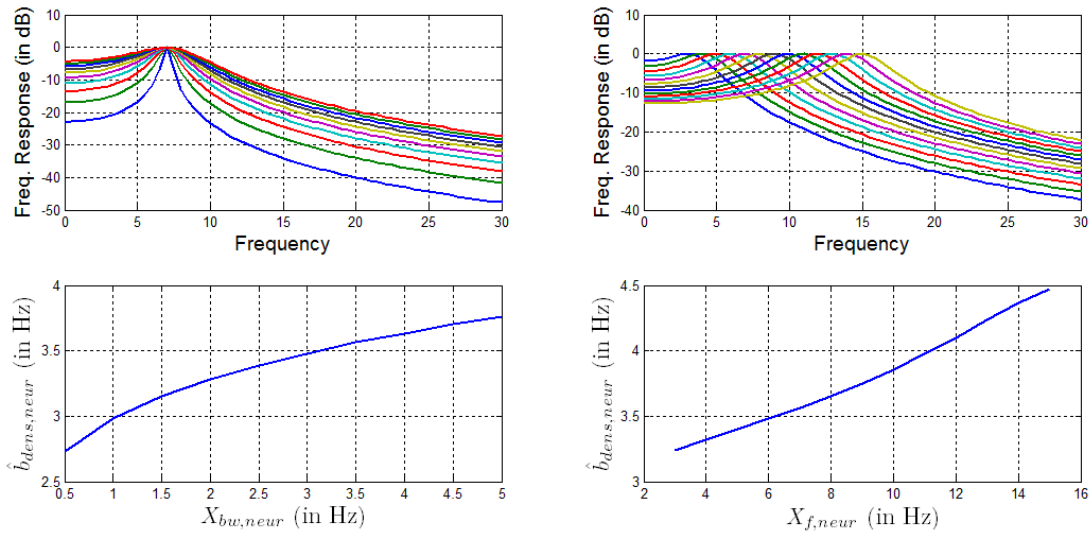


Figure 9.21 – Tremor bandwidth corresponding to the resonator parameters of the synthesizer

9.5 Conclusions

In this chapter, the cycle length tracking method as well as the cycle length perturbation analysis are validated by means of synthetic voiced speech sounds.

The results show a good agreement between the extracted cycle length time series and the reference cycle length time series produced by the synthesizer. Different kinds of behaviour have been observed depending on the kind of cycle length perturbations. In the presence of low-frequency modulation of the cycle lengths, the accuracy has been demonstrated and the method is able to deal with very high perturbation sizes (up to 10%). In the case of fast and large cycle-to-cycle length perturbations, the tracking method is able to extract reliably cycle lengths up to 4% of perturbation for the whole range of vocal frequencies.

Robustness with regard to the background noise has also been tested. The results indicate that for signal-to-noise ratios higher than 15dB the tracking method is able to obtain reliable cycle length sequences.

Finally, the method has been compared to an algorithm implemented in the *Praat* software. This algorithm, based on cross-correlation, is frame-based and delivers pitch contour estimates that are used for the tracking of cycle lengths. Although the robustness of *Praat* with regard to background noise is higher than the proposed method, the frequency search interval of the *Praat* algorithm has to be constrained to guarantee accuracy.

To conclude, the merit of the proposed cycle length tracking is its ability to track reliably vocal cycle lengths with no a priori assumption with regard to the vocal cycle length regularity and interval.

The results of the cycle length perturbation analyses validation indicate that the proposed vocal cues report adequately the corresponding synthesizer parameters. A selection is made between the different neurological tremor frequency and bandwidth cues proposed in chapter 8.

Key points

- The SCLT method as well as the vocal cycle length perturbation analysis method has been validated by means of stimuli that have been generated by a synthesizer of disordered voices.
- The SCLT method is able to deal with very high low-frequency cycle length perturbation sizes (up to 10%)
- In the case of fast and large cycle-to-cycle length perturbations, the SCLT method is able to extract reliably cycle lengths up to 4% of perturbation for the whole range of vocal frequencies
- In presence of background noise, the SCLT method is able to obtain reliable cycle length sequences for signal-to-noise ratios higher than 15dB
- The vocal cycle length perturbation analyses have been validated via a multiple linear regression model analysis. Some of the proposed vocal cues report adequately the corresponding synthesizer parameters



10. Parkinson and control speakers

Objectives of this chapter

- Analyze the vocal cycle length perturbations of two speech sound corpora sustained by control and Parkinson speakers
- Compare and discuss the discrepancies between corpora

Contents

10.1	Introduction	187
10.2	Corpora	188
10.2.1	Corpus from Bochum University Clinic	188
10.2.2	Corpus from Pays d'Aix Hospital	190
10.3	Vocal cues	191
10.4	Analysis of the corpus from Bochum University Clinic	192
10.4.1	Quartiles and histograms of vocal cues	192
10.4.2	Correlation	192
10.4.3	Comparison between control and Parkinson speakers	194
10.5	Analysis of the corpus from Pays d'Aix Hospital	196
10.5.1	Quartiles and histograms of vocal cues	196
10.5.2	Correlation	196
10.5.3	Comparison between control and Parkinson speakers	198
10.5.4	Effects of the use of medication	200
10.6	Comparison between corpora	202
10.6.1	Quartiles and histograms of vocal cues	202
10.6.2	Correlations	204
10.6.3	Three-way variance analysis	204
10.6.4	Perturbation size	205
10.6.5	Neurological tremor frequency and bandwidth	207
10.6.6	Vocal frequency F_0	208
10.7	Discussion and conclusion	209
10.7.1	Discrepancies between corpora	209
10.7.2	Patient attributes	209
10.7.3	Statistically significant effects of corpus, gender or pathology	209

10.1 Introduction

The salience-based cycle length tracking and the analysis of vocal cycle length perturbations are carried out for two corpora of vowels [a] sustained by control speakers and patients suffering from Parkinson's disease.

For each corpus, an ANOVA analysis is carried out to test for the effects of N explanatory variables on the mean of the cues. Here, a two-way ANOVA is used with 2 levels for each explanatory variable. The first variable, denoted "pathology" is related to the diagnosis. Its two levels are *Control* and *Parkinson*. The second independent variable, denoted "gender" has two levels that are *Male* and *Female*.

The method tests if the null hypothesis for all explanatory variable levels and their interaction is acceptable or may be rejected. In other terms, the analysis consists in determining if the gender or the pathology, or the interaction of these two variables, has an statistically significant effect on the average vocal cues.

For that, the analysis results reports probability values for the null hypotheses on the 2 main effects and their interactions. Basically, if the p-value associated with the factor "pathology" is lower than 5%, then the average vocal cue value differs statistically significantly between the control and Parkinson speakers. Moreover, if the p-value associated with the interaction between "gender" and "pathology" is also lower than 5% for this same vocal cue, one observes that male and female speakers do not evolve identically with regard to the pathology. In this case, a one-way ANOVA with explanatory variable "pathology" is separately carried out for male and female speakers.

A comparison between the corpora, as well as the pooled control and Parkinson speakers is also carried out.

10.2 Corpora

The corpora comprise vowels [a] sustained by control speakers and patients suffering from Parkinson's disease.

10.2.1 Corpus from Bochum University Clinic

The first corpus has been recorded at a sampling frequency of 44.1 kHz in WAV format in the same recording environment and by means of the same equipment at the Department of Neurology of Bochum University Clinic (Bochum, Germany). This corpus comprises 74 control speaker recordings (42♂, 32♀) and 205 Parkinson speaker recordings (129♂, 76♀).

Speaker chronological age as well as additional patient informations are available and reported hereafter :

- Duration since diagnosis (*DUR.*) : this is the time (in years) since the first diagnosis of the pathology. Notice that this duration does not report the time since the first symptoms but the duration since the diagnosis.
- Unified Parkinson's disease rating scale (*UPDRS*) : this scale is a popular scale to follow the longitudinal course of Parkinson's disease and is designed to monitor patient disability and impairment.

The UPDRS comprises 4 parts : *Mentation, behaviour and mood, Activities of daily living, Motor examination* and *Complications of therapy (in the past week)*. All these parts comprise several sections for which a score, based on a rating scale (0, 1, 2, 3, 4 for parts I to III or 0, 1 for part IV), is attributed. The final score is obtained by summing all individual scores. Notice that the speech quality of patients is only assessed in part II and III, as follows :

Part II : Activities of daily living

Speech

- 0 normal
 - 1 mildly affected, no difficulty being understood
 - 2 moderately affected, may be asked to repeat
 - 3 severely affected, frequently asked to repeat
 - 4 unintelligible most of time
-

Part III : Motor examination

Speech

- 0 normal
 - 1 slight loss of expression, diction, volume
 - 2 monotone, slurred but understandable, mod. impaired
 - 3 marked impairment, difficult to understand
 - 4 unintelligible
-

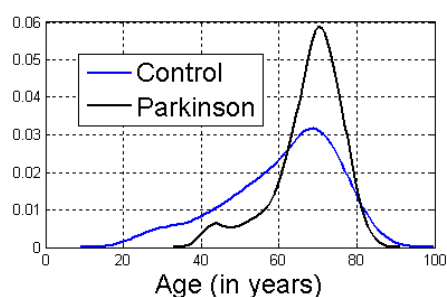
- Hoehn and Yahr scale (*HY*) : A commonly used scale for describing how the symptoms of Parkinson's disease progress. It is also used in conjunction with the UPDR scale. The modified Hoehn and Yahr scale comprises 7 stages of severity that are reported hereafter :

Stage	Description
1	Unilateral involvement only
1.5	Unilateral and axial involvement
2	Bilateral involvement without impairment of balance
2.5	Mild bilateral disease with recovery on pull test
3	Mild to moderate bilateral disease; some postural instability; physically independent
4	Severe disability; still able to walk or stand unassisted
5	Wheelchair bound or bedridden unless aided

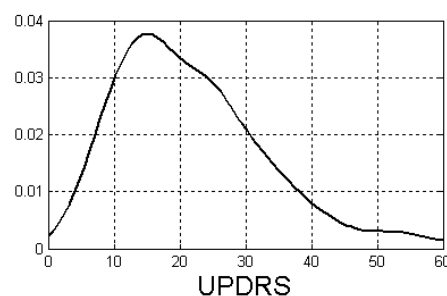
Table 10.1 reports the quartiles of ages, *UPDRS*, *HY* and duration since diagnosis (*DUR.*). Quartiles Q_1 to Q_3 report the values at 25%, 50%, and 75% of the feature value range. Figure 10.1 reports the distributions of these clinical cues.

	CONTROL		PARKINSON							
	AGE		AGE		UPDRS		HY		DUR.	
	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀
MIN	26	27	41	42	5	6	1	1	1	1
Q_1	56	47.5	65	64	14	14	2	2	2	2.25
Q_2	66	63	69	70	20.5	19	2.5	2.5	6	6
Q_3	71	69.5	72	75	29	27	3	3	10	9
MAX	83	80	82	83	53	61	4	4	18	30

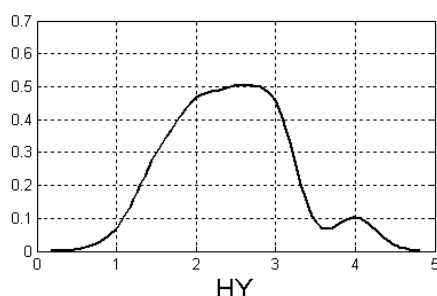
Table 10.1 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Quartiles of chronological ages, *UPDRS*, *HY* and duration since diagnosis (*DUR.*). Quartiles Q_1 to Q_3 report the values at 25%, 50%, and 75% of the feature value range.



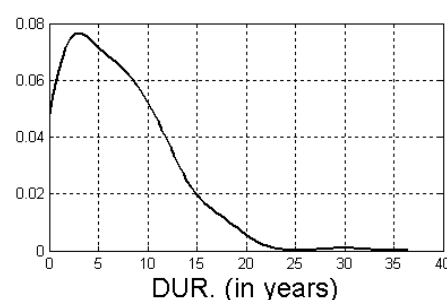
(a) Distribution of the chronological age of control and Parkinson speakers



(b) Distribution of the Unified Parkinson's disease rating scores (*UPDRS*)



(c) Distribution of the Hoehn and Yahr scores (*HY*)



(d) Distribution of the time since diagnosis (*DUR.*)

Figure 10.1 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Distributions of the chronological age for control and Parkinson speakers, as well as *UPDRS*, *HY* scores and duration since diagnosis for Parkinson speakers only

10.2.2 Corpus from Pays d'Aix Hospital

The second corpus has been recorded at the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) [Ghi+12]. This corpus comprises 123 control speaker recordings (50♂, 73♀) and 456 Parkinson speaker recordings (302♂, 154♀).

Speaker ages as well informations with regard to the use of L-DOPA in the clinical treatment of Parkinson's disease are available. Other information like the profession, the country of origin or the language of origin of speakers are available but not used in this study. Table 10.2 reports the quartiles of the age of control and Parkinson (with or without DOPA treatment) speaker. Figure 10.2 illustrates the distribution of age for control and Parkinson speakers.

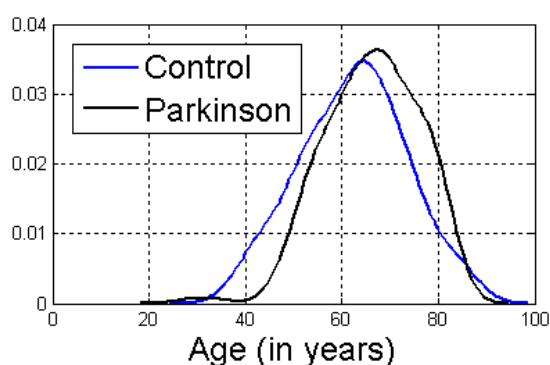


Figure 10.2 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Distribution of chronological age for control and Parkinson speakers.

	AGE							
	CONTROL		PARKINSON					
			Dopa OFF		Dopa ON		Dopa n/a	
	♂(50)	♀(73)	♂(137)	♀(70)	♂(149)	♀(77)	♂(16)	♀(7)
MIN	43	37	47	28	46	28	51	59
Q_1	58	53	59	60	59	60	58.5	66.25
Q_2	64.5	63	67	65.5	67	66	68	67
Q_3	70	70	74	73	74	73	75	68
MAX	86	86	85	85	85	85	81	75

Table 10.2 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Quartiles of age of control and Parkinson speakers (with or without DOPA treatment).

10.3 Vocal cues

The vocal cues that have been kept and computed for the analysis of the corpora are reported in Table 10.3. These cues are related to the perturbation sizes (jitter, neurological tremor and physiological tremor), the neurological tremor frequency and bandwidth.

Jitter size	σ_{jit}	Standard deviation of the jitter time series $x_{jit}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Neurological tremor depth	σ_{neur}	Standard deviation of the neurological tremor time series $x_{neur}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Physiological tremor depth	σ_{phys}	Standard deviation of the physiological tremor time series $x_{phys}(t)$ divided by the average of the intonation time series $x_{int}(t)$.	Eq. (8.4)
Neurological tremor frequency	$\hat{f}_{dens,neur}$	Abscissa of the center of gravity of the estimated neurological tremor frequency density $\hat{F}(f;h)$ in the frequency interval $[0, 15Hz]$.	Eq. (8.19)
Neurological tremor bandwidth	$\hat{b}_{dens,neur}$	Standard deviation of the estimated neurological tremor frequency density $\hat{F}(f;h)$ in the frequency interval $[0, 15Hz]$.	Eq. (8.20)
Average vocal frequency	F_0	Inverse of the average of the intonation time series $x_{int}(t)$.	Eq. (8.3)

Table 10.3 – Summary of retained vocal cues

10.4 Analysis of the corpus from Bochum University Clinic

10.4.1 Quartiles and histograms of vocal cues

Table 10.4 reports the quartiles of the size of jitter, the neurological and physiological tremor depths (all in %). The last three lines of the table report the neurological tremor frequency and bandwidth, and the average F_0 (all in Hz). Figure 10.3 reports the histograms of the vocal cues.

		Male		Female	
		CTRL	PARK	CTRL	PARK
σ_{jit} (%)	Q_0	0.14	0.15	0.18	0.15
	Q_1	0.22	0.30	0.25	0.33
	Q_2	0.42	0.42	0.36	0.45
	Q_3	0.74	0.76	0.58	0.87
	Q_4	2.21	3.19	1.38	2.38
σ_{neur} (%)	Q_0	0.48	0.52	0.32	0.53
	Q_1	0.66	0.74	0.49	0.82
	Q_2	1.08	1.06	0.72	1.19
	Q_3	1.41	1.41	1.13	1.62
	Q_4	2.84	3.01	1.83	4.12
σ_{phys} (%)	Q_0	0.39	0.35	0.15	0.30
	Q_1	0.55	0.60	0.40	0.64
	Q_2	0.79	0.85	0.65	0.91
	Q_3	1.35	1.15	1.05	1.36
	Q_4	2.48	2.45	1.52	3.66
$\hat{f}_{dens,neur}$ (Hz)	Q_0	3.34	3.33	3.34	3.20
	Q_1	4.19	4.26	4.03	3.99
	Q_2	4.58	4.79	4.39	4.60
	Q_3	5.24	5.43	4.66	5.20
	Q_4	6.96	7.05	6.42	6.80
$\hat{b}_{dens,neur}$ (Hz)	Q_0	1.50	1.84	1.89	1.67
	Q_1	2.35	2.42	2.18	2.25
	Q_2	2.81	2.80	2.57	2.63
	Q_3	3.22	3.14	2.79	3.01
	Q_4	3.77	3.72	3.34	3.65
F_0 (Hz)	Q_0	77.04	86.59	153.91	99.28
	Q_1	100.48	109.55	165.14	159.89
	Q_2	115.46	124.71	186.32	175.69
	Q_3	129.19	141.64	210.31	197.74
	Q_4	173.99	202.40	258.05	240.20

Table 10.4 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Quartiles of the size of jitter σ_{jit} , the neurological σ_{neur} and physiological σ_{phys} tremor depths (all in %). The last three lines of the Table report the tremor frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$, and the average F_0 (all in Hz). Quartiles Q_0 to Q_4 report the values at 2.5%, 25%, 50%, 75% and 97.5% of the feature value range.

10.4.2 Correlation

Table 10.5 reports Pearson’s linear correlation coefficients between the vocal perturbation cues and the patient attributes (age, UPDRS, time since diagnosis (DUR.) and HY). The correlation coefficient with a corresponding p-value $< 5\%$ are displayed in bold.

One observes that patient scores are only slightly correlated with perturbation cues ($r \leq 20\%$). The UPDRS index as well as duration since the diagnosis are correlated ($r \approx 60\%$) with the HY index. One observes also a correlation between the physiological σ_{phys} and neurological σ_{neur} tremor depths ($r = 69\%$) and between the neurological tremor frequency $\hat{f}_{dens,neur}$ and its bandwidth $\hat{b}_{dens,neur}$ ($r = 65\%$). The correlation coefficient between jitter σ_{jit} and physiological tremor σ_{phys} sizes is small and the other r values are feeble.

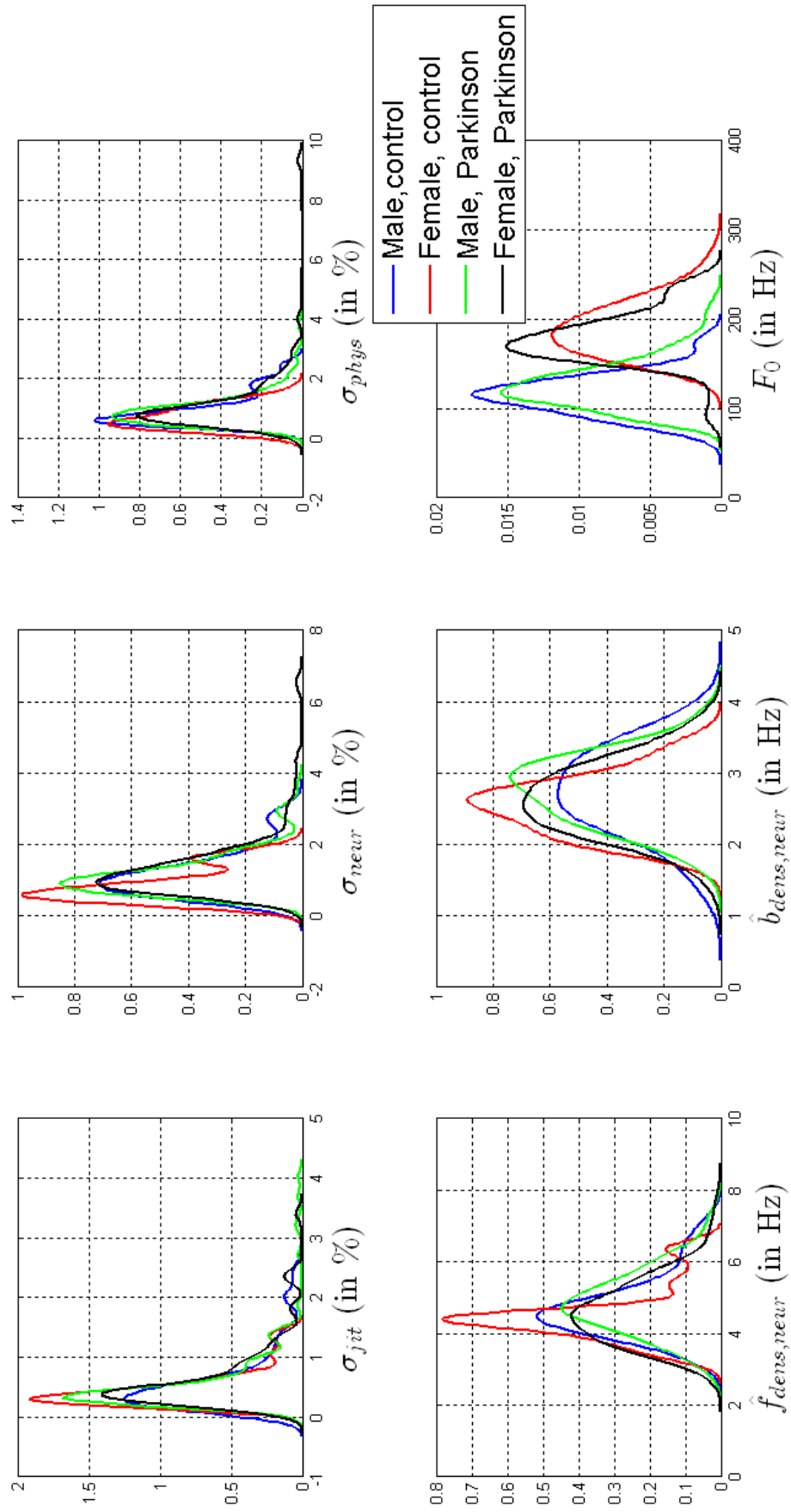


Figure 10.3 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Distributions of the vocal cues

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0	AGE	UPDRS	DUR.	HY
σ_{jit}	0.54	0.46	0.09	0.21	-0.08	0.01	0.02	-0.01	0.03
σ_{neur}		0.69	-0.06	-0.13	-0.04	0.09	0.11	0.14	0.20
σ_{phys}			-0.01	0.08	-0.15	0.02	0.02	0.11	0.09
$\hat{f}_{dens,neur}$				0.65	-0.07	-0.06	0.19	-0.02	0.06
$\hat{b}_{dens,neur}$					-0.12	-0.02	0.06	-0.06	-0.04
F_0						0.05	-0.02	-0.07	-0.01
AGE							0.00	0.24	0.32
UPDRS								0.19	0.61
DUR.									0.60

Table 10.5 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Pearson's linear correlation coefficients between the vocal perturbation cues and the patient attributes (age, UPDRS, time since diagnosis (DUR.) and HY). σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

Table 10.6 reports Pearson's linear correlation coefficients between vocal cues and age for the control speakers only. One observes that the age of the control speakers is moderately correlated with σ_{jit} , σ_{neur} , σ_{phys} and $\hat{f}_{dens,neur}$. One observes also that the perturbation sizes σ_{jit} , σ_{neur} and σ_{phys} are correlated ($r \geq 50\%$). Moreover, the neurological tremor frequency $\hat{f}_{dens,neur}$ is correlated with the neurological tremor bandwidth $\hat{b}_{dens,neur}$ ($r = 62\%$).

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0	AGE
σ_{jit}	0.50	0.50	0.25	0.35	-0.11	0.26
σ_{neur}		0.73	0.07	-0.18	-0.30	0.45
σ_{phys}			0.11	0.07	-0.28	0.48
$\hat{f}_{dens,neur}$				0.62	-0.14	0.25
$\hat{b}_{dens,neur}$					-0.18	0.14
F_0						-0.22

Table 10.6 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : Pearson's linear correlation coefficients between the vocal perturbation cues and chronological age for control speakers. σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

10.4.3 Comparison between control and Parkinson speakers

Table 10.7 reports the p-values in percent obtained via two-way and one-way ANOVA for each feature. All the p-values smaller than 5% are displayed in bold.

	Pathology (X_1)	Gender (X_2)	$X_1 \cdot X_2$
σ_{jit}	13.04	57.92	32.85
σ_{neur}	0.85	57.42	0.92
σ_{phys}	2.88	85.21	1.50
$\hat{f}_{dens,neur}$	40.22	3.66	84.93
$\hat{b}_{dens,neur}$	41.52	1.14	56.39
F_0	98.53	0.00	0.21

(a) Two-way ANOVA for each vocal cue

	Pathology	
	♂	♀
σ_{jit}	69.50	6.96
σ_{neur}	98.08	0.37
σ_{phys}	79.03	2.35
$\hat{f}_{dens,neur}$	43.12	66.62
$\hat{b}_{dens,neur}$	85.81	33.74
F_0	1.23	6.12

(b) One-way ANOVA for each vocal cue and gender

Table 10.7 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : p-values (expressed in %) obtained via two-way ANOVA (left) and one-way ANOVA (right). All the p-values smaller than 5% are displayed in bold.

One observes that the size of jitter σ_{jit} does not differ statistically significantly between control and Parkinson speakers or between male and female speakers. Neurological tremor depth σ_{neur} ($p < 1\%$) and physiological tremor depth σ_{phys} ($p < 5\%$) differ statistically significantly between control and Parkinson speakers. Neurological tremor frequency $\hat{f}_{dens,neur}$ ($p < 5\%$), neurological tremor bandwidth $\hat{b}_{dens,neur}$ ($p < 5\%$) and F_0 ($p < 0.1\%$) differ statistically significantly between male and female speakers. The interaction between variables "gender" and "pathology" is statistically significant for neurological tremor depth σ_{neur} ($p < 1\%$), physiological tremor depth σ_{phys} ($p < 5\%$) and F_0 ($p < 1\%$).

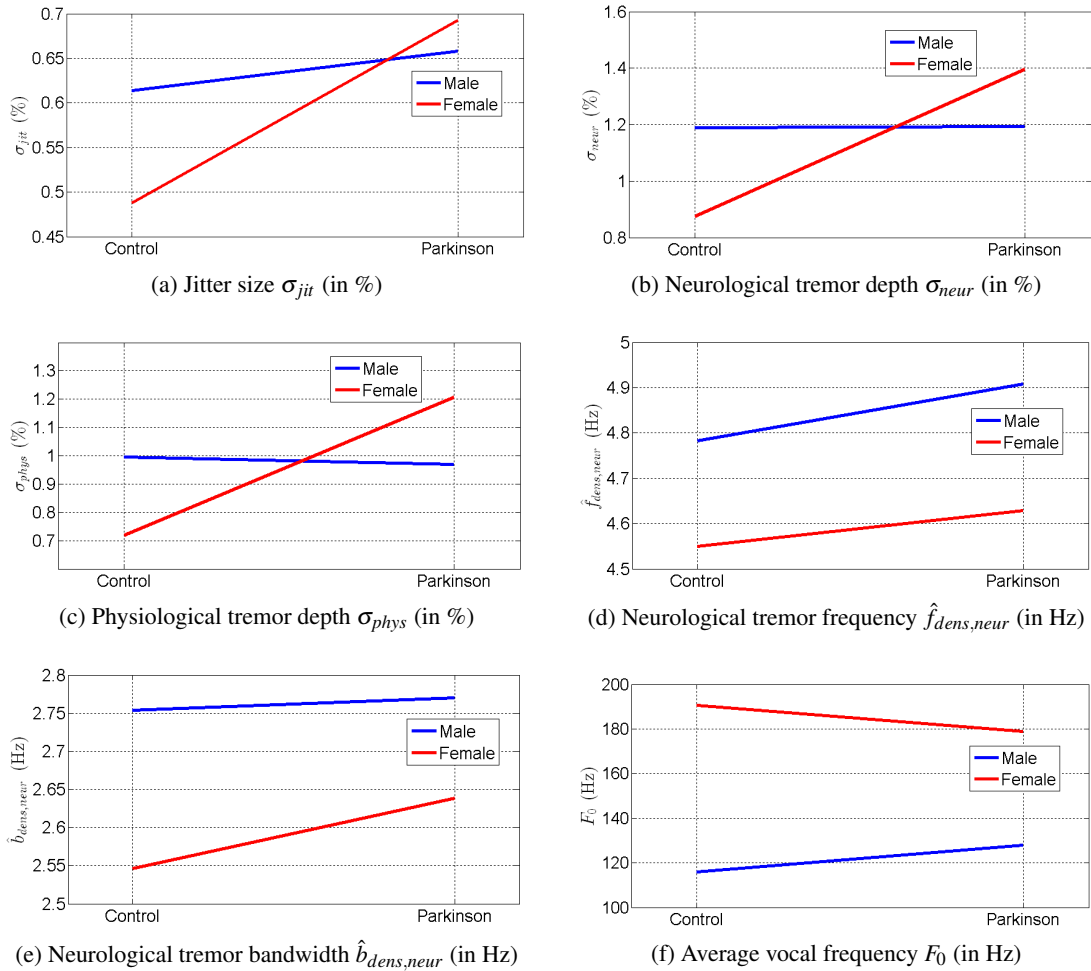


Figure 10.4 – Corpus from the Department of Neurology of Bochum University Clinic (Bochum, Germany) : variations of vocal cue averages of control and Parkinson speakers

Figure 10.4 illustrates for male (blue) and female (red) speakers the variations of vocal cue averages of control and Parkinson speakers. One observes that the neurological tremor depth σ_{neur} and physiological tremor depth σ_{phys} increase for female Parkinson speakers only (Figures 10.4b and 10.4c). The neurological tremor frequency $\hat{f}_{dens,neur}$ and its bandwidth $\hat{b}_{dens,neur}$ are statistically significantly higher for male speakers. The averages of the same cues increase for Parkinson speakers but no statistically significant difference is observed between control and Parkinson speakers (Figures 10.4d and 10.4e). Finally, the vocal frequency F_0 differs statistically significantly between male and female speakers. In addition, F_0 increases for male Parkinson speakers and decreases for female Parkinson speakers (Figure 10.4f).

10.5 Analysis of the corpus from Pays d'Aix Hospital

10.5.1 Quartiles and histograms of vocal cues

Table 10.8 reports the quartiles of the size of jitter, the neurological and physiological tremor depths (all in %). The last three lines of the table report the neurological tremor frequency and bandwidth, and the average F_0 (all in Hz). Figure 10.5 reports the histograms of the vocal cues.

		Male		Female	
		CTRL	PARK	CTRL	PARK
σ_{jit} (%)	Q_0	0.15	0.14	0.11	0.14
	Q_1	0.23	0.28	0.24	0.26
	Q_2	0.31	0.40	0.35	0.37
	Q_3	0.46	0.76	0.49	0.61
	Q_4	1.94	2.77	2.93	2.10
σ_{neur} (%)	Q_0	0.46	0.48	0.38	0.38
	Q_1	0.80	0.77	0.64	0.70
	Q_2	1.00	1.01	0.80	0.96
	Q_3	1.30	1.46	0.97	1.33
	Q_4	2.43	2.74	2.09	2.48
σ_{phys} (%)	Q_0	0.29	0.22	0.24	0.31
	Q_1	0.53	0.51	0.51	0.57
	Q_2	0.75	0.73	0.80	0.78
	Q_3	0.99	1.04	1.16	1.07
	Q_4	1.42	2.47	2.39	2.52
$\hat{f}_{dens,neur}$ (Hz)	Q_0	2.90	3.44	2.94	3.31
	Q_1	3.93	4.43	3.89	4.19
	Q_2	4.48	5.04	4.32	4.53
	Q_3	5.04	5.66	4.95	5.25
	Q_4	7.12	7.44	6.24	6.67
$\hat{b}_{dens,neur}$ (Hz)	Q_0	1.70	1.80	1.66	1.79
	Q_1	2.32	2.49	2.31	2.29
	Q_2	2.61	2.82	2.53	2.56
	Q_3	2.90	3.23	2.83	2.88
	Q_4	3.72	3.73	3.38	3.76
F_0 (Hz)	Q_0	90.09	90.31	134.82	134.99
	Q_1	105.59	113.13	163.53	169.01
	Q_2	115.84	129.57	178.77	183.39
	Q_3	129.53	147.33	196.47	200.58
	Q_4	162.78	234.76	267.95	248.30

Table 10.8 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Quartiles of the size of jitter σ_{jit} , the neurological σ_{neur} and physiological σ_{phys} tremor depths (all in %). The last three lines of the Table report the tremor frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$, and the average F_0 (all in Hz). Quartiles Q_0 to Q_4 report the values at 2.5%, 25%, 50%, 75% and 97.5% of the feature value range.

10.5.2 Correlation

Table 10.9 reports Pearson's linear correlation coefficients between the vocal perturbation cues and the patient attributes (age, dopa ON/OFF). No correlation has been observed between the perturbation cues and dopa ON or OFF. The correlation between patient age and the perturbation sizes (σ_{jit} , σ_{neur} , σ_{phys}) is feeble. One observes that the neurological tremor depth σ_{neur} is correlated with the jitter size σ_{jit} ($r = 52\%$) and the physiological tremor depth σ_{phys} ($r = 65\%$). The neurological tremor frequency $\hat{f}_{dens,neur}$ and its bandwidth $\hat{b}_{dens,neur}$ are also correlated ($r = 58\%$). The other r values are feeble.

Table 10.10 reports Pearson's linear correlation coefficients between vocal cues and age for control speakers only. No large correlation is observed, except between the neurological tremor

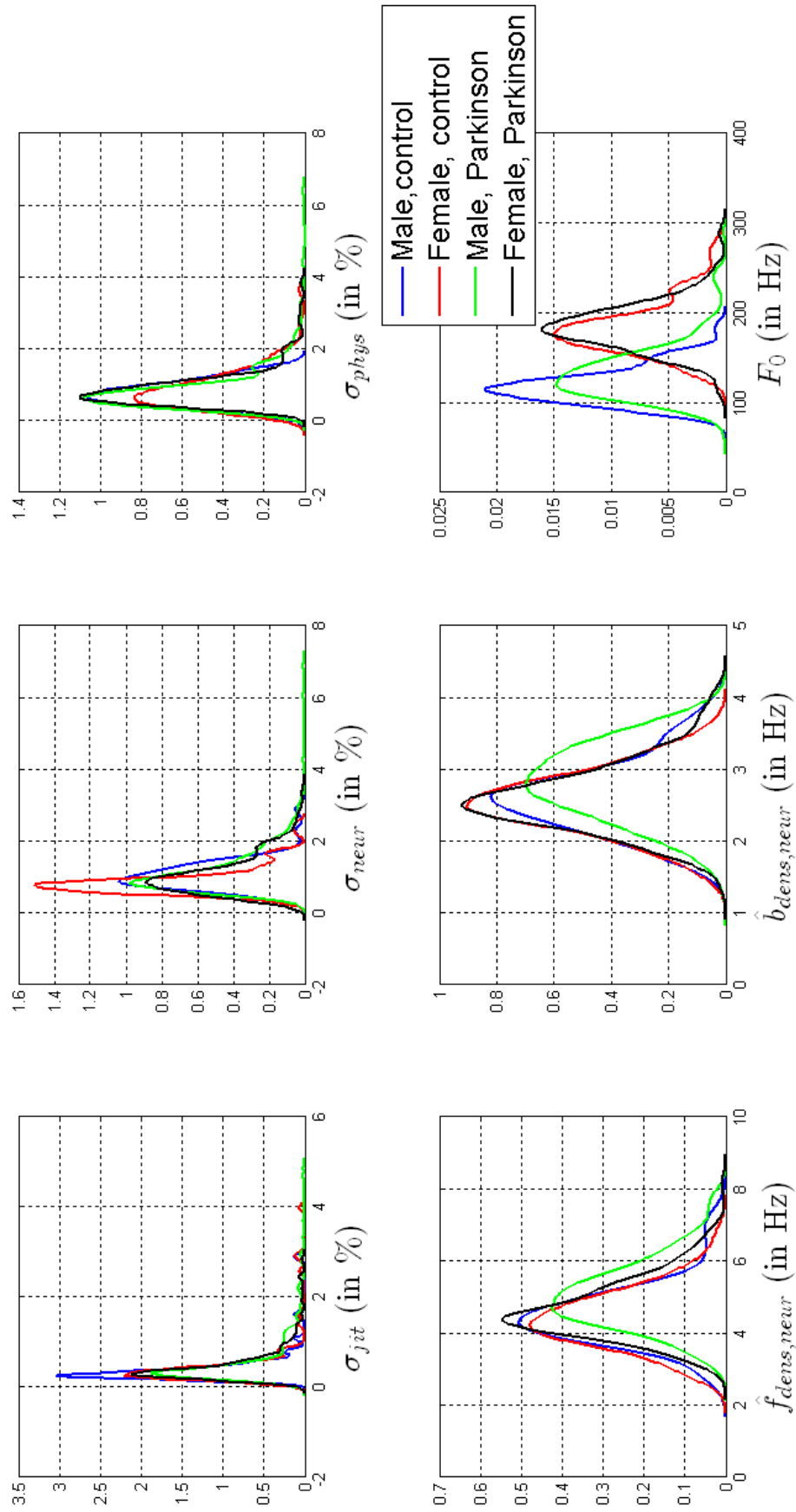


Figure 10.5 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Distributions of the vocal cues.

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0	AGE	DOPA
σ_{jit}	0.52	0.26	0.12	0.17	-0.13	0.13	-0.04
σ_{neur}		0.65	-0.09	-0.14	-0.05	0.25	-0.06
σ_{phys}			-0.08	-0.09	0.03	0.24	0.03
$\hat{f}_{dens,neur}$				0.58	-0.16	0.07	-0.05
$\hat{b}_{dens,neur}$					-0.14	-0.06	0.04
F_0						0.01	0.04
AGE							-0.01

Table 10.9 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Pearson's linear correlation coefficients between the vocal perturbation cues and the patient attributes (age, dopa ON/OFF). σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

frequency $\hat{f}_{dens,neur}$ and the neurological tremor bandwidth $\hat{b}_{dens,neur}$ ($r = 62\%$).

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0	AGE
σ_{jit}	0.36	0.09	0.21	0.26	0.05	0.05
σ_{neur}		0.22	-0.12	-0.21	-0.20	0.23
σ_{phys}			-0.07	-0.10	-0.02	0.16
$\hat{f}_{dens,neur}$				0.62	-0.02	0.01
$\hat{b}_{dens,neur}$					-0.02	-0.01
F_0						-0.04

Table 10.10 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : Pearson's linear correlation coefficients between the vocal perturbation cues and age for control speakers. σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

10.5.3 Comparison between control and Parkinson speakers

Table 10.11 reports the p-values obtained via two-way and one-way ANOVA for each feature. All the p-values smaller than 5% are displayed in bold. One observes that the physiological tremor depth σ_{phys} does not differ statistically significantly between control and Parkinson speakers or between male and female speakers. Neurological tremor depth σ_{neur} , frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$ as well as the average vocal frequency F_0 differ statistically significantly between control and Parkinson speakers and between male and female speakers. The interaction between variables "gender" and "pathology" is statistically significant for jitter size σ_{jit} ($p < 5\%$) and F_0 ($p < 5\%$).

	Pathology (X_1)	Gender (X_2)	$X_1 \cdot X_2$
σ_{jit}	12.54	85.04	4.79
σ_{neur}	1.06	1.87	62.23
σ_{phys}	32.56	23.48	32.74
$\hat{f}_{dens,neur}$	0.00	0.46	18.22
$\hat{b}_{dens,neur}$	0.72	0.39	16.04
F_0	1.75	0.00	1.95

(a) Two-way ANOVA for each vocal cue

(b) One-way ANOVA for each vocal cue and gender

Table 10.11 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : p-values (expressed in %) obtained via two-way ANOVA (left) and one-way ANOVA (right). All the p-values smaller than 5% are displayed in bold.

Figure 10.6 illustrates for male (blue) and female (red) speakers the variations of vocal cue

averages of control and Parkinson speakers. One observes that the neurological tremor depth σ_{neur} , frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$ are statistically significantly higher for male speakers and for Parkinson speakers (Figures 10.6b, 10.6d and 10.6e). Finally, one observes that the average vocal frequency F_0 increases for male Parkinson speakers.

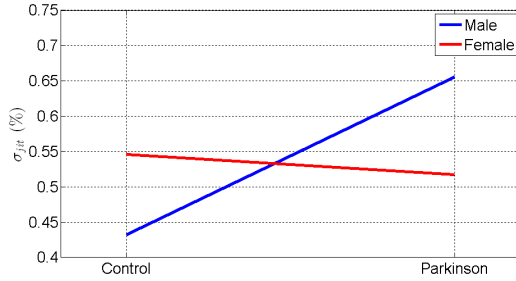
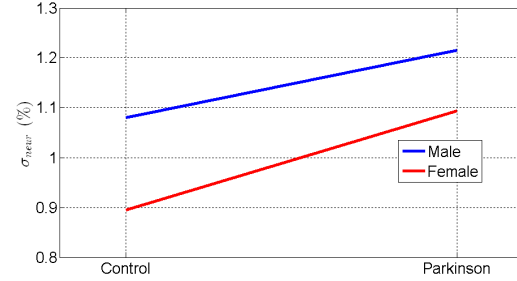
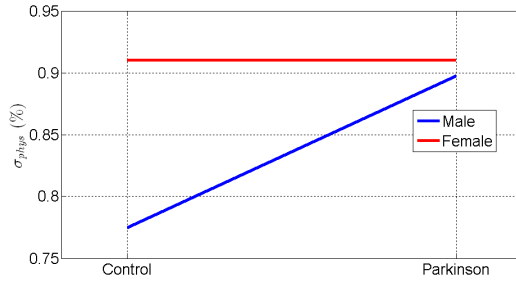
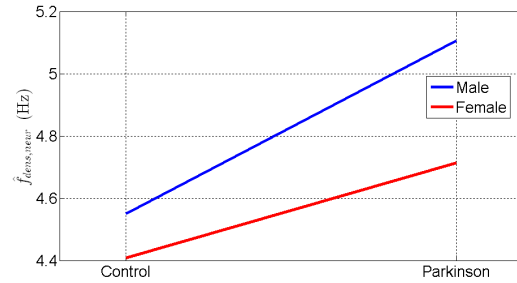
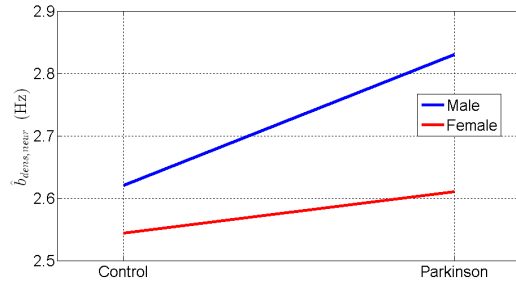
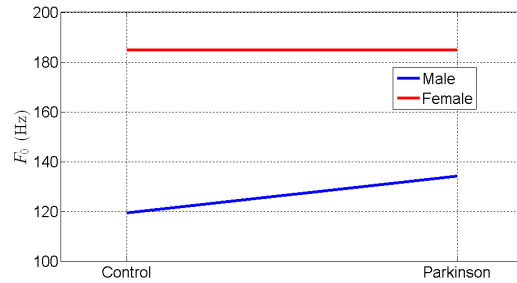
(a) Jitter size σ_{jit} (in %)(b) Neurological tremor depth σ_{neur} (in %)(c) Physiological tremor depth σ_{phys} (in %)(d) Neurological tremor frequency $\hat{f}_{dens,neur}$ (in Hz)(e) Neurological tremor bandwidth $\hat{b}_{dens,neur}$ (in Hz)(f) Average vocal frequency F_0 (in Hz)

Figure 10.6 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : variations of vocal cue averages of control and Parkinson speakers

10.5.4 Effects of the use of medication

A two-way ANOVA has been carried out to determine the effects of gender and medication (L-DOPA). Here, only Parkinson data have been considered. Table 10.12a reports the p-values for each feature. Additionally, a Wilcoxon matched pairs signed ranks test [She03] has been carried out considering a subset of patients for whom recordings with or without medication are available. Table 10.12b reports the p-values for each feature and gender. P-values smaller than 5% are displayed in bold.

	Dopa (X_1)	Gender (X_2)	$X_1 \cdot X_2$
σ_{jit}	40.82	3.25	80.39
σ_{neur}	11.06	8.93	29.58
σ_{phys}	64.12	86.43	90.86
$\hat{f}_{dens,neur}$	43.20	0.01	70.62
$\hat{b}_{dens,neur}$	39.96	0.01	79.39
F_0	30.58	0.00	94.64

(a) Two-way ANOVA for each vocal cue

	Dopa	
	♂	♀
σ_{jit}	46.61	3.56
σ_{neur}	3.91	0.10
σ_{phys}	92.22	66.03
$\hat{f}_{dens,neur}$	11.41	76.73
$\hat{b}_{dens,neur}$	69.85	8.67
F_0	6.03	1.41

(b) Wilcoxon matched pairs signed ranks test for each vocal cue and gender

Table 10.12 – Corpus from the Neurology Department of Pays d’Aix Hospital (Aix-en-Provence, France) : p-values (in %) obtained via two-way ANOVA (left) and Wilcoxon matched pairs signed ranks test (right). All the p-values smaller than 5% are displayed in bold.

For the first statistical test (ANOVA), no statistically significant differences are observed with regard to the use of L-DOPA. However, considering a subset of patients for whom recordings with or without medication are available, statistically significant differences between patients with or without medication are observed for jitter size σ_{jit} (female), neurological tremor depth σ_{neur} (male and female) and average vocal frequency F_0 (female). Figure 10.7 illustrates for male (blue) and female (red) speakers the variations of vocal cue averages of DOPA-ON and DOPA-OFF Parkinson speakers. Notice that these differences are feeble, but they may decrease the statistical effect of the pathology when DOPA-ON and DOPA-OFF recordings are pooled (section 10.5.3).

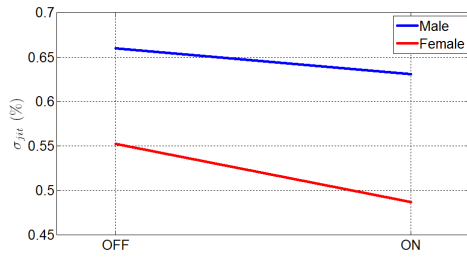
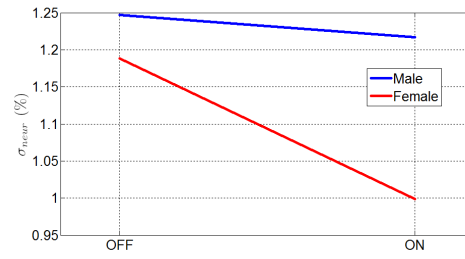
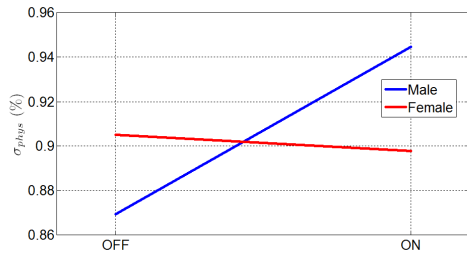
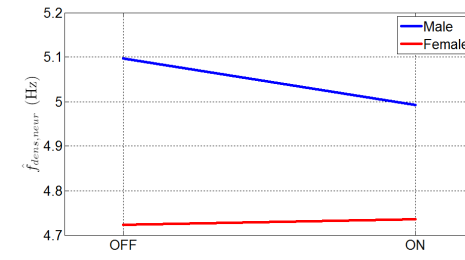
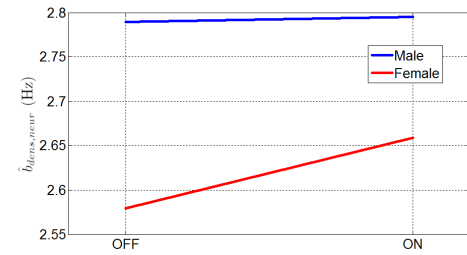
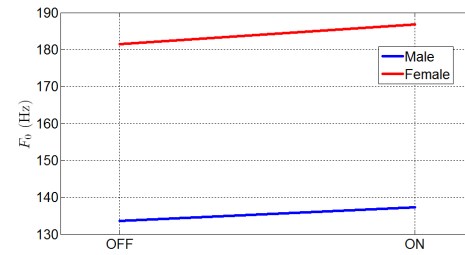
(a) Jitter size σ_{jit} (in %)(b) Neurological tremor depth σ_{neur} (in %)(c) Physiological tremor depth σ_{phys} (in %)(d) Neurological tremor frequency $\hat{f}_{dens,neur}$ (in Hz)(e) Neurological tremor bandwidth $\hat{b}_{dens,neur}$ (in Hz)(f) Average vocal frequency F_0 (in Hz)

Figure 10.7 – Corpus from the Neurology Department of Pays d'Aix Hospital (Aix-en-Provence, France) : vocal cue averages of DOPA-ON and DOPA-OFF Parkinson speakers

10.6 Comparison between corpora, as well as pooled control and Parkinson speakers

The two corpora are pooled. The number of pooled control speaker recordings has been 197 (92 ♂, 105 ♀) and the number of pooled Parkinson speaker recordings has been 661 (431 ♂, 230 ♀).

10.6.1 Quartiles and histograms of vocal cues

Table 10.13 reports the quartiles of the size of jitter, the neurological and physiological tremor depths (all in %). The last three lines of the table report the neurological tremor frequency and bandwidth, and the average F_0 (all in Hz). Figure 10.8 reports the histograms of the vocal cues.

		Male		Female	
		CTRL	PARK	CTRL	PARK
σ_{jit} (%)	Q_0	0.14	0.14	0.13	0.14
	Q_1	0.23	0.29	0.24	0.28
	Q_2	0.35	0.40	0.35	0.41
	Q_3	0.57	0.76	0.55	0.65
	Q_4	2.11	2.88	2.85	2.35
σ_{neur} (%)	Q_0	0.48	0.50	0.33	0.41
	Q_1	0.76	0.77	0.59	0.76
	Q_2	1.03	1.02	0.79	1.00
	Q_3	1.37	1.44	1.00	1.44
	Q_4	2.76	2.92	1.88	3.02
σ_{phys} (%)	Q_0	0.30	0.25	0.23	0.30
	Q_1	0.55	0.53	0.48	0.59
	Q_2	0.77	0.75	0.78	0.81
	Q_3	1.03	1.06	1.14	1.13
	Q_4	2.12	2.46	1.93	2.91
$\hat{f}_{dens,neur}$ (Hz)	Q_0	3.11	3.40	2.97	3.31
	Q_1	4.04	4.42	3.93	4.11
	Q_2	4.51	4.96	4.38	4.54
	Q_3	5.06	5.58	4.94	5.22
	Q_4	7.12	7.34	6.41	6.70
$\hat{b}_{dens,neur}$ (Hz)	Q_0	1.67	1.83	1.76	1.79
	Q_1	2.33	2.46	2.29	2.28
	Q_2	2.66	2.81	2.53	2.58
	Q_3	3.02	3.19	2.82	2.91
	Q_4	3.72	3.73	3.35	3.69
F_0 (Hz)	Q_0	83.76	87.70	135.48	123.05
	Q_1	104.20	112.07	163.72	165.13
	Q_2	115.80	128.34	184.41	179.98
	Q_3	129.36	146.11	200.61	198.43
	Q_4	173.60	220.38	265.76	243.78

Table 10.13 – Pooled data : Quartiles of the size of jitter σ_{jit} , the neurological σ_{neur} and physiological σ_{phys} tremor depths (all in %). The last three lines of the Table report the tremor frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$, and the average F_0 (all in Hz). Quartiles Q_0 to Q_4 report the values at 2.5%, 25%, 50%, 75% and 97.5% of the feature value range.

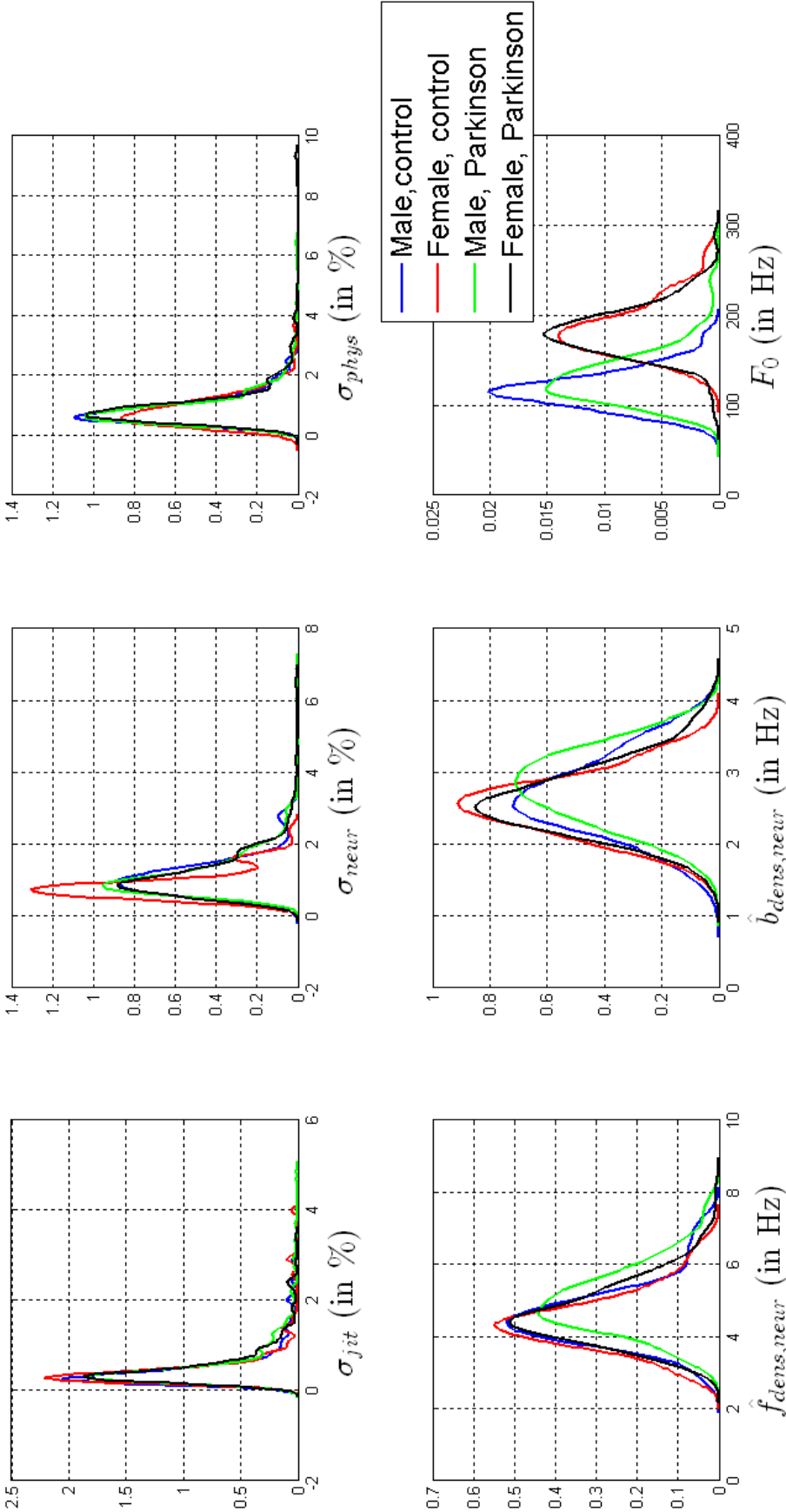


Figure 10.8 – Pooled data : Distributions of the vocal cues.

10.6.2 Correlations

Table 10.14 and 10.15 report Pearson's linear correlation coefficients between the vocal perturbation cues for the control and Parkinson speakers respectively.

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0
σ_{jit}	0.52	0.33	0.13	0.18	-0.12
σ_{neur}		0.66	-0.08	-0.13	-0.05
σ_{phys}			-0.05	-0.02	-0.04
$\hat{f}_{dens,neur}$				0.60	-0.15
$\hat{b}_{dens,neur}$					-0.15

Table 10.14 – Pooled corpora : Pearson's linear correlation coefficients between the vocal perturbation cues of the Parkinson speakers. σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

	σ_{neur}	σ_{phys}	$\hat{f}_{dens,neur}$	$\hat{b}_{dens,neur}$	F_0
σ_{jit}	0.40	0.22	0.22	0.29	-0.01
σ_{neur}		0.45	-0.02	-0.19	-0.25
σ_{phys}			0.00	-0.03	-0.13
$\hat{f}_{dens,neur}$				0.62	-0.08
$\hat{b}_{dens,neur}$					-0.09

Table 10.15 – Pooled corpora : Pearson's linear correlation coefficients between the vocal perturbation cues of the control speakers. σ_{jit} , σ_{neur} and σ_{phys} designate vocal jitter size, neurological tremor depth and physiological tremor depth. $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ refer to the frequency and the bandwidth of the neurological tremor

10.6.3 Three-way variance analysis

A three-way ANOVA is carried out to compare the results obtained for individual corpora. The first explanatory variable, denoted "corpus", has two levels : *Bochum* and *Aix-en-Provence*. The two other variables are "pathology" (with levels *Control* and *Parkinson*) and "gender" (with levels *Male* and *Female*). Table 10.16 reports the p-values obtained via three-way, two-way and one-way ANOVA for each vocal cue. All the p-values smaller than 5% are displayed in bold.

One observes that no statistically significant differences between corpora have been observed and that the interaction between explanatory variables "corpus" and "gender" is not significant. Interaction between variables "corpus" and "pathology" is statistically significant for neurological tremor frequency $\hat{f}_{dens,neur}$ only. Moreover, one observes that the interactions between all explanatory variables are statistically significant for all perturbation sizes.

One observes also that all vocal cues, except the average vocal frequency F_0 , differ statistically significantly between control and Parkinson speakers. The neurological tremor frequency $\hat{f}_{dens,neur}$, bandwidth $\hat{b}_{dens,neur}$ and average vocal frequency F_0 also differ statistically significantly between male and female speakers. In addition, interaction between explanatory variables "pathology" and "gender" are statistically significant for neurological tremor depth σ_{neur} and average vocal frequency F_0 .

	Corpus (X_1)	Pathology (X_2)	Gender (X_3)	$X_1 \cdot X_2$	$X_1 \cdot X_3$	$X_2 \cdot X_3$	$X_1 \cdot X_2 \cdot X_3$
σ_{jit}	15.06	3.39	58.19	79.52	74.78	66.33	4.87
σ_{neur}	10.22	0.02	6.52	40.51	38.64	1.04	4.59
σ_{phys}	8.04	1.00	63.14	13.35	40.86	8.29	0.50
$\hat{f}_{dens,neur}$	77.70	0.06	0.07	3.44	94.59	33.73	50.94
$\hat{b}_{dens,neur}$	54.30	2.29	0.02	32.18	79.93	69.22	19.12
F_0	32.25	13.82	0.00	14.58	35.26	0.02	36.40

(a) Three-way ANOVA for each vocal cue

	Pathology (X_1)	Gender (X_2)	$X_1 \cdot X_2$
σ_{jit}	5.90	49.83	34.56
σ_{neur}	0.04	1.81	3.67
σ_{phys}	6.59	54.64	29.54
$\hat{f}_{dens,neur}$	0.00	0.01	28.89
$\hat{b}_{dens,neur}$	1.05	0.00	48.67
F_0	2.57	0.00	0.02

	Pathology	
	♂	♀
σ_{jit}	5.76	45.53
σ_{neur}	31.06	0.01
σ_{phys}	53.31	7.37
$\hat{f}_{dens,neur}$	0.03	1.92
$\hat{b}_{dens,neur}$	2.41	16.95
F_0	0.00	28.76

(b) Two-way ANOVA for each vocal cue

(c) One-way ANOVA for each vocal cue and gender

Table 10.16 – Pooled corpora : p-values (expressed in %) obtained via three-way ANOVA (top), two-way ANOVA (bottom, left) and one-way ANOVA (bottom, right). All the p-values smaller than 5% are displayed in bold.

10.6.4 Perturbation size

Figure 10.9 reports the variations of the perturbation size averages for the two corpora (left column) or for control and Parkinson speakers (right column). In Figures 10.9a, 10.9c and 10.9e (left column), these variations are reported for control speakers (black) and Parkinson speakers (red). Figures 10.9b, 10.9d and 10.9f (right column) report the variations of these perturbation size averages for the corpora from Bochum (black) and Aix-en-Provence (red).

10.6.4.1 Jitter size σ_{jit}

Previous analyses have shown that no statistically significant differences were observed for σ_{jit} between male and female speakers or between control and Parkinson speakers for the two corpora. But a statistically significant interaction was observed between explanatory variables "pathology" and "gender" for Aix-en-Provence speakers. A statistically significant difference is observed however between the pooled control and Parkinson speakers. The reason is that jitter size σ_{jit} increases for Bochum female Parkinson speakers and for Aix-en-Provence male Parkinson speakers (Figure 10.9b). The large increase of F_0 of male Parkinson speakers at Aix-Hospital favours an increase of their vocal jitter size.

10.6.4.2 Neurological tremor depth σ_{neur}

Neurological tremor differed statistically significantly between control and Parkinson speakers for the 2 corpora. The same statistically significant difference is observed for the pooled control and Parkinson speakers. The statistically significant difference between Aix-en-Provence male and female speakers is not confirmed by the pooled male and female speakers. However, the statistically significant interaction between explanatory variables "pathology" and "gender" for the Bochum corpus is also statistically significant for the pooled data. Indeed, increasing neurological tremor depth σ_{neur} is observed for female Parkinson speakers (Figure 10.9d) but not for male Parkinson speakers.

In addition, one observes that neurological tremor depth σ_{neur} of female Parkinson speakers differ between the two corpora (Figure 10.9c).

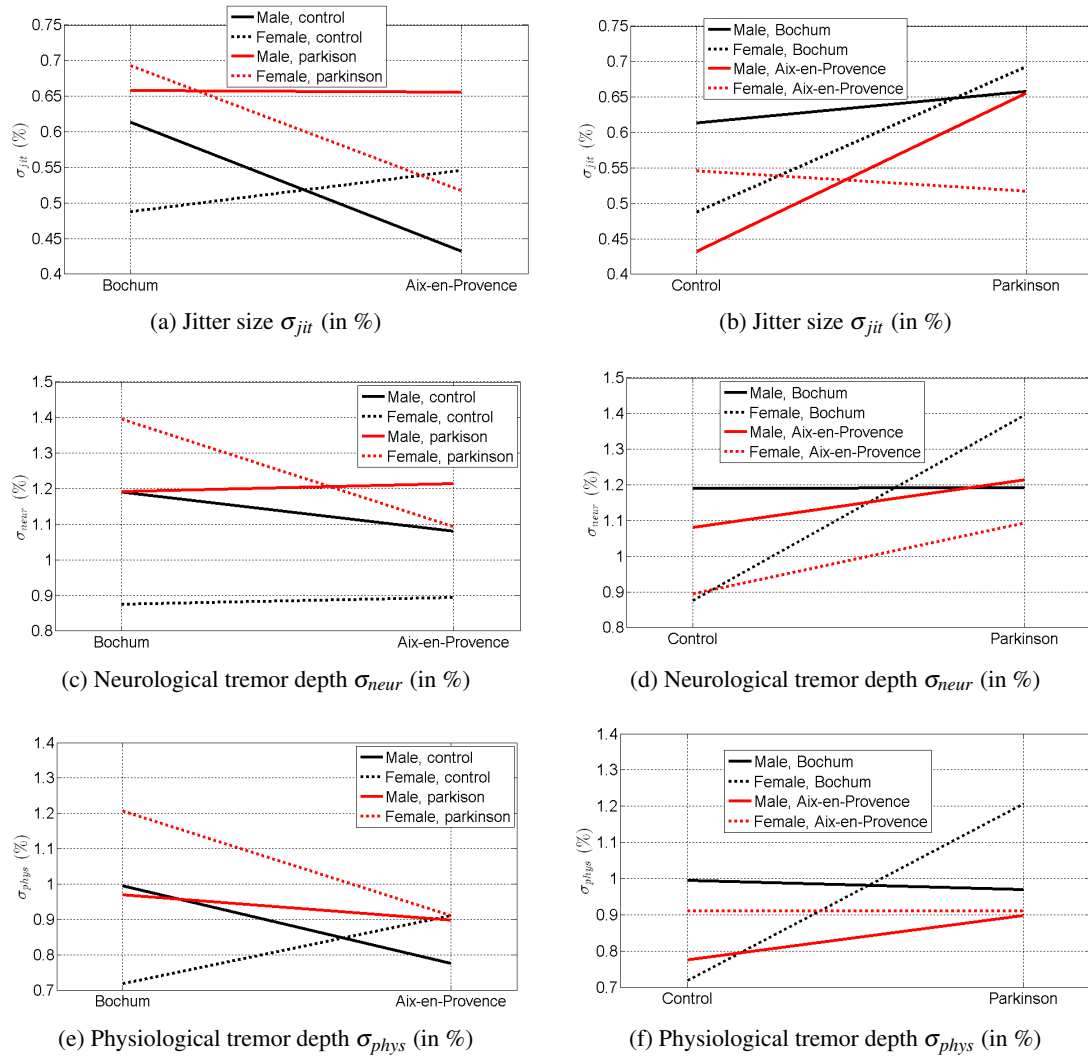


Figure 10.9 – Corpora : Variations of perturbation size averages for the two corpora (left column) and for control and Parkinson speakers (right column)

10.6.4.3 Physiological tremor depth σ_{phys}

The physiological tremor depth σ_{phys} difference between control and Parkinson speakers, as well as the interaction between explanatory variables "pathology" and "gender" were statistically significant for the Bochum corpus only. Physiological tremor depth σ_{phys} differs statistically significantly between pooled control and Parkinson speakers. The reason is that the physiological tremor depth σ_{phys} difference between Bochum female control and Parkinson speakers is large (Figure 10.9f). In addition, no interaction between "pathology" and "gender" is observed.

10.6.5 Neurological tremor frequency and bandwidth

Figure 10.10 illustrates the variations of neurological tremor frequency and bandwidth averages for the two corpora (left column) or for control and Parkinson speakers (right column).

In Figures 10.10a and 10.10c (left column), these variations are reported for control speakers (black) and Parkinson speakers (red). Figures 10.10b and 10.10d (right column) report the variations of neurological tremor frequency and bandwidth averages for the corpora from Bochum (black) and Aix-en-Provence (red).

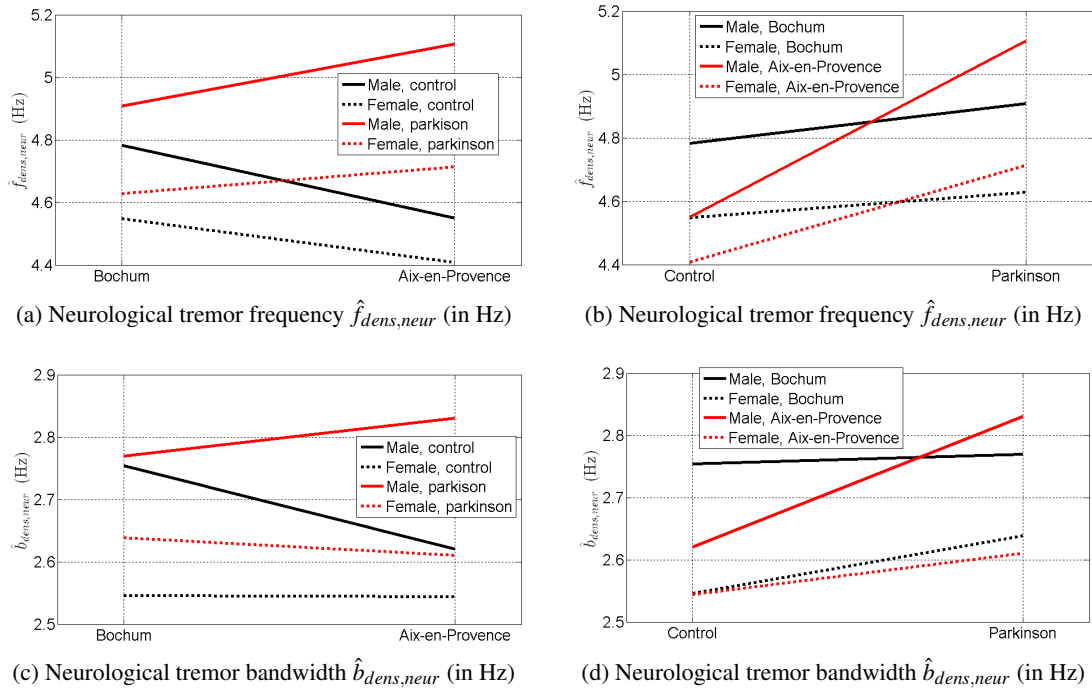


Figure 10.10 – Corpora : Variations of neurological tremor frequency and bandwidth averages for the two corpora (left column) and for control and Parkinson speakers (right column)

Neurological tremor frequency $\hat{f}_{dens,neur}$ and bandwidth $\hat{b}_{dens,neur}$ differed statistically significantly between male and female speakers for the two corpora and between Aix-en-Provence control and Parkinson speakers. These two cues also differ statistically significantly between pooled control and Parkinson speakers and between pooled male and female speakers. The reason is that the pooled Parkinson speakers are characterized by higher neurological tremor frequency and bandwidth than the pooled control speakers (Figure 10.10b and 10.10d), and that higher neurological frequencies and bandwidths are observed in male speakers (Figure 10.10a and 10.10c). In addition, the interaction between explanatory variables "corpus" and "pathology" is statistically significant for the neurological tremor frequency $\hat{f}_{dens,neur}$. The reason is that one observes in Figure 10.10a

that the neurological frequency $\hat{f}_{dens,neur}$ decreases for control speakers and increases for Parkinson speakers from one corpus to the other.

10.6.6 Vocal frequency F_0

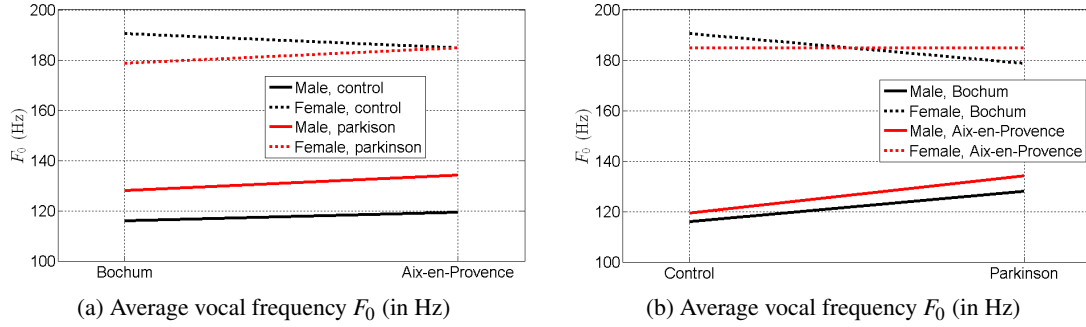


Figure 10.11 – Corpora : Variations of vocal frequency averages for the two corpora (left column) and for control and Parkinson speakers (right column)

Figure 10.11 illustrates the variations of vocal frequency averages for the two corpora (left column) or for control and Parkinson speakers (right column).

In Figure 10.11a, these variations are reported for control speakers (black) and Parkinson speakers (red). Figure 10.11b illustrates the variations of F_0 averages for the corpora from Bochum (black) and Aix-en-Provence (red).

One is relieved to observe that the vocal frequency F_0 is statistically significantly higher for female speakers ! No statistically significant difference is observed between pooled control and Parkinson speakers but the interaction between explanatory variables "pathology" and "gender" is statistically significant. The reason is that the average vocal frequency decreases for female Parkinson speakers and increases for male Parkinson speakers.

10.7 Discussion and conclusion

10.7.1 Discrepancies between corpora

Descriptive analysis shows that the distributions of chronological age differ between corpora, especially for young control speakers that are more numerous in the corpus from Bochum even though the medians of the Bochum control and Parkinson speakers are close. On the other hand, the control and Parkinson speakers from Aix-en-Provence have quasi-identical age distributions.

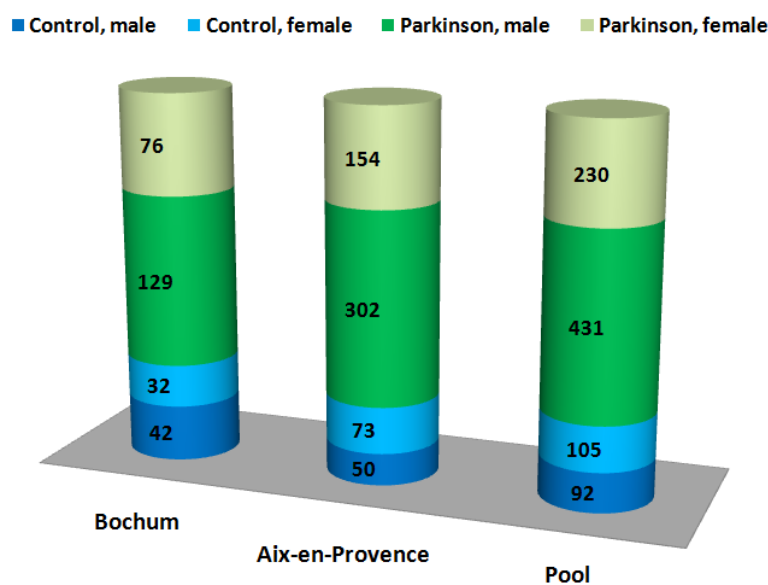


Figure 10.12 – Distribution of male, female, control and Parkinson speaker recordings

Another issue is related to the unequal distribution of male and female speakers between corpora or between control and Parkinson speakers (Figure 10.12). First, the percentages of control and Parkinson's speakers are approximatively identical for the two corpora and the number of male Parkinson speakers is always higher than the number of female Parkinson speakers. However, one observes different distributions of male and female control speakers, so that the Bochum corpus comprises more male than female speakers and inversely for the Aix-en-Provence corpus. Finally, the pooled control speakers comprise approximatively equal percentages of male and female speakers.

10.7.2 Patient attributes

Pearson's linear correlation analysis indicates that no significant correlation is observed between the vocal cues and patient attributes. A possible explanation is that *UPDRS* or *HY* scores describe global symptoms of Parkinson's disease and its progress and are therefore not focused on the assessment of a patient's voice.

10.7.3 Statistically significant effects of corpus, gender or pathology

10.7.3.1 Corpus

Even though the corpora have different gender distributions, no statistically significant main effects are assigned to the corpora.

10.7.3.2 Gender

The results of the two-way analysis of variance with explanatory variables "pathology" and "gender" have shown that the neurological tremor frequency $\hat{f}_{dens,neur}$, its bandwidth $\hat{b}_{dens,neur}$ and the average vocal frequency F_0 differ always statistically significantly between male and female speakers. Indeed, one observes that $\hat{f}_{dens,neur}$ and $\hat{b}_{dens,neur}$ averages are systematically larger for male speakers than female speakers. Inversely, the vocal frequency F_0 averages are systematically higher for female speakers than for male speakers.

These statistically significant differences of $\hat{b}_{dens,neur}$ between male and female speakers may explain the different correlation coefficients that are observed between the neurological tremor depth σ_{neur} and the physiological tremor depth σ_{phys} and jitter size σ_{jit} . Indeed, for the control speakers from Bochum, the correlation coefficient are high ($r = 50\%$ between σ_{neur} and σ_{jit} , and $r = 73\%$ between σ_{neur} and σ_{phys}) compared to the control speakers from Aix-en-Provence ($r \leq 36\%$). A possible explanation is that female control speakers are more numerous than male control speakers for the corpus from Aix-en-Provence. The $\hat{b}_{dens,neur}$ ensemble average is thus small because female control speakers are characterized by smaller bandwidths than male control speakers. As a consequence, the spectral energy is less dispersed around the neurological tremor frequency $\hat{f}_{dens,neur}$ so that the categorization of cycle length perturbations into physiological tremor, neurological tremor and jitter is less fuzzy. Therefore, assuming that the vocal jitter, the neurological tremor and physiological tremor cues report the symptoms of different physiological or neurological causes, no significant correlations are observed between perturbation size cues. Inversely, male control speakers are more numerous than female control speakers for the corpus from Bochum. The $\hat{b}_{dens,neur}$ ensemble average is thus higher than for the Aix-en-Provence corpus. As a consequence, the neurological tremor frequency density is more dispersed which suggests that more spectral overlap occurs with the vocal jitter and physiological tremor categories. Increasing amount of neurological tremor depth therefore affects positively the jitter size and the physiological tremor depth. Therefore, a correlation between perturbation sizes is observed.

The same reasoning may be applied to Parkinson speakers for individual or pooled corpora because for these speakers, male or female, the bandwidth $\hat{b}_{dens,neur}$ is larger and, anyway, male Parkinson speakers are far more numerous than female Parkinson speakers. Finally, moderate correlation coefficients between perturbation sizes for pooled control speakers are observed because the number of male and female control speakers are similar.

10.7.3.3 Pathology

The results of the two-way variance analysis with explanatory variables "pathology" and "gender" show that the neurological tremor depth differs statistically significantly between control and Parkinson speakers for individual or pooled corpora. Indeed, Parkinson speakers are characterized by higher neurological tremor depth than control speakers. Statistically significant interaction between variables "pathology" and "gender" suggests that the neurological tremor depth σ_{neur} in male and female does not evolve identically with pathology. Indeed, the difference between control and Parkinson speakers is higher for female speakers than for male speakers.

Physiological tremor depth σ_{phys} and jitter size σ_{jit} differ statistically significantly between pooled control and Parkinson speakers, but these differences are not significant for individual corpora. This observation is confirmed by the statistically significant interaction between variables "corpus", "pathology" and "gender". That triple interaction, in the absence of dual interaction, suggests that the statistical significance of the difference between control and Parkinson speakers is corpus-dependent and therefore not likely to be generally valid. In addition, when repeating the ANOVA analysis of the pooled corpus while dropping explanatory variable "corpus", the statistically significant difference between control and Parkinson speakers disappears for physiological

tremor depth and jitter.

Neurological tremor frequency $\hat{f}_{dens,neur}$ differs statistically significantly between control and Parkinson speakers for the corpus from Aix-en-Provence and for the pooled data, for which a statistically significant interaction is observed between variables "gender" and "pathology".


The neurological tremor bandwidth $\hat{b}_{dens,neur}$ is correlated with the neurological tremor frequency $\hat{f}_{dens,neur}$. Statistically significant differences between control and Parkinson speakers and male speakers are observed.

10.7.3.4 Effect of gender and pathology on vocal frequency F_0

A statistically significant interaction between the variable "pathology" and "gender" is observed for vocal frequency F_0 . Indeed, the vocal frequency of male Parkinson speakers increases while the vocal frequency of female Parkinson speakers decreases. As a consequence, the perturbation size cues σ_{phys} , σ_{neur} and σ_{jit} , expressed in % and normalized by T_0 , are amplified for male Parkinson ($F_0 \nearrow \Rightarrow T_0 \searrow \Rightarrow \frac{1}{T_0} \nearrow$) and attenuated for female speakers ($F_0 \searrow \Rightarrow T_0 \nearrow \Rightarrow \frac{1}{T_0} \searrow$). Consequently, the statistically significant differences of neurological tremor depth σ_{neur} between female control and Parkinson speakers cannot be attributed to the effect of F_0 on the depth cues.

Key points

- The neurological tremor depth is statistically significantly higher for female Parkinson speakers than for female control speakers
- The neurological tremor frequency differs statistically significantly between male and female speakers and increases statistically significantly for male Parkinson speakers compared to male control speakers
- The vocal frequency increases for male Parkinson speakers and decreases for female Parkinson speakers



11. Conclusions & perspectives

Objectives of this chapter

- Remind the motivations and objectives of the study
- State the key results
- Discuss improvements and perspectives

Contents

11.1	Objectives and motivations	215
11.2	Key results	215
11.3	Improvements & perspectives	216

11.1 Objectives and motivations

The general framework of the thesis is the assessment of disordered voices. The assessment of voice and laryngeal function is based on auditory ratings and acoustic analyses of speech sounds. Acoustic feature-based assessment methods are indeed popular because they are non-invasive and enable clinicians to monitor the voice of patients quantitatively.

The goal of this study is the analysis of vocal tremor and vocal jitter in Parkinson speakers and control speakers. Few studies have, indeed, been devoted to vocal tremor in human speakers in general and Parkinson speakers in particular. Also, a large majority of the existing studies have involved small corpora with tens of speakers at most. Idem, no studies have addressed jointly vocal jitter and vocal tremor (neurological & physiological) as well as declination.

Here, cycle length jitter and vocal tremor frequencies and depths are obtained for two corpora of vowels comprising 123 and 74 control and 456 and 205 Parkinson speaker recordings respectively.

11.2 Key results

1. The cycle length tracking based on speech cycle peak salience and dynamic programming is able to track reliably vocal cycle lengths without strong a priori assumption with regard to cycle length regularity and cycle length intervals.
2. The proposed method is able to track reliably cycle lengths in the presence of low-frequency modulations up to 10% and fast perturbations up to 4%, over the whole range of vocal frequencies. Reliable cycle length sequences are obtained for signal-to-noise ratios higher than 15dB.
3. The perturbation analysis is based on empirical mode decomposition to split the cycle length time series into temporal sub-series corresponding to jitter, neurological tremor, physiological tremor and trend. Instantaneous frequencies of individual modes are obtained via AM-FM decomposition and the weighted category average is defined in the complex plane to decrease the effect of mode mixing.
4. The vocal cycle length perturbation analysis has been validated by means of synthetic stimuli and multiple linear regression. The validation stage suggest discarding several cues because they do not report the reference parameter values adequately.
5. Two corpora of Parkinson and control speakers have been analyzed. The results show that the neurological tremor modulation depth σ_{neur} is statistically significantly higher for female Parkinson speakers than for female control speakers.
6. Neurological tremor frequency differs statistically significantly for male and female speakers.
7. Neurological tremor frequency increases statistically significantly for the pooled Parkinson speakers compared to the pooled control speakers.
8. A statistically significant interaction between explanatory variables "gender" and "pathology" is observed for vocal frequency F_0 , which increases for male Parkinson speakers and decreases for female Parkinson speakers, compared to control speakers.

11.3 Improvements & perspectives

In this study, a method for the tracking of vocal cycle lengths in sustained voiced speech sounds is proposed. It relies on a sample salience analysis and the cycle length sequence is obtained via dynamic programming. One asset is that no strong a priori assumptions are made with regard to the cycle length regularity or range. Currently, the analysis of a fragment of a sustained speech sound that has been selected manually is carried out. A possible improvement would be the implementation of an automatic voice activity detector to avoid time-consuming and repetitive segmentation tasks. Ideally, this voice activity detector should be able to detect and isolate a fragment of speech sound which is not contaminated by environmental noise or voice breaks.

The cycle length perturbation analysis has been carried out to assess the voice quality of control speakers and patients that suffers from Parkinson's disease. The method could be applied to other pathologies involving vocal tremor like *essential tremor*, *amyotrophic lateral sclerosis* or *spasmodic dysphonia*.

The *salience-based cycle length tracking* may also be used in other applications involving the tracking of pseudo-periodic cycle patterns. As an example, three unexpected but interesting applications of the developed sample salience analysis have been discovered while browsing the Internet. The applications are related to the tracking of walking cycles on the basis of accelerometer signals in the framework of gait analysis. The idea is that, when feet strike the ground during walking or running activities, a peak with large acceleration salience is observed at the beginning of each cycle. Therefore, cycles can be detected by locating such remarkable events.

- Biometric authentication based on gait recognition : The authentication via accelerometer-based biometric gait recognition offers a user-friendly alternative to common authentication methods on smartphones. It has the great advantage that the authentication can be performed without user interaction. The idea is that, when the user is walking, his walk-pattern can be extracted from the acceleration signal recorded using the integrated sensors of a smartphone. This pattern can be used for authentication. In this framework, the sample salience analysis method has been used and cited in PhD thesis [Nic12], scientific papers [MN12] [Nic+11] [MM13] and different projects [Mal09] [MD09].
- A PhD thesis that concerns the Transmission Power Management for Wireless Health Applications uses salience analysis to analyze data from body sensors and provides an adaptive transmit power selection algorithm that reacts to the user's mobility [Ami12].
- A US patent [AS13] uses sample salience analysis in the framework of the monitoring of exercise or other forms of physical activity to improve health and manage obesity. Here, sample salience analysis has been applied to measured acceleration data to count a number of instances of the activity that have occurred, such as walking or running.

The salience analysis could be generalized to enable multi-dimensional analysis of data. For instance, in topography, the concept of bi-dimensional salience, expressed in km^2 and defined as the spatial area over which a topographic map sample is a maximum, would enable the categorization of mountain summits and/or the selection of the best panoramas.

Finally, the proposed perturbation analysis may be used to analyze the variability of financial time series. As an example, the method has been applied to financial data to analyze the volatility of financial assets. The financial asset price volatility refers to the degree to which prices vary over a certain length of time. A preliminary study [Mer11] has been realized in this framework. The goal consisted in decomposing the asset return time series in several time sub-series that refer to

short, medium and long term stockbroking horizon choice. The investment risk was then assessed by computing the standard deviation of the normalized variability time series.

The salience analysis has also been applied to financial data in [Van12] to analyze the information diffusion impact on stock price dynamics. The stock price time series is the result of two phenomena : the incorporation of information in the stock's valuation and noise (i.e. non-information related changes in price). The objectives were to describe the process in stock price changes due to new information release. Salience analysis as well as empirical mode decomposition have been used to identify the discrete events in high frequency intraday stock price data.



Bibliography

- [AZ91] H Ackermann and W Ziegler. "Cerebellar voice tremor: an acoustic analysis." In: *Journal of Neurology, Neurosurgery & Psychiatry* 54.1 (1991), pp. 74–76.
- [AS13] N. Amini and M. Sarrafzadeh. *Exercise-Based Entertainment And Game Controller To Improve Health And Manage Obesity*. US Patent App. 13/427,738. Apr. 2013. URL: <https://www.google.com/patents/US20130090213>.
- [Ami12] Navid Amini. "Transmission Power Management for Wireless Health Applications". PhD thesis. University of California, Los Angeles, 2012.
- [Ana+12] Supraja Anand, Rahul Shrivastav, Judith M Wingate, and Neil N Chheda. "An acoustic-perceptual study of vocal tremor". In: *Journal of Voice* 26.6 (2012), 811–e1.
- [Ann+10] M. Anniko, M. Bernal-Sprekelsen, V. Bonkowsky, P. Bradley, and S Iurato. *Otorhinolaryngology, Head and Neck Surgery*. Springer-Verlag Berlin Heidelberg, 2010.
- [Aro+92] Arnold E Aronson, William S Winholtz, Lorraine Olson Ramig, and Sandra R Silber. "Rapid voice tremor, or "flutter," in amyotrophic lateral sclerosis". In: *Annals of Otolaryngology, Rhinology & Laryngology* 101.6 (1992), pp. 511–518.
- [Ass] National Spasmodic Dysphonia Association. *Spasmodic Dysphonia*. URL: <https://www.dysphonia.org/spasmodic-dysphonia.php>.
- [BJN06] Byron J. Bailey, Jonas T. Johnson, and Shawn D. Newlands. *Head & Neck Surgery - Otolaryngology, 4th Edition*. Lippincott Williams & Wilkins, 2006.
- [BA83] Diane M Bless and James H Abbs. *Vocal fold physiology: contemporary research and clinical issues*. College-Hill, 1983.
- [Boa92] Boualem Boashash. "Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications". In: *Proceedings of the IEEE* 80.4 (1992), pp. 540–568.
- [Boe93] Paul Boersma. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Amsterdam. 1993, pp. 97–110.

- [BW01] Paul Boersma and David Weenink. “Praat, a system for doing phonetics by computer”. In: *Glott International* 5.9/10 (2001), pp. 341–345.
- [BHM77] Stephen Bradley, Arnolfo Hax, and Thomas Magnanti. *Applied mathematical programming*. Addison Wesley, 1977.
- [BS03] Eugene H Buder and Edythe A Strand. “Quantitative and Graphic Acoustic Analysis of Phonatory ModulationsThe Modulogram”. In: *Journal of speech, language, and hearing research* 46.2 (2003), pp. 475–490.
- [Cno+08] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Lefebvre, and F. Greniez. “Low-frequency vocal modulations in vowels produced by parkinsonian subjects”. In: *Speech Communication* 50.4 (2008), pp. 288–300.
- [Cno07] Laurence Cnockaert. “Analyse du tremblement vocal et application à des locuteurs parkinsoniens”. PhD thesis. Thèse de doctorat en sciences de l’ingénieur. Université libre de Bruxelles ULB, Bruxelles (Décembre 2007), 2007.
- [Com20] Pierre Combescure. “listes de dix phrases phonétiquement équilibrées”. In: *Revue d’acoustique* 56 (20), p. 1981.
- [DK02] Alain De Cheveigné and Hideki Kawahara. “YIN, a fundamental frequency estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930.
- [Dej10] Philippe H Dejonckere. “Assessment of voice and respiratory function”. In: *Surgery of larynx and trachea*. Springer, 2010, pp. 11–26.
- [DBB98] Günther Deuschl, Peter Bain, and Mitchell Brin. “Consensus Statement of the Movement Disorder Society on Tremor”. In: *Movement Disorders* 13.S3 (1998), pp. 2–23. ISSN: 1531-8257.
- [DWI02] Christopher Dromey, Paul Warrick, and Jonathan Irish. “The influence of pitch and loudness changes on the acoustics of vocal tremor”. In: *Journal of speech, language, and hearing research* 45.5 (2002), pp. 879–890.
- [Dub12] Thomas Dubuisson. “Glottal source estimation and automatic detection of dysphonic speakers”. PhD thesis. University of Mons, Faculty of Engineering, TCTS Lab, 2012.
- [Fou15] International Essential Tremor Foundation. *Essential tremor. Patient handbook*. 2015. URL: <http://www.essentialtremor.org/>.
- [Fra10] Samia Fraj. “Synthèse des voix pathologiques”. PhD thesis. Université Libre de Bruxelles, 2010.
- [Ghi+12] A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. De Looze, D. Robert, F. Viallet, and A. Giovanni. “How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?” In: *Speech Communication* 54.5 (2012). Advanced Voice Function Assessment, pp. 664–679.
- [GM12] Giuliana Grimaldi and Mario Manto. *Mechanisms and emerging therapies in tremor disorders*. Springer Science & Business Media, 2012.
- [HM18] Gray Henry and Goss Charles Mayo. *Anatomy of the human body*. Twentieth. Philadelphia : Lea and Febiger, 1918.
- [Hes83] Wolfgang Hess. *Pitch determination of speech signals*. Springer, 1983.
- [Hir81] M. Hirano. *Clinical Examination of Voice*. Vol. 5. Disorders of Human Communication. Wien, Austria: Springer-Verlag, 1981.

- [Hir95] H Hirose. “Voice quality in patients with neurological disorders”. In: *Voice Quality Control* (1995).
- [Hua+98] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis”. In: *Proceeding of the The Royal Society* (1998).
- [HW08] Norden E Huang and Zhaohua Wu. “A review on Hilbert-Huang transform: Method and its applications to geophysical studies”. In: *Reviews of Geophysics* 46.2 (2008).
- [Hua+09] Norden E. Huang, Zhaohua Wu, Steven R. Long, Kenneth C. Arnold, Xianyao Chen, and Karin Blank. “On Instantaneous Frequency”. In: *Advances in Adaptive Data Analysis* 01.02 (2009), pp. 177–229.
- [KL62] C. C. Kelly and K. L. Lochbaum. “Speech Synthesis”. In: *Proc. Fourth ICA* (1962).
- [Kla80] Dennis H Klatt. “Software for a cascade/parallel formant synthesizer”. In: *Journal of the Acoustical Society of America* 67.3 (1980), pp. 971–995.
- [KGG03] Jody Kreiman, Brian Gabelman, and Bruce R Gerratt. “Perception of vocal tremor”. In: *Journal of speech, language, and hearing research* 46.1 (2003), pp. 203–214.
- [Log+78] Jeri A Logemann, Hilda B Fisher, Benjamin Boshes, and E Richard Blonsky. “Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients”. In: *Journal of Speech and Hearing Disorders* 43.1 (1978), pp. 47–57.
- [Lud+86] Christy L Ludlow, Celia J Bassich, Nadine P Connor, and David C Coulter. “Phonatory characteristics of vocal fold tremor”. In: *Journal of Phonetics* 14.3-4 (1986), pp. 509–515.
- [Mal09] Alexis Malozemoff. *MUMT 501-Final Project Report: Accelerometer-based gait recognition*. Tech. rep. McGill University, Montreal, 2009.
- [MD09] Alexis J Malozemoff and Philippe Depalle. *MUMT 502 Project Report: Gait Recognition Using Accelerometers and Sound*. Tech. rep. McGill University, Montreal, 2009.
- [MM00] JH McAuley and CD Marsden. “Physiological and pathological tremors and rhythmic central motor control”. In: *Brain* 123.8 (2000), pp. 1545–1567.
- [Mer11] Christophe Mertens. “Application de la méthode de décomposition en modes empiriques à l’estimation de la volatilité des titres financiers”. MA thesis. Solvay Brussels School of Economics and Management. Université Libre de Bruxelles, 2011.
- [Mer+11a] Christophe Mertens, Francis Grenez, Victor Boucher, and Jean Schoentgen. “Analysis of glottal cycle tremor and jitter by empirical mode decomposition”. In: *Proceedings Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications - Firenze, Italy*. 2011, pp. 127–130.
- [Mer+09] Christophe Mertens, Francis Grenez, L. Crevier-Buchman, and Jean Schoentgen. “Saliency Analysis for Glottal Cycle Detection in Disordered Speech”. In: *Proceedings Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications - Firenze, Italy*. 2009, pp. 99–102.
- [Mer+10] Christophe Mertens, Francis Grenez, L. Crevier-Buchman, and Jean Schoentgen. “Reliable tracking based on speech sample saliency of vocal cycle length perturbations”. In: *Proceedings Interspeech - Makuhari, Japan*. 2010, pp. 2566–2569.

- [Mer+11b] Christophe Mertens, Francis Grenez, L. Crevier-Buchman, and Jean Schoentgen. *Détection des perturbations des durées de cycles basée sur la proéminence des échantillons du signal de parole*. Quatrièmes Journées de Phonétique Clinique - Strasbourg, France. 2011.
- [MGS08] Christophe Mertens, Francis Grenez, and Jean Schoentgen. *Investigation of the intelligibility and vocal disturbances of speech recorded via a throat microphone: preliminary results*. COST 2103 Advanced Voice Function Assessment - Aachen, Allemagne. 2008.
- [MGS09a] Christophe Mertens, Francis Grenez, and Jean Schoentgen. *Détection des cycles vocaux par la méthode des proéminences*. 3èmes Journées de Phonétiques cliniques - Aix-en-Provence, France. 2009.
- [MGS09b] Christophe Mertens, Francis Grenez, and Jean Schoentgen. "Preliminary evaluation of speech sample salience analysis for speech cycle detection". In: *Proc. 3rd Advanced Voice Function Assessment International Workshop - Madrid, Spain*. 2009, pp. 29–32.
- [MGS09c] Christophe Mertens, Francis Grenez, and Jean Schoentgen. "Speech sample salience analysis for speech cycle detection". In: *Proceedings Interspeech - Brighton, U.K.* 2009, pp. 939–942.
- [MGS12a] Christophe Mertens, Francis Grenez, and Jean Schoentgen. "Analysis of vocal tremor and jitter by empirical mode decomposition of glottal cycle length time series". In: *Proceedings Interspeech - Portland, U.S.A.* 2012, pp. 1634–1637.
- [MGS12b] Christophe Mertens, Francis Grenez, and Jean Schoentgen. *Analysis of vocal tremor by empirical mode decomposition of speech cycle length time series*. 8th International Conference on Voice Physiology and Biomechanics - Erlangen, Germany. 2012.
- [Mer+08] Christophe Mertens, Francis Grenez, Jean Schoentgen, and L. Crevier-Buchman. *Speech data recording with a view to the study of vocal tremor and vocal jitter in normophonic and dysphonic speakers*. Tech. rep. Funder: COST 2013. 2008.
- [MSG10] Christophe Mertens, Jean Schoentgen, and Francis Grenez. *Tracking of vocal cycle length perturbations via speech sample saliences*. Advances in Quantitative Laryngology, Voice and Speech Research - Erlangen, Allemagne. 2010.
- [Mer+13a] Christophe Mertens, Jean Schoentgen, Francis Grenez, and Sabine Skodda. "Acoustic analysis of vocal tremor in Parkinson speakers". In: *Proceedings 8th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications - Firenze, Italy*. 2013, pp. 19–22.
- [Mer+13b] Christophe Mertens, Jean Schoentgen, Francis Grenez, and Sabine Skodda. "Acoustic and Perceptual Analysis of Vocal Tremor". In: *Proceedings Interspeech - Lyon, France*. 2013, pp. 2257–2260.
- [MM13] Muhammad Muaaz and René Mayrhofer. "An analysis of different approaches to gait recognition using cell phone based accelerometers". In: *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. ACM. 2013, p. 293.
- [MN12] Muhammad Muaaz and Claudia Nickel. "Influence of different walking speeds and surfaces on accelerometer-based biometric gait recognition". In: *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*. IEEE. 2012, pp. 508–512.
- [MY08] K Murty and B Yegnanarayana. "Epoch extraction from speech signals". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.8 (2008), pp. 1602–1613.

- [Nay+07] Patrick Naylor, Anastasis Kounoudes, Jon Gudnason, Mike Brookes, et al. “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 15.1 (2007), pp. 34–43.
- [NS] National Institute of Neurological Disorders and Stroke. *Amyotrophic Lateral Sclerosis (ALS) Fact Sheet*. URL: http://www.ninds.nih.gov/disorders/amyotrophiclateral sclerosis/detail_ALS.htm.
- [Nic12] Claudia Nickel. “Accelerometer-based biometric gait recognition for authentication on smartphones”. PhD thesis. Technischen Universitat Darmstadt, 2012.
- [Nic+11] Claudia Nickel, Mohammad O Derawi, Patrick Bours, and Christoph Busch. “Scenario test of accelerometer-based biometric gait recognition”. In: *Security and Communication Networks (IWSCN), 2011 Third International Workshop on*. IEEE. 2011, pp. 15–21.
- [OS04] Alan V Oppenheim and Ronald W Schafer. “From frequency to quefrency: A history of the cepstrum”. In: *Signal Processing Magazine, IEEE* 21.5 (2004), pp. 95–106.
- [PS14] Robert J Podesva and Devyani Sharma. *Research methods in linguistics*. Cambridge University Press, 2014.
- [RFG03] G. Rilling, P. Flandrin, and P. Goncalves. “On empirical mode decomposition and its algorithm”. In: *Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing, Grado (Italy)*. 2003.
- [Ros+74] Myron J Ross, Harry L Shaffer, Asaf Cohen, Richard Freudberg, and Harold J Manley. “Average magnitude difference function pitch extractor”. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 22.5 (1974), pp. 353–362.
- [Rub+15] L.L. Rubchinsky, A.S. Kuznetsov, V.L. Wheelock, and K.A. Sigvardt. *Tremor*. 2015. URL: www.scholarpedia.org/article/Tremor.
- [Say12] K. Sayood. *Introduction to data compression*. The Morgan Kaufmann Series in Multimedia Information and Systems, 2012.
- [Sch02] Jean Schoentgen. “Modulation frequency and modulation level owing to vocal microtremor”. In: *The Journal of the Acoustical Society of America* 112.2 (2002), pp. 690–700.
- [SD83] Bruce G Secrest and George R Doddington. “An integrated pitch tracking algorithm for speech systems”. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’83*. Vol. 8. IEEE. 1983, pp. 1352–1355.
- [Sha+10] Jun Shao, Julia K MacCallum, Yu Zhang, Alicia Sprecher, and Jack J Jiang. “Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters”. In: *Journal of communication disorders* 43.1 (2010), pp. 35–44.
- [She03] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.
- [SG86] B. W. Silverman and P. J. Green. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.
- [SY95] Roel Smits and B Yegnanarayana. “Determination of instants of significant excitation in speech using group delay function”. In: *Speech and Audio Processing, IEEE Transactions on* 3.5 (1995), pp. 325–333.
- [Sto02] Brad H. Story. “An overview of the physiology, physics and modeling of the sound source for vowels”. In: *Acoustical Science and Technology*, 23,4 (2002), pp. 195–206.

- [Tal95] D Talkin. “A robust algorithm for ptch tracking”. In: *Speech Coding and Synthesis* (1995).
- [Tit93] Ingo R. Titze. *Principles of voice production*. Prentice-Hall Inc, 1993.
- [Van12] Aibek Van den ackerveken. “Information diffusion impact on stock price dynamics”. MA thesis. Ecole Polytechnique de Bruxelles. Université Libre de Bruxelles, 2012.
- [WJ95] M. P. Wand and M. C. Jones. *Kernel smoothing*. London: Chapman and Hall, 1995.
- [WR92] William S Winholtz and Lorraine Olson Ramig. “Vocal tremor analysis with the vocal demodulator”. In: *Journal of speech, language, and hearing research* 35.3 (1992), pp. 562–573.
- [YG88] Eyal Yair and Isak Gath. “On the use of pitch power spectrum in the evaluation of vocal tremor”. In: *Proceedings of the IEEE* 76.9 (1988), pp. 1166–1175.

Index

A

AM-FM decomposition	117
Carrier	116
Examples	117
Instantaneous envelope	116
Instantaneous frequency	119
Numerical differentiation, 120	
Amyotrophic lateral sclerosis	15
Anatomy	5
Ligament	5
Mucosa	5
Muscles	5
Vocal folds	5
Autocorrelation function	26
Average magnitude difference function	27

C

Cepstrum	28
Continuous wavelet transform	100
Examples	30, 100
Corpora	188

D

Dynamic programming	
Backtracking	67
Examples	68, 84
Initialization	66
Optimal path search	67
Optimization network	63

E

Electroglottography	11
Electromyography	11
Empirical mode decomposition	111
Boundary effects	113
Empirical modes	111
Examples	111
Mode mixing	121, 149
Sifting	113
Stopping criterion	115
Essential tremor	15
Cycle length perturbation analysis ...	127
Cycle length tracking	79

F	
FIR filter	80
Flowglottograph	11
Formants	9
Fourier analysis	95
Basis functions	95
Examples	98, 100
Short-term analysis	100
Spectrogram	100
Windowing	96
Fourier series	95
Coefficients	96
Example	96
Fourier transform	96

G	
Glottal excitation	8

H	
Harmonic product spectrum	29

I	
Instantaneous amplitude	94, 102
Instantaneous envelope	103, 116
Instantaneous frequency	105, 116, 119
Instantaneous phase function	94, 102, 116

K	
Kernel smoothing	141
Kernel width	141

M	
Multiple linear regression model	177

N	
Nervous system	7

Central nervous system	7
Nervous feedback loops	8
Peripheral nervous system	7
Neurological tremor	
Bandwidth (density)	143
Bandwidth (scalar quant.)	146
Depth	135
Envelope	137
Frequency	137
Frequency (density)	143
Frequency density estimate	141, 149
Frequency estimate	139
Frequency estimate (scalar quant.) ...	145
Phase	137
Time series	131
Variability	139

P	
Parkinson's disease	15
Cycle length perturbation analysis ...	127
Cycle length tracking	79
Hoehn and Yahr scale	188
Key results	215
Possible symptoms	15
UPDR scale	188
Perturbation analysis	
Categorization	131
Neurological tremor bandwidth .	136, 139, 143, 145
Neurological tremor frequency .	136, 139, 143, 145
Perturbation sizes	135
Validation	177
Vocal frequency	135
Physiological tremor	
Depth	135
Time series	131
Pneumotachograph	11

S	
Salience	42
Final salience	46, 52
Local salience	45
Running salience	46, 52
Salience allocation	43
Basic algorithm	43

Examples	55, 82	Deferent	95
Global salience allocation	49	Epicycle	95
Partial salience allocation	47	Trajectory	93, 95
Sliding analysis window length	54, 82		
Validation	52		
Salience-based cycle length tracking	79		
Cycle length time series	86		
Cycle length tracking	63, 84		
Parameters, 84			
Preprocessing	80		
Salience allocation	45, 82		
Validation	164		
Scalar quantization	145		
Spasmodic dysphonia	15		
Spirometer	11		
Synthesizer of disordered voices	161		
Glottal airflow model	163		
Perturbation model	161		

T

Topographic isolation	41
Topographic prominence	41
Triplet	64
Triplet sequence	64
Average inter-peak distance	66
Number of cycles	66
Overall length perturbation	65

V

Videokymography	11
Videolaryngostroboscopy	11
Vocal frequency	135
Vocal jitter	14
Size	135
Time series	131
Vocal timbre	9
Vocal tract	9
Vocal tremor	14
Causes	14
Definition	14
Literature review	16

W

Wheel mechanism	94
---------------------------	----

Abstract

The study concerns the analysis of vocal cycle length perturbations in normophonic and dysphonic speakers.

A method for tracking cycle lengths in voiced speech is proposed. The speech cycles are detected via the saliences of the speech signal samples, defined as the length of the temporal interval over which a sample is a maximum. The tracking of the cycle lengths is based on a dynamic programming algorithm that does not request that the signal is locally periodic and the average period length known a priori.

The method is validated on a corpus of synthetic stimuli. The results show a good agreement between the extracted and the synthetic reference length time series. The method is able to track accurately low-frequency modulations and fast cycle-to-cycle perturbations of up to 10% and 4% respectively over the whole range of vocal frequencies. Robustness with regard to the background noise has also been tested. The results indicate that the tracking is reliable for signal-to-noise ratios higher than 15dB.

A method for analyzing the size of the cycle length perturbations as well as their frequency is proposed. The cycle length time series is decomposed into a sum of oscillating components by empirical mode decomposition the instantaneous envelopes and frequencies of which are obtained via AM-FM decomposition. Based on their average instantaneous frequencies, the empirical modes are then assigned to four categories (declination, physiological tremor, neurological tremor as well as cycle length jitter) and added within each. The within-category size of the cycle length perturbations is estimated via the standard deviation of the empirical mode sum divided by the average cycle length. The neurological tremor modulation frequency and bandwidth are obtained via the instantaneous frequencies and amplitudes of empirical modes in the neurological tremor category and summarized via a weighted instantaneous frequency probability density, compensating for the effects of mode mixing.

The method is applied to two corpora of vowels comprising 123 and 74 control and 456 and 205 Parkinson speaker recordings respectively. The results indicate that the neurological tremor modulation depth is statistically significantly higher for female Parkinson speakers than for female control speakers. Neurological tremor frequency differs statistically significantly between male and female speakers and increases statistically significantly for the pooled Parkinson speakers compared to the pooled control speakers. Finally, the average vocal frequency increases for male Parkinson speakers and decreases for female Parkinson speakers, compared to the control speakers.