University of Mons

Doctoral School MUSICS
Signal Processing

P H D T H E S I S

to obtain the title of

PhD in Applied Sciences

of University of Mons
Specialty : Speech Processing

Defended by

Thomas Drugman

Advances in Glottal Analysis and its Applications

Thesis Advisor: Thierry Dutoit

prepared at University of Mons, Faculty of Engineering,
TCTS Lab

Jury :

Prof. Christine Renotte	-	University of Mons
Prof. Thierry Dutoit	-	University of Mons
Prof. Francis Moiny	-	University of Mons
Dr. Vincent Pagel	-	Acapela Group
Ass. Prof. Baris Bozkurt	-	Izmir Institute of Technology, Turkey
Ass. Prof. Yannis Stylianou	-	University of Crete, Greece
Prof. Christophe d'Alessandro	-	University of Paris South, France

To my family, my friends, and all the good souls I have met across the world.

Shoot for the moon. Even if you miss, you will land among the stars.

[Les Brown]

Abstract

From artificial voices in GPS to automatic systems of dictation, from voice-based identity verification to voice pathology detection, speech processing applications are nowadays omnipresent in our daily life. By offering solutions to companies seeking for efficiency enhancement with simultaneous cost saving, the market of speech technology is forecast to be particularly promising in the next years.

The present thesis deals with advances in glottal analysis in order to incorporate new techniques within speech processing applications. While current systems are usually based on information related to the vocal tract configuration, the airflow passing through the vocal folds, and called glottal flow, is expected to exhibit a relevant complementarity. Unfortunately, glottal analysis from speech recordings requires specific complex processing operations, which explains why it has been generally avoided.

The main goal of this thesis is to provide new advances in glottal analysis so as to popularize it in speech processing. First, new techniques for glottal excitation estimation and modeling are proposed and shown to outperform other state-of-the-art approaches on large corpora of real speech. Moreover, proposed methods are integrated within various speech processing applications: speech synthesis, voice pathology detection, speaker recognition and expressive speech analysis. They are shown to lead to a substantial improvement when compared to other existing techniques.

More specifically, the present thesis covers three separate but interconnected parts. In the first part, new algorithms for robust pitch tracking and for automatic determination of glottal closure instants are developed. This step is necessary as accurate glottal analysis requires to process pitch-synchronous speech frames. In the second part, a new non-parametric method based on Complex Cepstrum is proposed for glottal flow estimation. In addition, a way to achieve this decomposition asynchronously is investigated. A comprehensive comparative study of glottal flow estimation approaches is also given. Relying on this expertise, the usefulness of glottal information for voice pathology detection and expressive speech analysis is explored. In the third part, a new excitation modeling called Deterministic plus Stochastic Model of the residual signal is proposed. This model is applied to speech synthesis where it is shown to enhance the naturalness and quality of the delivered voice. Finally, glottal signatures derived from this model are observed to lead to an increase of identification rates for speaker recognition purpose.

Keywords: Information Technology, Voice Technology, Speech Processing, Speech Analysis, Speech Synthesis, Speaker Recognition, Voice Pathology, Expressive Speech, Glottal flow, Source-tract Separation, Pitch Estimation, Glottal Closure Instant, Excitation Modeling.

Acknowledgements

First and foremost this thesis would not have been possible unless the support from the Fonds National de la Recherche Scientifique (FNRS). I also owe my deepest gratitude to Prof. Thierry Dutoit, my supervisor, for his precious guidance, advices, dynamism, friendship, and above all for his patience with regard to my jokes. It was an honor for me to work under his supervision.

I would like to thank Ass. Prof. Baris Bozkurt, from the Izmir Institute of Technology (Turkey), Prof. Abeer Alwan, from the University of California, Los Angeles, and Dr. Pierre Divenyi, from the Speech and Hearing Reasearch Department, Matinez, California, for welcoming me so warmly in their laboratories. It was a pleasure to share so many things with them, both at the scientific and human points of view. I am also grateful to Mr. Geoffrey Wilfart, Dr. Vincent Pagel, Mr. Fabrice Malfrère and Mr. Olivier Deroo, from the Acapela Group, for the industrial partnership. I would like to extend my appreciation to Prof. Patrick Naylor and Dr. Mark Thomas, from the Imperial College of London, and Dr. Jon Gudnason, from the University of Iceland, Reykjavik, for their fruitful collaboration. I am also thankful to Prof. Patrice Mégret for his precious help with LaTeX.

I am indebted to my colleagues to support my sense of humor and my hyperactivity. I promise them not to drink coffee anymore. It was a real pleasure to work in such a familiar and productive atmosphere. For both their help and friendship, I would like to thank Alex Brq#1, Thom Jack Black, Benj m'Poney, Djé Bien Monté, Maria my Deer, Onur Patlejan, Fredo Collina, Nico Ramses, Matt Kool & the Gang, Jean-Marc Tu-me-l'avais-promis, Ben Souss la Racaille, Nico Ké-des-pecs, Joelle la Schtroumpfette and ma Nathalie Préférée.

I am thankful to my friends for their help and all the good moments we shared. Among others, special thanks to Damskoutz Belote-Rebelote, Bobo Funky House Maldini, Jess ma Grande Soeur, Fa le Lapin, Roma le p'tit Frérino, Vinz le Para, Da le Gentil, Caro m'Bidon, Zacky Big Louloute, Enzouille, mon Cricri d'Amour, Jo Pérusse, Pitch Très Chouette, Francesco Del Bosqué, Cristofo Cacciatore, Paoluzzo Il Più Figo, Chri Tutto-dipende-da-te, Mitz Forte Bine, Freddytu El Messi de Venenzuela, Andresito Cerveza, Béné Tit Chat, Da le Méchant, Kev Ted, Darkanno, Gwé Vandekamp, Ri Jonagold, Tonton Roucha and Shevaless.

I would like to express my deepest gratitude to my parents, grand-parents, my brother Sylvanucci, and Stéphanie. They have always been there, both in good and hard times. They were present to initiate my delights and to sweep my sorrows. I will always be indebted for everything they made for me. Without their constant support, I would not be what I am today. I owe them so much.

Finally, an infinity of thanks to Victoire for backing me during the writing of the thesis. She is simply a sunshine, and through her joy of life brought an ocean of love and happiness in me.

Contents

List of Figures

List of Tables

Acronyms

AC	AutoCorrelation
ANN	Artificial Neural Network
ARMA	AutoRegressive Moving Average
ARX	AutoRegressive eXogenous
CALM	Causal-Anticausal Linear Model
CC	Complex Cepstrum
CCD	Complex Cepstrum-based Decomposition
CELP	Code-Excited Linear Prediction
CGD	Chirp Group Delay
CMOS	Comparative Mean Opinion Score
CoG	Center of Gravity
CPIF	Closed Phase Inverse Filtering
CPU	Central Processing Unit
CRD	Cumulative Relative Dispersion
CZT	Chirp Z-Transform
DAP	Discrete All Pole
DFT	Discrete Fourier Transform
DSM	Deterministic plus Stochastic Model
DTFT	Discrete Time Fourier Transform
DYPSA	Dynamic Programming Phase Slope Algorithm
EGG	ElectroGlottoGraph
FAR	False Alarm Rate
FFE	F0 Frame Error
FFT	Fast Fourier Transform
FM	Fourier Magnitude
FPE	Fine Pitch Error
GCI	Glottal Closure Instant
GIA	Global Industry Analysts
GMM	Gaussian Mixture Model
GOI	Glottal Opening Instant
GPE	Gross Pitch Error
GRBAS	Grade, Roughness, Breathiness, Aesthenia, Strain

HCI	Human-Computer Interaction
HE	Hilbert Envelope
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HRF	Harmonic Richness Factor
IAIF	Iterative Adaptive Inverse Filtering
IDA	Identification Accuracy
IDR	Identification Rate
IDTFT	Inverse Discrete Time Fourier Transform
IFFT	Inverse Fast Fourier Transform
LF	Liljencrants-Fant
LOMA	Lines Of Maximum Amplitude
LPC	Linear Predictive Coding
MBE	Multi-Band Excitation
MCGF	Mel-Cesptral representation of the Glottal Flow
ME	Mixed Excitation
MFCC	Mel-Frequency Cepstral Coefficients
MGC	Mel-Generalized Cepstrum
MI	Mutual Information
MLSA	Mel-Log Spectrum Approximation
ModGD	Modified Group Delay
MR	Miss Rate
MSD	Multi-Space probability Density
NAQ	Normalized Amplitude Quotient
NUU	Non-Uniform Units
OQ	Open Quotient
PCA	Principal Component Analysis
PE	Perceptive Energy
PPGD	Product of the Power and Group Delay
PSP	Parabolic Spectrum Parameters
QOQ	Quasi-Open Quotient
RCT	Relative Computation Time
RIR	Room Impulse Response
ROC	Receiver Operating Characteristic
ROM	Read-Only Memory
RTSE	Relative Time Squared Error
SD	Spectral Distortion
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SHRP	Subharmonic to Harmonic Ratio Pitch tracker
SIFT	Simplified Inverse Filter Tracking
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level

SPS	Statistical Parametric Synthesis
SRH	Summation of Residual Harmonics
SSH	Summation of Speech Harmonics
SWT	Stationary Wavelet Transform
TDGF	Time Domain representation of the Glottal Flow
TDPSOLA	Time Domain Pitch Synchronous Overlap-Add
TTS	Text-To-Speech
VAD	Voice Activity Detection
VDE	Voicing Decision Error
VSCC	Voice Source Cepstrum Coefficients
WGN	White Gaussian Noise
YAGA	Yet Another GCI Algorithm
ZCZT	Zeros of the Chirp Z-Transform
ZFR	Zero Frequency Resonator
ZZT	Zeros of the Z-Transform

Chapter 1

General Introduction

Contents

1.1 Speech Technology: What For?

1.1.1 Did you say *"Speech Processing"*?

Speech is certainly the most natural communication mode which humans use to interact with each other. This can be explained by the fact that speech is characterized by a high delivery rate of information. This information can be analyzed at several non-exclusive levels of description. At the *acoustic* level [1], speech is studied as a mechanical wave that is an oscillation of pressure. *Phonetics* [2] deals with the physical properties of speech sounds (phones): how it is produced by the articulatory system, and how it is perceived by the auditory system. The *phonological* level [2] is the necessary interface between phonetics and linguistical descriptions of higher levels. It introduces an abstract and functional unit called *phoneme*, which has the property to convey a meaning. *Morphology* [3] focuses on the formation and composition of words, while *syntax* [4] studies the formation and composition of phrases and sentences from these words. The *semantical* level [5] is concerned with how meaning is inferred from words and concepts. As for the last description level, *pragmatics* [6] analyses how meaning is inferred from context. In this thesis, it is focused on the **acoustic analysis** of speech.

These several levels explain why language acquisition is such a highly complex task [7], which is carried out by infants a long time after the learning of skills such as walking. Nonetheless, after this long acquisition step, speech turns out to be one of the most efficient means of communication. Therefore there is an important demand to incorporate speech as a possible modality in Human-Computer Interactions (HCIs, [8]). This motivates the interest for a large variety of *Speech Processing* applications.

In the broad sense, Speech Processing refers to the study of speech signals and the ensuing processing methods. The main applications of Speech Processing can be categorized as follows:

- *Speech Recognition:* Speech Recognition refers to the task for a machine to recognize and understand speech [9]. Main challenges are to reach high recognition rates for any speaker and in any environment.

- *Speech Synthesis:* In Speech Synthesis, also called Text-to-Speech (TTS), the goal is to produce the automatic lecture of an unknown text [10]. Challenges are typically expressed in terms of naturalness and intelligibility of the produced voice.

- *Speaker Recognition:* Automatic Speaker Recognition refers to the use of a machine in order to recognize a person from a spoken phrase [11]. Speaker Recognition is made of two main subfields [12]: Speaker *Verification* (i.e to verify a person's claimed identity from his voice) and Speaker *Identification* (i.e there is no a priori identity claim, and the system decides who the person is).

- *Voice Analysis:* Voice Analysis is the study of speech sounds for the purpose of characterizing non-standard speech, i.e exhibiting an affect, a voice disorder and so forth. Goals can be, for example, to detect, quantify and qualify a voice pathology within a medical context [13], or to study and synthesize expressive speech [14].

- *Speech Enhancement:* Speech Enhancement refers to the *cleaning* process which aims at reducing the presence of noise in a corrupted signal, or the task of enhancing its intelligibility [15].

- *Speech Coding:* Speech Coding is the art of reducing the bit rate required to describe a speech signal while preserving its quality [16]. This is a particular form of data compression (and sound compression), important in the telecommunication area.

The positioning of this thesis with regard to these Speech Processing applications is described in Section 1.3.

1.1.2 The Speech Technology Market

As an important component of Information Technology, the field of speech technology has exploded with the advent and dazzling growth of telecommunication techniques. According to the Global Industry Analysts (GIA), world speech technology market is forecast to reach US$20.9 billion by the year 2015 [17]. In that study, GIA focuses on the three following product segments: Speech Recognition, Speech Synthesis and Speaker Recognition. The GIA report profiles 211 companies (including 231 divisions/subsidiaries) including many key and niche players [17]. To give an idea of the location of speech technology companies around the world, Table 1.1 shows the repartition of the players considered in [17].

Country	**Number of market players**
United States	145
Canada	13
Germany	13
United Kingdom	16
Rest of Europe	24
Asia-Pacific	15
Middle-East	5

Table 1.1 - *Global repartition of speech technology market players considered in [17].*

It is worth noting that these latter statistics do not encompass the budget spent by armies, most of all by the US army which is known to invest collocal amounts of money in speech technology applications.

A particularly attractive asset of the speech technology market is that it benefits from several medium to long-term advantages [17]. This has allowed the speech market to stay afloat and perform remarkably well even in the harsh financial crisis.

Besides the private sector, speech technology has caught the interest of a large research community. To illustrate this, the Interspeech conference is the world's largest and most comprehensive conference on issues surrounding the science and technology of spoken language processing both in humans and in machines. For its 2010 edition held in Makuhari, Japan, 1324 scientific articles were submitted[1] (among which 779 were included in the technical program), which reflects the strong interest in developing new speech technology solutions.

1.2 Speech Production and Modeling

1.2.1 Speech Production

Organs intervening in the phonation process can be categorized into three main groups [18]: the lungs, larynx and vocal tract. The lungs are the source of energy and their role is to provide an airflow arising in the trachea. This airflow then reaches the larynx where it is modulated. The larynx houses the vocal folds (see Figure 1.1), which are an essential component of phonation. The space comprised between the vocal folds is defined as the glottis. The glottal modulation provides either a periodic or a noisy source to the vocal tract. The vocal tract (see Figure 1.2) is made of the oral, nasal and pharyngeal resonant cavities. Its role is to *"color"* the sound, i.e to define its timbre, by spectrally shaping the glottal airflow [1]. The airflow modulated by the glottis and colored by the vocal tract is then radiated by the lips. This variation of air pressure causes a traveling wave which is perceived as speech by the listener [1].

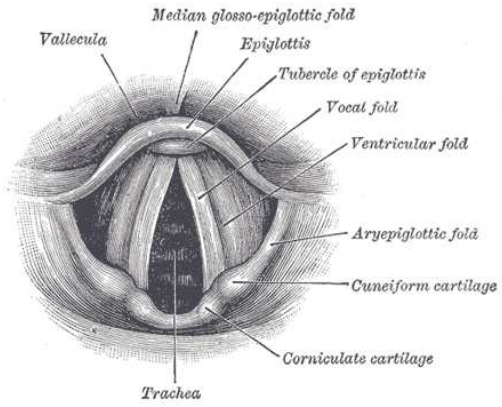

Figure 1.1 - *Tranversal view of the larynx. Glottis is defined as the space comprised between the vocal folds. From Gray's Anatomy of the Human Body, 20th edition.*

This thesis focuses on the glottal component of the speech signal. The glottal behaviour manipulates both pitch and volume, as well as the voice quality. Pitch represents the perceived fundamental frequency of a sound, and allows the construction of *"melody"* in speech (i.e if a sound is *"higher"* or

[1] Interspeech Conference 2010, *http://www.interspeech2010.org/*

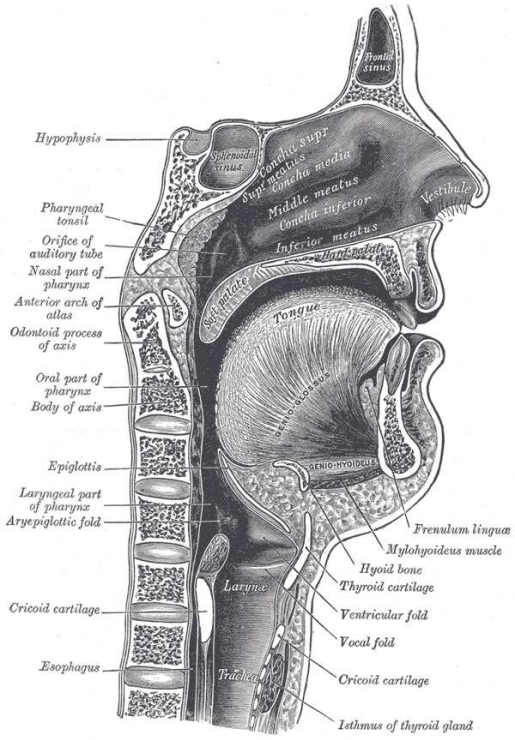

Figure 1.2 - *Representation of the phonation apparatus. Speech results from an airflow evicted from the lungs, arising in the trachea, passing through the glottis, filtered by the vocal tract cavities and finally radiated by the lips. From Gray's Anatomy of the Human Body, 20th edition.*

"lower"). The estimation of pitch from the speech signal is studied in Chapter 2. Pitch and volume are important features for controlling stress and intonation of speech. As for the voice quality, it refers to the laryngeal phonation style [19] and provides paralinguistic information. For example, voice quality makes the difference between a soft, normal or loud phonation.

As previously mentioned, one may distinguish two modes in the glottal behaviour. During the production of *voiced sounds*, the airflow arising from the trachea causes a quasi-periodic vibration of the vocal folds [18]. On the opposite, when the glottal excitation is noisy, resulting sounds are qualified as *unvoiced*. In this thesis, it is focused on the glottal characteristics during the production of voiced speech.

Figure 1.3 gives an illustration of speech analysis on the sentence *"The cranberry bog gets very pretty in Autumn."* uttered by a male speaker. Plot (a) shows the speech waveform as captured by a microphone. The second plot is the so-called *spectrogram* of the speech signal, i.e a representation of the energy distribution in the time-frequency domain. Plot (c) displays the corresponding pitch track. It can be observed that pitch only exists for voiced regions of speech, where the signal is pseudo-periodic.

An important difficulty in glottal analysis is the difficulty in observing the glottal behaviour. Some devices such as electroglottographs (EGG) or laryngographs measure the impedance between the vocal folds [20], which is an image of the glottal opening. Another approach is the use of high-speed imaging (typically around 3000 images/second, [21]) recorded by introducing a laryngoscope positioned to visualize the larynx. A crucial drawback of these apparatus is that they are particularly uncomfortable for the speaker. In addition, although they are informative about the glottal behaviour, they only provide an image (e.g the glottal impedance, or surface) of the real glottal flow. For these reasons, this thesis only focuses on designing techniques which solely rely on the audio speech signal captured by a

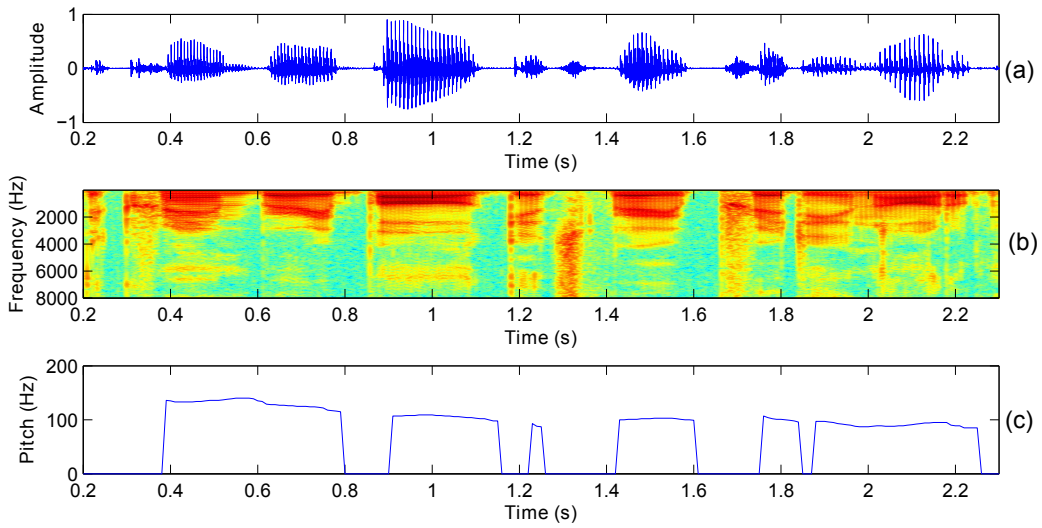

Figure 1.3 - *Illustration of a speech sound. The sentence* "The cranberry bog gets very pretty in Autumn." *has been uttered by a male speaker.* (a)*: The speech waveform,* (b)*: its spectrogram,* (c)*: the pitch track.*

microphone, as it is the case for the great majority of Speech Processing applications.

A more thorough introduction on the processing of the glottal flow is given in Chapter 4.

1.2.2 Speech Modeling

A large number of speech models rely on a source-filter approach [18]. In such an approach, the source refers to the excitation signal produced by the vocal folds at the glottis, and the filtering operation to the spectral coloring carried out by the vocal tract cavities. In several speech processing applications, separating these two contributions is important as it could lead to their distinct characterization and modeling. This is advantageous since these two components act on different properties of the speech signal.

The actual excitation signal is the airflow arising from the trachea and passing through the vocal folds, and is called the glottal flow [1]. However, its estimation directly from the speech waveform is a typical blind separation problem since neither the glottal nor the vocal tract contributions are observable. This makes the glottal flow estimation a particularly complex issue. Part II of this thesis adresses the problem of automatically estimating the glottal flow directly from the speech waveform, and how it can be applied in Voice Analysis.

Due to the aforementioned hindrances, using the real glottal flow in usual speech processing systems is commonly avoided. For this reason, it is generally preferred to consider, for the filter, the contribution of the spectral envelope of the speech signal, and for the source, the residual signal obtained by inverse filtering. Although not exactly motivated by a physiological interpretation, this approach has the advantage of being more practical while giving a sufficiently good approximation to the actual *deconvolution problem*, i.e the problem of source-tract (or source-filter) separation.

Methods parameterizing the spectral envelope (i.e the filter), such as the well-known Linear Predictive Coding (LPC) or Mel Frequency Cepstral (MFCC) features [22], are widely used in almost every field of Speech Processing. Figure 1.4 illustrates the LPC technique for the windowed segment

of real speech shown in plot (a). This speech frame is a segment of the sentence analyzed in Figure 1.3. It is observed in plot (b) that the spectral envelope of the speech signal (thin line) is correctly estimated via the LPC modeling (solid line). Note that the speech spectrum, as displayed in plot(b), can be seen as a vertical slice in the spectrogram of Figure 1.3(b), at the considered analysis time. Various peaks can be noticed in this spectrum. They correspond to the vocal tract resonances, called *formants*. The first four formants, denoted F_1 to F_4, are indicated on the plot. The residual signal obtained via inverse filtering (i.e after removing the contribution of the spectral envelope) is displayed in plot (c), and is referred to as the LPC *source* or *excitation* signal. For this example, the speech and the residual signals are periodic, and consequently the corresponding sound is qualified as voiced. Particular instants of significant excitation are observed in the residual signal. These are referred to as Glottal Closure Instants (GCIs). Their automatic determination from the speech signal is studied in Chapter 3. It is also seen in plot (d) that the amplitude spectrum of the residual signal is almost flat, which is a result of the "whitening" process (i.e the effects of the vocal tract resonances have been removed) achieved by inverse filtering.

Contrarily to methods capturing the spectral envelope (i.e modeling the filter), techniques modeling the excitation signal are still not well established and there might be a lot to be gained by incorporating such a modeling in several speech processing applications. This is the object of Part III which proposes a new model of the residual signal, and investigates its application to Speech Synthesis and Speaker Recognition.

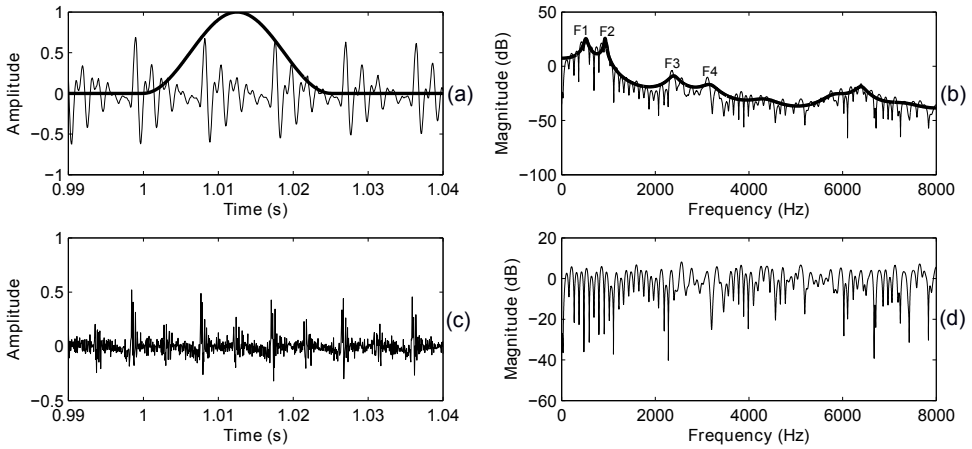

Figure 1.4 - *Illustration of the LPC method.* (a): *the speech signal (thin line) and the applied window (solid line),* (b): *the magnitude spectrum of the speech signal (thin line) and its LPC spectral envelope (solid line), with the four first formants indicated for information,* (c): *the LPC residual signal obtained by inverse filtering,* (d): *the magnitude spectrum of the residual signal.*

1.3 Contributions and Structure of the Thesis

The contribution of the present thesis is schematized in Figure 1.5. The main outter circle (in purple) is Speech Analysis, which is a field aiming at developping tools of signal processing applied to the speech signal. As explained in Section 1.2.1, during the mechanism of phonation, an airflow is evicted from the lungs, arises in the trachea and is modulated by its passage through the space delimited by the vocal folds and called glottis [1]. Glottal Analysis (the middle circle in green in Figure 1.5) then

refers to the study of methods using information from the glottal component of speech. This thesis (inner circle in red in Figure 1.5) will present some advances made in the field of Glottal Analysis.

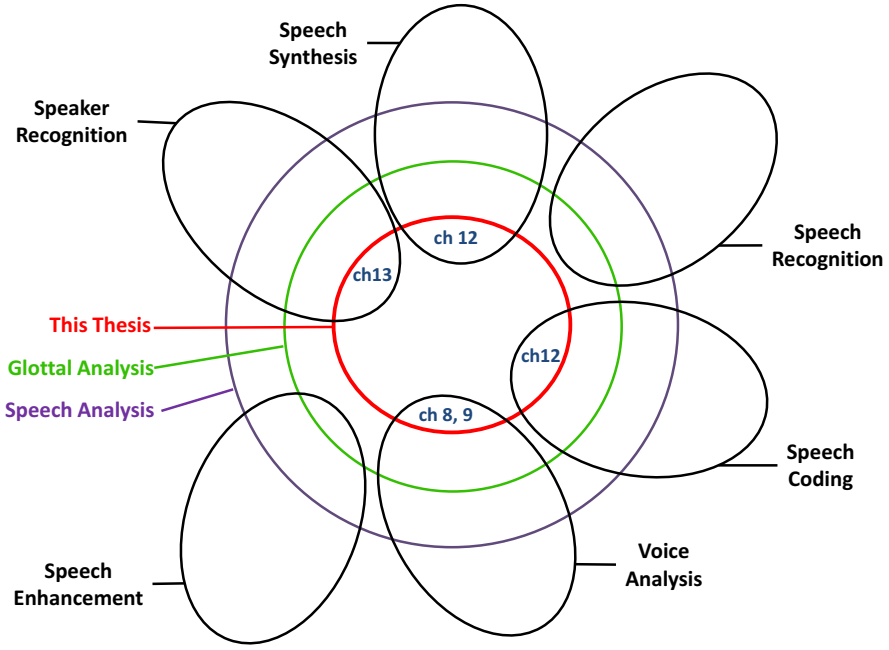

Figure 1.5 - *Schematic representation of the contribution of the present thesis. Its goal is to develop new methods of Speech Analysis using glottal information, and to integrate them in several applications of Speech Processing: Voice Analysis in Chapters 8 and 9, Speech Synthesis and Speech Coding in Chapter 12, and Speaker Recognition in Chapter 13.*

The six black ellipses in Figure 1.5 represent the six main applications of Speech Processing introduced in Section 1.1.1. It is worth noting that although these latter are not independent and should therefore exhibit some overlap, this is not displayed in Figure 1.5 for the sake of clarity. Since methods designed in Speech Analysis are fundamental tools of signal processing, they can be applied to all fields of Speech Processing. The goal of this thesis is to develop new techniques of Glottal Analysis, and to integrate them in several applications of Speech Processing: Voice Analysis in Chapters 8 and 9, Speech Synthesis and Speech Coding in Chapter 12, and Speaker Recognition in Chapter 13. Albeit glottal information could be useful for Speech Recognition and Speech Enhancement, these issues are not tackled in the frame of this thesis.

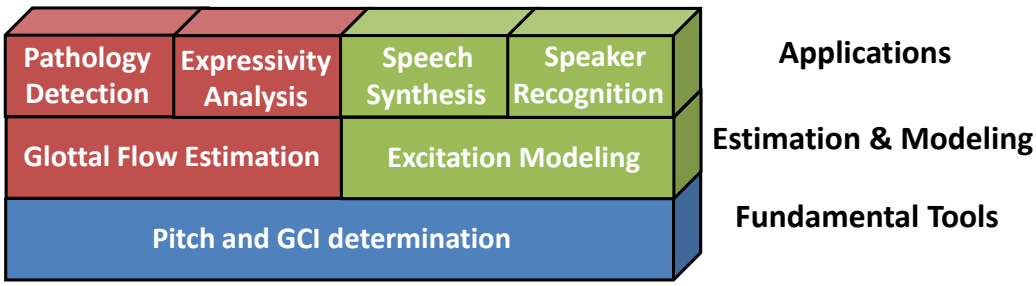

Figure 1.6 - *Schematic structure of the thesis in three layers. Part I is represented in blue, Part II in red and Part III in green.*

The structure of this thesis is schematized in Figure 1.6, according to three abstraction levels: the development of fundamental tools, the proposition of new techniques for glottal flow estimation and modeling, and their integration within various speech processing applications. The thesis is divided into three parts, represented by three different colours in Figure 1.6:

- **Part I** investigates the development of efficient and robust tools using only the speech recordings, which are necessary for precise Glottal Analysis. Chapter 2 focuses on robust pitch tracking, where, even in adverse conditions, the goal is to determine the voiced regions of speech and to extract the pitch contour. Chapter 3 studies the automatic detection of Glottal Closure Instants (GCIs), these particular moments of significant excitation of the vocal tract. The accurate, reliable and robust estimation of both the pitch and GCI locations is required in several speech processing systems, and in particular for Glottal Analysis. Indeed, for such applications, it is preferable to process speech frames synchronized on GCIs and whose length is proportional to the pitch period. Algorithms proposed in Chapters 2 and 3 are thus fundamental tools for Speech Analysis.

- **Part II** addresses the issue of glottal flow estimation from the speech waveform. An introduction on this topic as well as a presentation of the state-of-the-art methods are first given in Chapter 4. The three following chapters focus on the development and assessment of a non-parametric technique of glottal flow estimation which exploits phase properties of the speech signal. The theoretical framework as well as a first evaluation of this method are given in Chapter 5. This approach is quantitavely compared to other existing techniques of glottal flow estimation in Chapter 6. Finally, Chapter 7 aims at removing the constraint of GCI-synchronization in the proposed method. Based on the study led in these three chapters, the remainder of Part II then targets incorporating glottal features within two specific applications of Speech Processing: automatic voice pathology detection in Chapter 8 and expressive speech analysis in Chapter 9. Finally, Chapter 10 highlights the main results of glottal flow estimation and applicability obtained in Part II.

- **Part III** proposes a new model of the residual signal. The theoretical framework and properties of this model are thoroughly described in Chapter 11. The two next chapters are then targeted at applying this new excitation model within two important fields of Speech Processing. Chapter 12 integrates this model in a vocoder for Speech Synthesis, while Chapter 13 investigates the potential use of glottal signatures derived from this model for Speaker Recognition. In both applications, it is shown that the proposed model outperforms other state-of-the-art excitation modelings.

Finally Chapter 14 concludes and summarizes the main contributions of this thesis.

Bibliography

[1] T. Quatieri. *Discrete-time speech signal processing.* Prentice-Hall, 2002.

[2] M. Davenport and S. Hannahs. *Introducing Phonetics and Phonology.* A Hodder Arnold Publication, first edition, 1998.

[3] R. Lieber. *Introducing Morphology.* Cambridge University Press, first edition, 2010.

[4] A. Carnie. *Syntax: A Generative Introduction.* Wiley-Blackwell, second edition, 2006.

[5] N. Riemer. *Introducing Semantics.* Cambridge University Press, first edition, 2010.

[6] Y. Huang. *Pragmatics.* Oxford University Press, 2006.

[7] S. Crain and D. Lillo-Martin. *An Introduction to Linguistic Theory and Language Acquisition.* Wiley-Blackwell, 1999.

[8] C. Benoit, J. Martin, C. Pelachaud, L. Schomaker, and B. Suhm. Audio-visual and multimodal speech systems. 1998.

[9] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[10] T. Dutoit. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht, 1997.

[11] D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 4072–4075, 2002.

[12] J. Campbell. Speaker recognition: A tutorial. *Proc. of the IEEE*, 85(9):1437–1462, 1997.

[13] J. Stemple, L. Glaze, and B. Klaben. Clinical voice pathology: Theory and management. *Singular Editions, 3rd Edition*, 2000.

[14] N. Campbell. *Expressive/Affective Speech Synthesis.* Springer Handbook on Speech Processing, 2007.

[15] J. Benesty, S. Makino, and J. Chen. *Speech Enhancement.* Springer, Berlin, Heidelberg, 2005.

[16] B. Atal, V. Cuperman, and A. Gersho. *Advances in Speech Coding.* Springer, first edition, 1990.

[17] Global Industry Analysts Inc. Speech technology - a global strategic business report. page 812, 2011.

[18] J. Flanagan. *Speech Analysis, Synthesis, and Perception.* Springer-Verlag, second edition, 1972.

[19] C. d'Alessandro. Voice source parameters and prosodic analysis. In *Method in empirical prosody research*, pages 63–87, 2006.

[20] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. *J. Acoust. Soc. Am.*, 115:1321–1332, 2004.

[21] Y. Shue and A. Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5134–5137, 2010.

[22] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel generalized cepstral analysis Ů a unified approach to speech spectral estimation. In *ICSLP*, 1994.

Part I

Pitch Estimation and Glottal Closure Instant Determination

Chapter 2

Robust Pitch Tracking Based on Residual Harmonics

Contents

Abstract

This chapter focuses on the problem of pitch tracking in noisy conditions. A method exploiting the harmonics of the residual signal is presented. The proposed criterion is used both for pitch estimation, as well as for determining the voicing segments of speech. In the experiments, the method is compared to six state-of-the-art pitch trackers on the Keele and CSTR databases. The proposed technique is shown to be particularly robust to additive noise, leading to a significant improvement in adverse conditions.

This chapter is based upon the following publication:

- Thomas Drugman, Abeer Alwan, *Robust Pitch Tracking Based on Residual Harmonics*, Interspeech Conference, Firenze, Italy, 2011.

Many thanks to Prof. Abeer Alwan from University of California, Los Angeles, for her helpful guidance and fruitful discussions.

2.1 Introduction

Pitch tracking refers to the task of estimating the contours of the fundamental frequency for voiced segments. Such a system is of particular interest in several applications of speech processing, such as speech coding, analysis, synthesis or recognition. While most current pitch trackers perform well in clean conditions, their performance rapidly degrades in noisy environments and the development of accurate and robust algorithms still remains a challeging open problem.

Techniques estimating the fundamental frequency from speech signals can be classified according to the features they rely on [1]. Some methods use properties in the time domain, others focus on the periodicity of speech as manifested in the spectral domain, while a last category exploits both spaces. Besides, this information can be processed in a deterministic way, or using a statistical approach [1]. This chapter proposes a pitch tracking method exploiting the harmonics contained in the spectrum of the residual signal. The idea of using a summation of harmonics for detecting the fundamental frequency is not new. In [2], Hermes proposed the use of a subharmonic summation so as to account for the phenomenon of virtual pitch. This approach was inspired by the use of spectral and cepstral comb filters [3]. In [4], Sun suggested the use of the Subharmonic-to-Harmonic Ratio for estimating the pitch frequency and for voice quality analysis. The method proposed in this chapter is different in several points. First, the spectrum of the residual signal (and not of the speech signal) is inspected. As in the Simplified Inverse Filter Tracking (SIFT) algorithm (which relies on the autocorrelation function computed on the residual signal, [5]), flattening the amplitude spectrum allows to minimize the effects of both the vocal tract resonances and of the noise. Secondly, the harmonic-based criterion used for the pitch estimation is different from those employed in the two aforementioned approaches. Besides the proposed criterion is also used for discriminating between voiced and unvoiced regions of speech. Note that harmonic-based Voice Activity Detector (VAD) has also been exploited in [6].

The structure of the chapter is the following. Section 2.2 describes the principle of the proposed technique. An extensive quantitative assessment of its performance in comparison with other state-of-the-art techniques is given in Section 2.3, focusing particularly on noise robustness. Section 2.3.1 presents the adopted experimental protocol. The implementation details of the proposed method are discussed in Section 2.3.2. Methods compared in this work are presented in Section 2.3.3 and results of the evaluation are provided in Section 2.3.4.

2.2 Pitch tracking based on residual harmonics

The proposed method relies on the analysis of the residual signal. For this, an auto-regressive modeling of the spectral envelope is estimated from the speech signal $s(t)$ and the residual signal $r(t)$ is obtained by inverse filtering. This whitening process has the advantage of removing the main contributions of both the noise and the vocal tract resonances. For each Hanning-windowed frame, covering several cycles of the resulting residual signal $r(t)$, the amplitude spectrum $R(f)$ is computed. $R(f)$ has a relatively flat envelope and, for voiced segments of speech, presents peaks at the harmonics of the fundamental frequency F_0. From this spectrum, and for each frequency in the range $[F_{0,min}, F_{0,max}]$, the Summation of Residual Harmonics (SRH) is computed as:

$$SRH(f) = R(f) + \sum_{k=2}^{N_{harm}} [R(k \cdot f) - R((k - \frac{1}{2}) \cdot f)]. \tag{2.1}$$

Considering only the term $R(k \cdot f)$ in the summation, this equation takes the contribution of the N_{harm} first harmonics into account. It could then be expected that this expression reaches a maximum

for $f = F_0$. However, this is also true for the harmonics present in the range $[F_{0,min}, F_{0,max}]$. For this reason, the substraction by $R((k - \frac{1}{2}) \cdot f)$ allows to significantly reduce the relative importance of the maxima of SRH at the even harmonics. The estimated pitch value F_0^* for a given residual frame is thus the frequency maximizing $SRH(f)$ at that time.

Figure 2.1 displays the typical evolution of SRH for a segment of female voice. The pitch track (around 200 Hz) clearly emerges. Moreover, no particularly high value of SRH is observed during the unvoiced regions of speech. Therefore, SRH can also be used to provide voicing decisions by a simple local thresholding. More precisely, a frame is determined to be voiced if $SRH(F_0^*)$ is greater than a fixed threshold θ. Note that for the comparison with θ, the residual spectrum $R(f)$ needs to be normalized in energy for each frame.

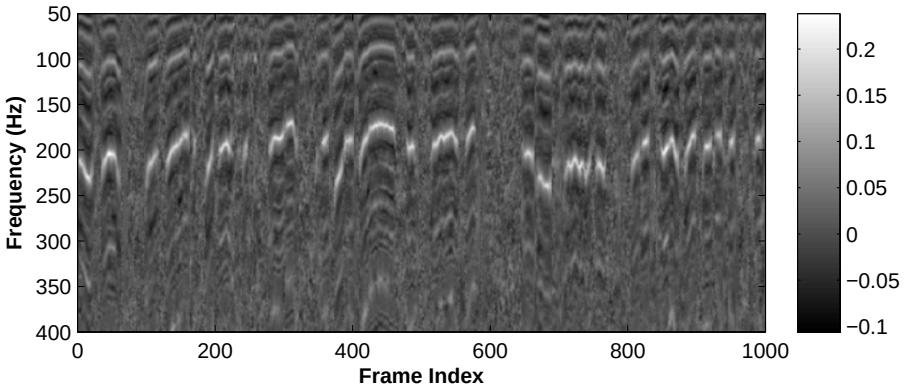

Figure 2.1 - *Evolution of SRH for a segment of clean speech uttered by a female speaker.*

It is worth noting that, in Equation 2.1, the risk of ambiguity with odd harmonics is not addressed. This may be problematic for low-pitched voices for which the third harmonic may be present in the initial range $[F_{0,min}, F_{0,max}]$. Albeit we made several attempts to incorporate a correction in Equation 2.1 by substracting a term in $R((k \pm \frac{1}{3}) \cdot f)$, no improvement was observed (this was especially true in noisy conditions). For this reason, the proposed algorithm works in two steps. In the first step, the described process is performed using the full range $[F_{0,min}, F_{0,max}]$, from which the mean pitch frequency $F_{0,mean}$ of the considered speaker is estimated. In the second step, the final pitch tracking is obtained by applying the same process but in the range $[0.5 \cdot F_{0,mean}; 2 \cdot F_{0,mean}]$. It can be indeed assumed that a normal speaker will not exceed these limits. Note that this idea of restricting the range of F_0 for a given speaker is similar to what has been proposed in [7] (for the choice of the window length).

Figure 2.2 illustrates the proposed method for a segment of female speech, both in clean conditions, and with a Jet noise at 0dB of Signal-to-Noise Ratio (SNR). In the top plot, the pitch ground truth and the estimated fundamental frequency F_0^* are displayed. A close agreement between the estimates and the reference can be noticed during voiced speech. Interestingly, this is true for both clean and noisy speech (except on a short period of 5 frames where F_0^* is half the actual fundamental frequency). It is worth noting that no post-correction of the pitch estimation, using for example dynamic programing, was applied. In the bottom plot, the values of $SRH(F_0^*)$, together with the ideal voiced-unvoiced decisions, are exhibited since they are used for determining the voicing boundaries. It is observed that $SRH(F_0^*)$ conveys a high amount of information about the voicing decisions. However, in adverse conditions, since the relative importance of harmonics becomes weaker with the presence of noise, the values of $SRH(F_0^*)$ are smaller during voiced regions, making consequently the decisions more difficult.

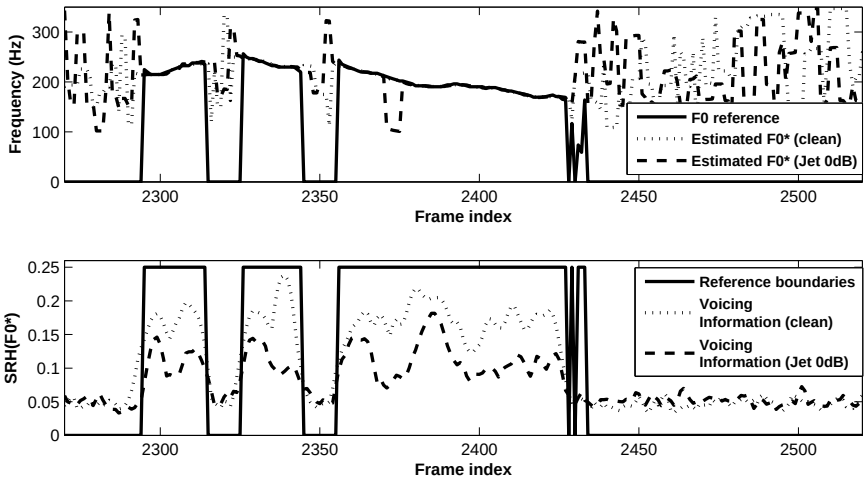

Figure 2.2 - *Illustration of the proposed method in clean and noisy speech (using a jet noise with a SNR of 0 dB).* Top plot : *The pitch ground truth and the estimates* F_0^*. Bottom plot : *The ideal voicing decisions and the values of* $SRH(F_0^*)$.

2.3 Experiments

2.3.1 Experimental Protocol

The experimental protocol is divided into two steps: training and testing. The goal of the training phase is to optimize the several parameters used by the proposed algorithm described in Section 2.2. During the testing, the proposed method is compared to other state-of-the-art methods of pitch tracking, both in clean and noisy conditions. For assessing the performance of a given method, the four following measures are used [8]:

- The **Voicing Decision Error (VDE)** is the proportion of frames for which an error of the voicing decision is made.
- The **Gross Pitch Error (GPE)** is the proportion of frames, where the decisions of both the pitch tracker and the ground truth are voiced, for which the relative error of F_0 is higher than a threshold of 20%.
- The **Fine Pitch Error (FPE)** is defined as the standard deviation (in %) of the distribution of the relative error of F_0 for which this error is below a threshold of 20%.
- The **F0 Frame Error (FFE)** is the proportion of frames for which an error (either according to the GPE or the VDE criterion) is made. FFE can be seen as a single measure for assessing the overall performance of a pitch tracker.

The noisy conditions are simulated by adding to the original speech signal a noise at 0 dB of SNR. The noise signals were taken from the Noisex-92 database [9]. Since the main scope of this chapter is the study of the robustness of pitch trackers, several types of noise were considered: speech babble, car interior, factory, jet cockpit, and white noise.

During the **training** phase, the APLAWD database [10] is used. It consists of ten repetitions of five phonetically balanced English sentences spoken by each of five male and five female talkers, with a

total duration of about 20 minutes. The pitch ground truth was extracted by using the autocorrelation function on the parallel electroglottographic recordings.

For the **testing**, both the Keele and CSTR databases were used, for comparison purpose with other studies. The Keele database [11] contains speech from 10 speakers with five males and five females, with a bit more of 30 seconds per speaker. As for the CSTR database [12], it contains five minutes of speech from one male and one female speaker. For all datasets, recordings sampled at 16kHz were considered, and the provided pitch references were used as a ground truth.

2.3.2 Parameter Optimization for the Proposed Method

In this training step, each parameter is optimized so as to minimize the overall FFE, averaged over all speakers of the APLAWD database, and for both clean and noisy conditions. According to this objective framework, the optimal parameter values are the following. The LPC order for obtaining the residual signal by inverse filtering is set to 12, although it was observed not to have a critical impact in the range between 10 and 18. A too high order tends to overfit the spectral envelope, which may be detrimental in noisy conditions, while a too low value does not sufficiently remove the contributions of both the vocal tract and the noise. The optimal length for framing the residual signal is chosen to be 100 ms (while the frame shift is fixed to 10 ms). To illustrate this, Figure 2.3 shows the impact of the window length on the FFE for clean and noisy conditions. It turns out that a length of 100 ms makes a good compromise for being efficient in any environment. This means that our algorithm requires a large contextual information for performing well. Note that we observed that this does not affect the capabilities of the proposed method to track rapidly-varying pitch contours, maintaining low values of both GPE and FPE. The optimal number of harmonics used in Equation 2.1 is $N_{harm} = 5$. Considering more harmonics is detrimental in adverse conditions, as the noise affects strongly the periodicity of the speech signal, and only the few first harmonic peaks emerge in the spectrum. Finally, the optimal threshold θ used for the voicing decisions is 0.07, as it gave the best tradeoff between false positive and false negative decisions.

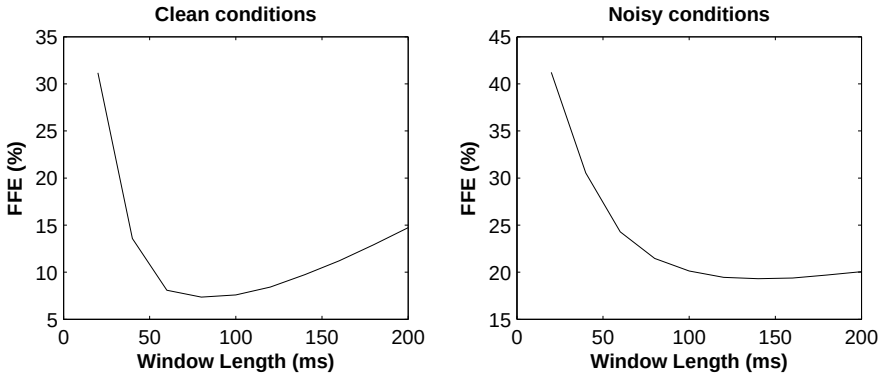

Figure 2.3 - *Influence of the window length on FFE, averaged in clean and noisy conditions.*

2.3.3 Methods compared in this work

In the following, the proposed technique (SRH) is compared to the seven following methods of pitch estimation and tracking:

- **Get_F0**: Included in the ESPS package, this method is an implementation of the RAPT algorithm [13]. In this work, we used the version available in Wavesurfer

<http://www.speech.kth.se/wavesurfer/>.

- **SHRP**: This spectral technique is based on the Subharmonic to Harmonic Ratio, as proposed in [4]. For our tests, we used the implementation available in <http://mel.speech.nwu.edu/sunxj/pda.htm>.
- **TEMPO**: This technique is based on a fixed point analysis [14] and is available in the STRAIGHT toolkit <http://www. wakayama-u.ac.jp/∼kawahara/PSSws/>.
- **AC**: This method relies on an accurate autocorrelation function and is implemented in the Praat toolbox <http://www.praat.org>. It was shown to outperform the original autocorrelation based and the cepstrum-based techniques [15].
- **CC**: This approach makes use of the crosscorrelation function [16] and is also implemented in the Praat toolbox.
- **YIN**: This algorithm is one of the most popular and most efficient method of pitch estimation. It is based on the autocorrelation technique with several modifications that combine to prevent errors [17]. Since YIN only provides F0 estimates, it is here *coupled with the voiced-unvoiced decisions taken by our proposed SRH approach.* The YIN implementation can be freely found at <http://www.auditory.org/postings/2002/26.html>.
- **SSH**: The *Summation of Speech Harmonics* technique is given for comparison purpose as the proposed approach applied this time on the speech signal, and not on its residual as done in SRH. The contribution of the spectral envelope mainly due to the vocal tract is therefore not removed. Note that for SSH the optimal value of the threshold θ is 0.18.

All methods were used with their default parameter values for which they were optimized. The frame shift is fixed to 10 ms, and the range of F0 set to [50 Hz,400 Hz].

2.3.4 Results

Figures 2.4 and 2.5 show a comparison of the FFE (as it is an overall measure for assessing the performance of a pitch tracker) for all methods and in all conditions, respectively for female and male speakers. In clean speech, it is seen that the proposed SSH and SRH methods give a performance comparable to other techniques, while Get_F0 outperforms all other approaches for both male and female speakers. On the opposite, the advantage of SRH is clearly noticed for adverse conditions. In 9 out of the 10 noisy cases (5 noise types and 2 genders), SRH provides better results than existing methods, showing generally an appreciable robustness improvement. The only unfavourable case is the estimation with a Babble noise for male speakers. This may be explained by the fact that this noise highly degrades the speech spectral contents at low frequencies. The five first residual harmonics used by SRH may then be strongly altered, leading to a degradation of performance. Inspecting the performance of SSH, it turns out that it exhibits among the worst results for female speakers in noisy environments, but is almost as efficient as SRH for male voices.

Tables 2.1 and 2.2 present the detailed results of pitch tracking respectively for clean speech, and for noisy conditions (averaged over all noise types at 0dB of SNR). On clean recordings, Get_F0 provides the best results in terms of VDE and FFE on both genders, while the best GPE is obtained by the proposed method SRH for female voices, and by TEMPO for male speakers. Regarding its efficiency in terms of FPE, albeit having the slightly largest values, SRH has a performance sensibly comparable to the state-of-the-art, confirming its ability to also capture the pitch contour details. On noisy speech,

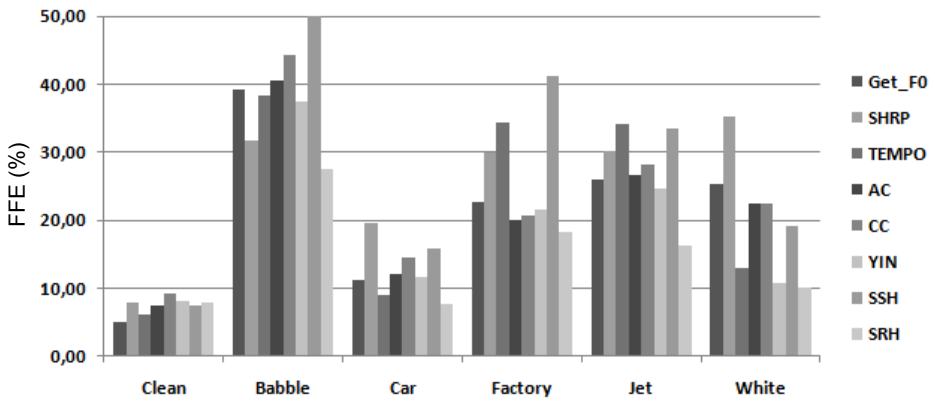

Figure 2.4 - *F0 Frame Error (%) for* ***female speakers*** *and for all methods in six conditions: clean speech and noisy speech at 0dB of SNR with five types of noise.*

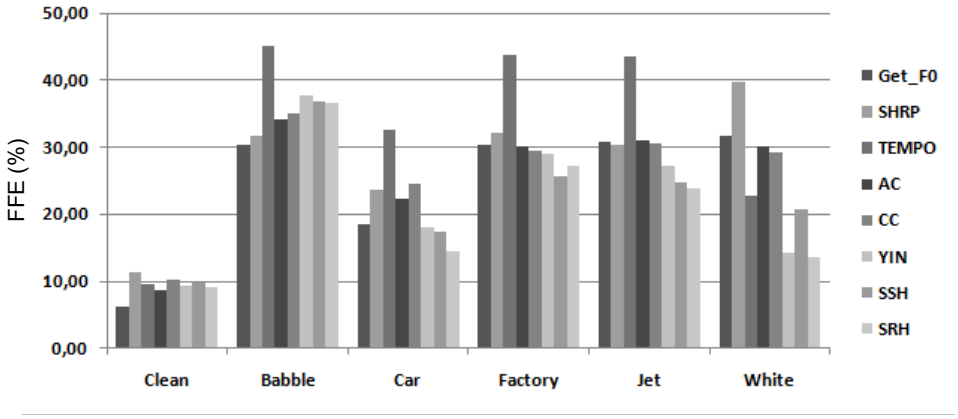

Figure 2.5 - *F0 Frame Error (%) for* ***male speakers*** *and for all methods in six conditions: clean speech and noisy speech at 0dB of SNR with five types of noise.*

SRH clearly outperforms all other approaches, especially for female speakers where the FFE is reduced of at least 8.5% (except for YIN which uses the proposed VAD from SRH). This gain is also substantial for male voices with regard to existing approaches (consequently leaving out of comparison the SSH and the modified YIN techniques), with a decrease of 5.3% of FFE, and of 5.7% regarding the errors on the voicing decisions. It is worth noting the remarkably good performance of SRH for female voices in noisy environments, providing very low values of VDE and GPE (and thus FFE). All methods (except SSH in adverse conditions) are also observed to give better results for female speakers than for male voices. Finally, it is interesting to emphasize that, while relying on the same voicing decisions, YIN leads in all conditions to a greater GPE than SRH, especially for noisy recordings. This confirms the quality of SRH both as a VAD and for pitch contour estimation.

	Female				Male			
	VDE	GPE	FPE	FFE	VDE	GPE	FPE	FFE
Get_F0	**3.74**	2.78	2.95	**4.92**	**5.34**	1.79	3.06	**6.11**
SHRP	7.01	2.03	**2.52**	7.83	10.2	2.74	3.17	11.4
TEMPO	5.38	1.51	3.05	6.01	9.28	**0.93**	3.13	9.66
AC	6.81	1.50	2.68	7.41	8.02	1.40	**2.77**	8.59
CC	8.41	1.76	2.77	9.15	9.25	2.23	3.44	10.2
YIN	7.29	1.88	2.95	8.06	8.34	2.47	2.93	9.38
SSH	5.81	4.67	2.76	7.49	8.87	2.45	3.31	9.88
SRH	7.29	**1.29**	3.10	7.81	8.34	1.95	3.46	9.15

Table 2.1 - *Detailed pitch tracking results in **clean** conditions for both male and female speakers.*

	Female				Male			
	VDE	GPE	FPE	FFE	VDE	GPE	FPE	FFE
Get_F0	20.8	14.8	2.4	24.9	27.7	**2.7**	2.7	28.3
SHRP	27.0	11.5	1.9	29.3	30.1	6.8	2.8	31.5
TEMPO	25.2	4.4	3.9	25.8	36.8	16.7	3.8	37.6
AC	20.5	14.2	2.4	24.3	28.2	5.9	**2.4**	29.6
CC	21.1	18.0	2.7	26.1	27.8	7.9	3.0	29.8
YIN	**15.1**	19.0	3.0	21.2	**22.1**	11.9	2.8	25.2
SSH	24.2	39.1	**1.9**	32.1	23.3	6.3	2.8	25.1
SRH	**15.1**	**2.7**	2.6	**16.0**	**22.1**	4.0	2.7	**23.1**

Table 2.2 - *Detailed pitch tracking results in **noisy** conditions (averaged over all noise types at 0 dB of SNR), for both male and female speakers.*

2.4 Conclusion

This chapter described a simple method of pitch tracking by focusing on the spectrum of the residual signal. A criterion based on the Summation of Residual Harmonics (SRH) is proposed both for pitch estimation and for the determination of voicing boundaries. A comparison with six state-of-the-art pitch trackers is performed in both clean and noisy conditions. A clear advantage of the proposed approach is its robustness to additive noise. In 9 out of the 10 noisy experiments, SRH is shown to lead to a significant improvement, while its performance is comparable to other techniques in clean conditions.

Bibliography

[1] W. Hess, S. Furui, and M. Sondhi. Pitch and voicing determination. In *Advances in Speech Signal Processing, New York*, pages 3–48, 1992.

[2] D. J. Hermes. Measurement of pitch by subharmonic summation. In *JASA*, volume 83, pages 257–264, 1988.

[3] P. Martin. A comparison of pitch detection by cepstrum and spectral comb analysis. In *Proc. ICASSP*, pages 180–183, 1982.

[4] X. Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proc. ICASSP*, volume 1, pages 333–336, 2002.

[5] J. D. Markel. The SIFT algorithm for fundamental frequency estimation. In *IEEE Trans. Audio Electroacoust.*, volume AE-20, pages 367–377, 1972.

[6] L. Tan, B. Borgstrom, and A. Alwan. Voice activity detection using harmonic frequency components in likelihood ratio test. In *Proc. ICASSP*, pages 4466–4469, 2010.

[7] B. Yegnanarayana and K.S.R. Murty. Event-based instantaneous fundamental frequency estimation from speech signals. In *IEEE Trans. Audio, Speech and Language Processing*, volume 17, pages 614–624, 2009.

[8] W. Chu and A. Alwan. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *Proc. ICASSP*, pages 3969–3972, 2009.

[9] Online. Noisex-92. In *http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.html*.

[10] G. Lindsey, A. Breen, and S. Nevard. SPAR's archivable actual-word databases. Technical report, University College London, 1987.

[11] F. Plante, G. Meyer, and W.A. Ainsworth. A pitch extraction reference database. In *Eurospeech*, pages 837–840, 1995.

[12] P. Bagshaw, S. Hiller, and M. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Eurospeech*, pages 1003–1006, 1993.

[13] D. Talkin. A robust algorithm for pitch tracking (RAPT). In *Speech coding and synthesis, Eds.: Elsevier Science*, pages 495–518, 1995.

[14] H. Kawahara, H. Katayose, A. de Cheveigne, and R. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *Proc. Eurospeech*, volume 6, pages 2781–2784, 1999.

[15] P. Boersma. Accurate short-term analysis of fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. Inst. Phonetic Sci.*, volume 17, pages 97–110, 1993.

[16] R. Goldberg and L. Riek. A practical handbook of speech coders. In *Boca Raton, FL: CRC*, 2000.

[17] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

Chapter 3

Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review

Contents

Abstract

The pseudo-periodicity of voiced speech can be exploited in several speech processing applications. This requires however that the precise locations of the Glottal Closure Instants (GCIs) are available. The focus of this chapter is the evaluation of automatic methods for the detection of GCIs directly from the speech waveform. A new procedure to determine GCIs, called the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm, is proposed. The procedure is divided into two successive steps. First a mean-based signal is computed, and intervals where GCIs are expected to occur are extracted from it. Secondly, at each interval a precise position of the GCI is assigned by locating a discontinuity in the Linear Prediction residual. SEDREAMS is compared to four state-of-the-art GCI detection algorithms using six different databases with contemporaneous electroglottographic recordings as ground truth, and containing many hours of speech by multiple speakers. The four techniques to which SEDREAMS is compared are the Hilbert Envelope-based detection (HE), the Zero Frequency Resonator-based method (ZFR), the Dynamic Programming Phase Slope Algorithm (DYPSA) and the Yet Another GCI Algorithm (YAGA). The efficacy of these methods is first evaluated on clean speech, both in terms of reliabililty and accuracy. Their robustness to additive noise and to reverberation is also assessed. A further contribution of this chapter is the evaluation of their performance on a concrete application of speech processing: the causal-anticausal decomposition of speech. It is shown that for clean speech, SEDREAMS and YAGA are the best performing techniques, both in terms of identification rate and accuracy. ZFR and SEDREAMS also show a superior robustness to additive noise and reverberation.

This chapter is based upon the following publications:

- Thomas Drugman, Thierry Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Interspeech Conference, Brighton, United Kingdom, 2009.
- Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, Thierry Dutoit, *Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*

Many thanks to Dr. Mark Thomas (Imperial College of London), Dr. Jon Gudnason (University of Iceland) and Prof. Patrick Naylor (Imperial College of London) for their fruitful collaboration on the quantitative review of methods for Glottal Closure Instant detection.

3.1 Introduction

Glottal-synchronous speech processing is a field of speech science in which the pseudoperiodicity of voiced speech is exploited. Research into the tracking of pitch contours has proven useful in the field of phonetics [1] and speech quality assessment [2]; however more recent efforts in the detection of Glottal Closure Instants (GCIs) enable the estimation of both pitch contours and, additionally, the boundaries of individual cycles of speech. Such information has been put to practical use in applications including prosodic speech modification [3], speech dereverberation [4], glottal flow estimation [5], speech synthesis [6], [7], data-driven voice source modelling [8] and causal-anticausal deconvolution of speech signals [9].

Increased interest in glottal-synchronous speech processing has brought about a corresponding demand for automatic and reliable detection of GCIs from both clean speech and speech that has been corrupted by acoustic noise sources and/or reverberation. Early approaches that search for maxima in the autocorrelation function of the speech signal [10] were found to be unreliable due to formant frequencies causing multiple maxima. More recent methods search for discontinuities in the linear production model of speech [11] by deconvolving the excitation signal and vocal tract filter with Linear Predictive Coding (LPC) [12]. Preliminary efforts are documented in [5]; more recent algorithms use known features of speech to achieve more reliable detection [13, 14, 15]. Deconvolution of the vocal tract and excitation signal by homomorphic processing [16] has also been used for GCI detection although its efficacy compared with LPC has not been fully researched. Various studies have shown that, while linear model-based approaches can give accurate results on clean speech, reverberation can be particularly detrimental to performance [4, 17].

Methods that use smoothing or measures of energy in speech signal are also common. These include the Hilbert Envelope [18], Frobenius Norm [19], Zero-Frequency Resonator (ZFR) [20] and SEDREAMS [21]. Smoothing of the speech signal is advantageous because the vocal tract resonances, additive noise and reverberation are attenuated while the periodicity of the speech signal is preserved. A disadvantage lies in the ambiguity of the precise time instant of the GCI; for this reason LP residual can be used in addition to smoothed speech to obtain more accurate estimates [14, 21]. Smoothing on multiple dyadic scales is exploited by wavelet decomposition of the speech signal with the Multiscale Product [22] and Lines of Maximum Amplitudes (LOMA) [23] to achieve both accuracy and robustness. The YAGA algorithm [15] employs both multiscale processing and the linear speech model.

The aim of this chapter is two-fold. First a new algorithm of GCI determination from the speech signal, called SEDREAMS, is proposed. The second goal is to provide a review and objective evaluation with four contemporary methods for GCI detection, namely Hilbert Envelope-based method [18], DYPSA [14], ZFR [20] and YAGA [15] algorithms. These techniques were chosen as they were shown to be currently among the best performing GCI estimation methods, and since they rely on very different approaches. They are here evaluated together with SEDREAMS against reference GCIs provided by an Electroglottograph (EGG) signal on six databases, of combined duration 232 minutes, containing contemporaneous recordings of EGG and speech. Performance is also evaluated in the presence of additive noise and reverberation. A novel contribution of this chapter is the application of the algorithms to causal-anticausal deconvolution [9], which provides additional insight into their performance in a real-world problem.

The remainder of this chapter is organised as follows. In Section 3.2 the four state-of-the-art algorithms under test are described. The new proposed approach for GCI estimation is presented in Section 3.3. The evaluation techniques are described in Section 3.4. Sections 3.5 and 3.6 discuss the performance results on clean and noisy/reverberant speech respectively. Section 3.7 compares the methods in terms of computational complexity. Finally conclusions are given in Section 3.8.

3.2 Methods Compared in this Chapter

This section presents four of the main representative state-of-the-art methods for automatically detecting GCIs from speech waveforms. These techniques are detailed here below and their reliability, accuracy, robustness and computational complexity will be compared in Sections 3.5, 3.6 and 3.7 to our new method of GCI detection described in Section 3.3. It is worth noting at this point that all methods assume a positive polarity of the speech signal. Polarity should then be verified and corrected if required, using an algorithm such as [24].

3.2.1 Hilbert Envelope-based method

Several approaches relying on the Hilbert Envelope (HE) have been proposed in the literature [25, 26, 27]. In this chapter, a method based on the HE of the Linear Prediction (LP) residual signal (i.e the signal whitened by inverse filtering after removing an auto-regressive modeling of the spectral envelope) is considered.

Figure 3.1 illustrates the principle of this method for a short segment of voiced speech (Fig.3.1(a)). The corresponding synchronized derivative of the ElectroGlottoGraph (dEGG) is displayed in Fig.3.1(e), as it is informative about the actual positions of both GCIs (instants where the dEGG has a large positive value) and GOIs (instants of weaker negative peaks between two successive GCIs). The LP residual signal (shown in Fig.3.1(b)) contains clear peaks around the GCI locations. Indeed the impulse-like nature of the excitation at GCIs is reflected by discontinuities in this signal. It is also observed that for some glottal cycles (particularly before 170 ms or beyond 280 ms) the LP residual also presents clear discontinuities around GOIs. The resulting HE of the LP residual, containing large positive peaks when the excitation presents discontinuities, and its Center of Gravity (CoG)-based signal are respectively exhibited in Figures 3.1(c) and 3.1(d). Denoting $H_e(n)$ the Hilbert envelope of the residue at sample index n, the CoG-based signal is defined as:

$$CoG(n) = \frac{\sum_{m=-N}^{N} m \cdot w(m) H_e(n+m)}{\sum_{m=-N}^{N} w(m) H_e(n+m)} \tag{3.1}$$

where $w(m)$ is a windowing function of length $2N+1$. In this work a Blackman window whose length is 1.1 times the mean pitch period of the considered speaker was used. We empirically reported in our experiments that using this window length led to a good compromise between misses and false alarms (i.e to the best reliability performance). Once the CoG-based signal is computed, GCI locations correspond to the instants of negative zero-crossing. The resulting GCI positions obtained for the speech segment are indicated in the top of Fig.3.1(e). It is clearly noticed that the possible ambiguity with the discontinuities around GOIs is removed by using the CoG-based signal.

3.2.2 The DYPSA algorithm

The Dynamic Programming Phase Slope Algorithm (DYPSA) [14] estimates GCIs by the identification of peaks in the linear prediction residual of speech in a similar way to the HE method. It consists of two main components: estimation of GCI candidates with the group delay function of the LP residual and N-best dynamic programming. These components are defined as follows.

Group Delay Function

The group delay function is the average slope of the unwrapped phase spectrum of the short time Fourier transform of the LP residual [28] [29]. It can be shown to accurately identify impulsive features in

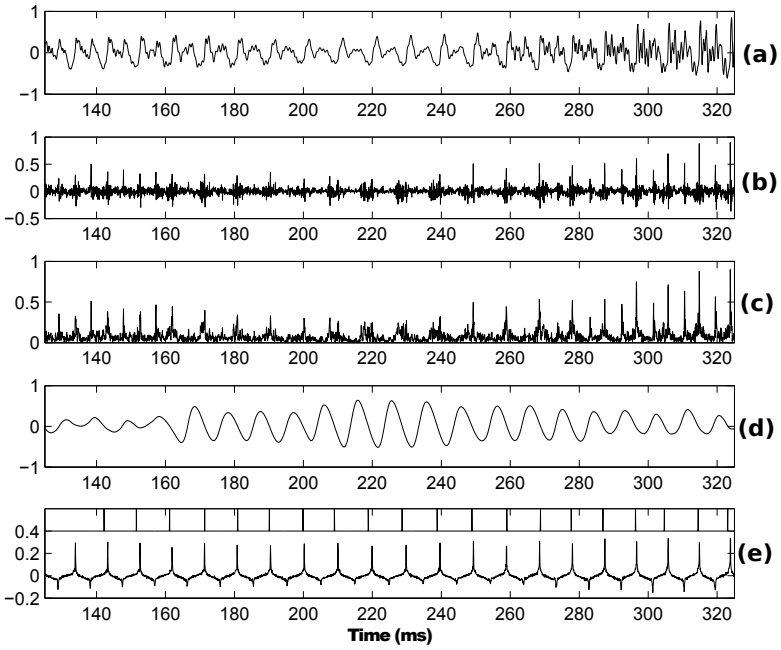

Figure 3.1 - *Illustration of GCI detection using the Hilbert Envelope-based method on a segment of voiced speech.* (a): *the speech signal,* (b): *the LP residual signal,* (c): *the Hilbert Envelope (HE) of the LP residue,* (d): *the Center of Gravity-based signal computed from the HE,* (e): *the synchronized differenced EGG with the GCI positions located by the HE-based method.*

a function provided their minimum separation is known. GCI candidates are selected based on the negative-going zero crossings of the group delay function. Consider an LP residual signal, $e(n)$, and an R-sample windowed segment $x_n(r)$ beginning at sample n:

$$x_n(r) = w(r)e(n+r) \text{ for } r = 0, \ldots, R-1 \tag{3.2}$$

where $w(r)$ is a windowing function. The group delay of $x_n(r)$ is given by [28]:

$$\tau_n(k) = \frac{-\mathrm{d}\arg(X_n)}{\mathrm{d}\omega} = \Re\left(\frac{\tilde{X}_n(k)}{X_n(k)}\right) \tag{3.3}$$

where $X_n(k)$ is the Fourier transform of $x_n(r)$ and $\tilde{X}_n(k)$ is the Fourier transform of $rx_n(r)$. If $x_n(r) = \delta(r-r_0)$, where $\delta(r)$ is a unit impulse function, it follows from Equation (3.3) that $\tau_n(k) \equiv r_0 \forall k$. In the presence of noise, $\tau_n(k)$ becomes noisy, therefore an averaging procedure is performed over k. Different approaches are reviewed in [29]. The *Energy-Weighted Group Delay* is defined as:

$$d(n) = \frac{\sum_{k=0}^{R-1} |X_n(k)|^2 \tau_n(k)}{\sum_{k=0}^{R-1} |X_n(k)|^2} - \frac{R-1}{2}. \tag{3.4}$$

Manipulation yields the simplified expression:

$$d(n) = \frac{\sum_{r=0}^{R-1} r x_n^2(r)}{\sum_{r=0}^{R-1} x_n^2(r)} - \frac{R-1}{2} \tag{3.5}$$

which is an efficient time-domain formulation and can be viewed as a centre of gravity of $x_n(r)$, bounded in the range $[-(R-1)/2, (R-1)/2]$. The location of the negative-going zero crossings of $d(n)$ give an accurate estimation of the location of a peak in a function.

It can be shown that the signal $d(n)$ does not always produce a negative-going zero crossing when an impulsive feature occurs in $e(n)$. In such cases, it has been observed that $d(n)$ consistently exhibits local minima followed by local maxima in the vicinity of the impulsive feature [14]. A *phase-slope projection* technique is therefore introduced to estimate the time of the impulsive feature by finding the midpoint between local maxima and minima where no zero crossing is produced, then projecting a line onto the time axis with negative unit slope.

Dynamic Programming

Erroneous GCI candidates are removed using known characteristics of voiced speech by minimising a cost function so as to select a subset of the GCI candidates which most likely correspond to true GCIs. The subset of candidates is selected according by minimising the following cost function:

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \boldsymbol{\lambda}^T \mathbf{c}_{\Omega}(r), \tag{3.6}$$

where Ω is a subset with GCI candidates of size $|\Omega|$ selected to produce minimum cost, $\boldsymbol{\lambda} = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S]^T = [0.8\ 0.5\ 0.4\ 0.3\ 0.1]^T$ is a vector of weighting factors, the choice of which is described in [14], and $\mathbf{c}(r) = [c_A(r)\ c_P(r)\ c_J(r)\ c_F(r)\ c_S(r)]^T$ is a vector of cost elements evaluated at the rth element of Ω. The cost vector elements are:

- *Speech waveform similarity*, $c_A(r)$, between neighbouring candidates, where candidates not correlated with the previous candidate are penalised.
- *Pitch deviation*, $c_P(r)$, between the current and the previous two candidates, where candidates with large deviation are penalised.
- *Projected candidate cost*, $c_J(r)$, for the candidates from the phase-slope projection, which often arise from erroneous peaks.
- *Normalised energy*, $c_F(r)$, which penalises candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $c_S(r)$, where candidates arising from zero-crossings with gradients close to unity are favoured.

3.2.3 The Zero Frequency Resonator-based technique

The Zero Frequency Resonator-based (ZFR) technique relies on the observation that the impulsive nature of the excitation at GCIs is reflected across all frequencies [20]. The GCI positions can be detected by confining the analysis around a single frequency. More precisely, the method focuses the analysis on the output of zero frequency resonators to guarantee that the influence of vocal-tract resonances is minimal and, consequently, that the output of the zero frequency resonators is mainly controlled by the excitation pulses. The zero frequency-filtered signal (denoted $y(n)$ here below) is obtained from the speech waveform $s(n)$ by the following operations [20]:

1. Remove from the speech signal the dc or low-frequency bias during recording:

$$x(n) = s(n) - s(n-1) \tag{3.7}$$

2. Pass this signal two times through an ideal zero-frequency resonator:

$$y_1(n) = x(n) + 2 \cdot y_1(n-1) - y_1(n-2) \tag{3.8}$$

$$y_2(n) = y_1(n) + 2 \cdot y_2(n-1) - y_2(n-2) \tag{3.9}$$

   The two passages are necessary for minimizing the influence of the vocal tract resonances in $y_2(n)$.

3. As the resulting signal $y_2(n)$ is exponentially increasing or decreasing after this filtering, its trend is removed by a mean-substraction operation:

$$y(n) = y_2(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2(n+m) \tag{3.10}$$

   where the window length $2N + 1$ was reported in [20] to be not very critical, as long as it is in the range of about 1 to 2 times the average pitch period $\bar{T}_{0,mean}$ of the considered speaker. Accordingly, we used in this study a window whose length is $1.5 \cdot \bar{T}_{0,mean}$. Note also that this operation of mean removal has to be repeated three times in order to avoid any residual drift of $y(n)$.

An illustration of the resulting zero frequency-filtered signal is displayed in Fig. 3.2(b) for our example. This signal is observed to possess two advantageous properties: 1) it oscillates at the local pitch period, 2) the positive zero-crossings of this signal correspond to the GCI positions. This is confirmed in Fig. 3.2(c), where a good agreement is noticed between the GCI locations identified by the ZFR technique and the actual discontinuities in the synchronized dEGG.

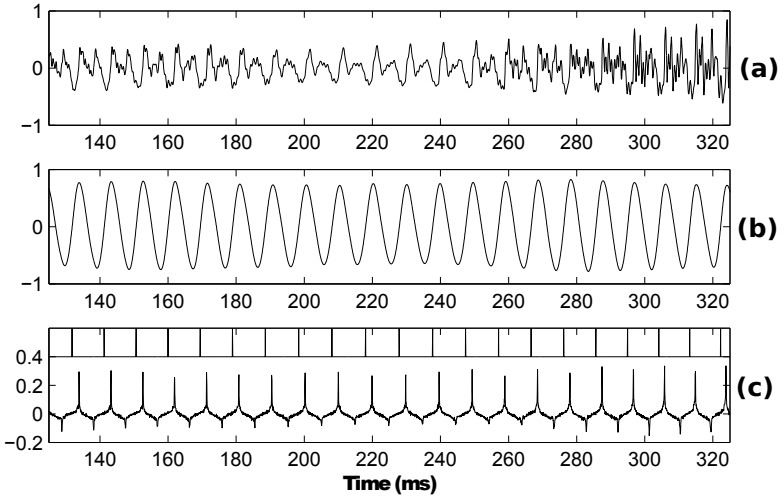

Figure 3.2 - *Illustration of GCI detection using the Zero Frequency Resonator-based method on a segment of voiced speech.* (a): *the speech signal,* (b): *the zero frequency-filtered signal,* (c): *the synchronized dEGG with the GCI positions located by the ZFR-based method.*

3.2.4 The YAGA algorithm

The Yet Another GCI Algorithm (YAGA) [15], like DYPSA, is an LP-based approach that employs N-best dynamic programming to find the best path through a set of candidate GCIs. The algorithms differ in the way in which the candidate set is estimated. Candidates are derived in DYPSA using a linear prediction residual, calculated by inverse-filtering a preemphasised speech signal with the LP coefficients. GCIs are manifest as impulsive features that may be detected with the group delay function. In YAGA, candidates are derived from an estimate of the voice source signal $u'(n)$ by using the same LP coefficients to inverse-filter the non-preemphasized speech signal. This differs crucially in that it exhibits discontinuities at both GCIs and GOIs, although GOIs are not considered in this chapter. The speech signal $s(n)$ and voice source signal $u'(n)$ are shown for a short speech sample in Fig. 3.3 (a) and (b) respectively.

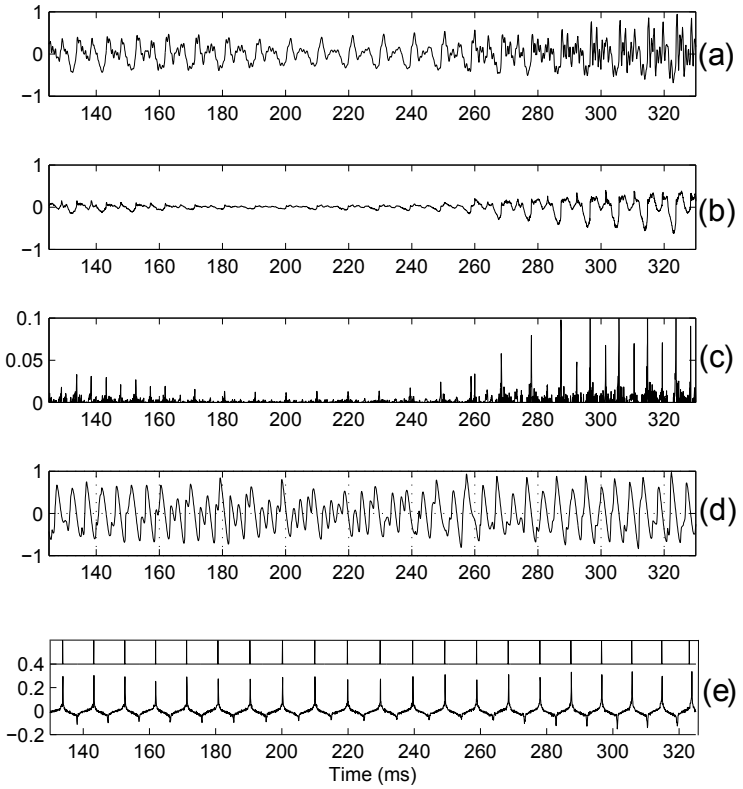

Figure 3.3 - *Illustration of GCI detection using the YAGA algorithm on a segment of voiced speech.* (a): *the speech signal,* (b): *the corresponding voice source signal,* (c): *the multiscale product of the voice source,* (d): *the group-delay function,* (e): *the synchronized dEGG with the GCI positions located by the YAGA algorithm.*

The impulsive nature of the LPC residual is well-suited to detection with the group delay method as discussed in Section 3.2.2. In order for the group delay method to be applied to voice source signal, a discontinuity detector that yields an impulse-like signal is required. Such a detector might be achieved by a 1st-order differentiator, however it is known that GCIs and GOIs are not instantaneous discontinuities but are instead spread over time [22]. The Stationary Wavelet Transform (SWT) is a multiscale analysis tool for the detection of discontinuities in a signal by considering the product of the signal at different scales [30]. It was first used in the context of GCI detection in [22] by application

to the speech signal. YAGA employs a similar approach on the voice source signal, which is expected to yield better results as it is free from unwanted vocal tract resonances. The SWT of signal $u'(n)$, $1 \leq n \leq N$ at scale j is:

$$\begin{aligned} d_j^s(n) &= W_{2^j} u'(n), \\ &= \sum_k g_j(k) a_{j-1}^s(n-k), \end{aligned} \tag{3.11}$$

where the maximum scale J is bounded by $\log_2 N$ and $j = 1, 2, \ldots, J-1$. The approximation coefficients are given by:

$$a_j^s(n) = \sum_k h_j(k) a_{j-1}^s(n-k), \tag{3.12}$$

where $a_0^s(n) = u'(n)$ and $g_j(k)$, $h_j(k)$ are detail and approximation filters respectively that are upsampled by two on each iteration to effect a change of scale [30]. Filters are derived from a biorthogonal spline wavelet with one vanishing moment [30]. The multiscale product, $p(n)$, is formed by:

$$p(n) = \prod_{j=1}^{j_1} d_j(n) = \prod_{j=1}^{j_1} W_{2^j} u'(n), \tag{3.13}$$

where it is assumed that the lowest scale to include is always 1. The de-noising effect of the approximation filters each scale in conjunction with the multiscale product means that $p(n)$ is near-zero except at discontinuities across the first j_1 scales of $u'(n)$ where it becomes impulse-like. The value of j_1 is bounded by J, but in practice $j_1 = 3$ gives good localization of discontinuities in acoustic signals [31].

The multiscale product of the voice source signal in Fig. 3.3 (b) is shown in plot (c). Impulse-like features can be seen in the vicinity of discontinuities of $u'(n)$; such features are then detected by the negative-going zero-crossings of the group delay function in plot (d) that form the candidate set of GCIs. In order to distinguish between GCIs, GOIs and false candidates, an N-best dynamic programming algorithm is applied. The cost function employed is similar to that of DYPSA with an improved waveform similarity measure and an additional element to reliably differentiate between GCIs and GOIs.

3.3 A New Method for GCI Detection: the SEDREAMS Algorithm

We here propose a new technique for automatically determining the GCI locations from the speech signal: the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm. We have shown in [21] that it is a reliable and accurate method for locating both GCIs and GOIs (although in a less accurate way) from the speech waveform. Since the present study only focuses on GCIs, the determination of GOI locations by the SEDREAMS algorithm is omitted. The two steps involved in this method are: *i)* the determination of short intervals where GCIs are expected to occur and *ii)* the refinement of the GCI locations within these intervals. These two steps are described in the following subsections. The SEDREAMS algorithm will then be compared in the rest of this chapter to the four state-of-the-art methods presented in Section 3.2.

3.3.1 Determining intervals of presence using a mean-based signal

As highlighted by the ZFR technique [20], a discontinuity in the excitation is reflected over the whole spectral band, including the zero frequency. Inspired by this observation, the analysis is focused on a mean-based signal. Denoting the speech waveform as $s(n)$, the mean-based signal $y(n)$ is defined as:

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m) \tag{3.14}$$

where $w(m)$ is a windowing function of length $2N+1$. While the choice of the window shape is not critical (a typical Blackman window is used in this study), its length influences the time response of this filtering operation, and may then affect the reliability of the method. The impact of the window length on the misidentification rate is illustrated in Figure 3.4 for the female speaker SLT from the CMU ARCTIC database [32]. Optimality is seen as a trade-off between two opposite effects. A too short window causes the appearance of spurious extrema in the mean-based signal, giving birth to false alarms. On the other hand, a too large window smooths it, affecting in this way the miss rate. However we clearly observed for the three speakers a valley between 1.5 and 2 times the average pitch period $T_{0,mean}$. Throughout the rest of this thesis we used for SEDREAMS a window whose length is $1.75 \cdot T_{0,mean}$.

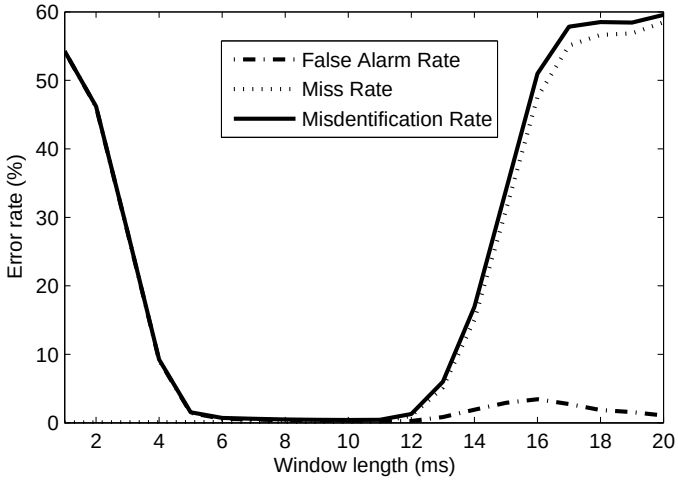

Figure 3.4 - *Effect of the window length used by SEDREAMS on the misidentification rate for the speaker SLT, whose average pitch period is 5.7 ms.*

A segment of voiced speech and its corresponding mean-based signal using an appropriate window length are illustrated in Figs. 3.5(a) and 3.5(b). Interestingly it is observed that the mean-based signal oscillates at the local pitch period. However the mean-based signal is not sufficient in itself for accurately locating GCIs. Indeed, consider Fig. 3.6 where, for five different speakers, the distributions of the actual GCI positions (extracted from synchronized EGG recordings) are displayed within a normalized cycle of the mean-based signal. It turns out that GCIs may occur at a non-constant relative position within the cycle. However, once minima and maxima of the mean-based signal are located, it is straightforward to derive short intervals of presence where GCIs are expected to occur. More precisely, as observed in Fig. 3.6, these intervals are defined as the timespan starting at the minimum of the mean-based signal, and whose length is 0.35 times the local pitch period (i.e the period between two consecutive minima). Such intervals are illustrated in Fig.3.5(c) for our example.

3.3.2 Refining GCI locations using the residual excitation

Intervals of presence obtained in the previous step give fuzzy short regions where a GCI should happen. The goal of the next step is to refine, for each of these intervals, the precise location of the GCI occuring

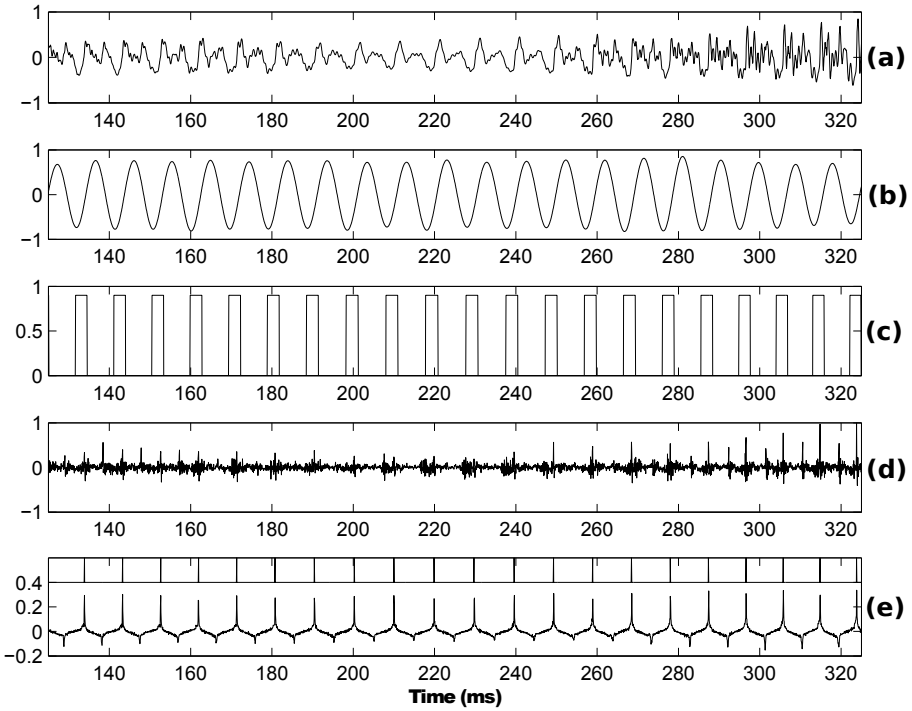

Figure 3.5 - *Illustration of GCI detection using the SEDREAMS algorithm on a segment of voiced speech.* (a): *the speech signal,* (b): *the mean-based signal,* (c): *intervals of presence derived from the mean-based signal,* (d): *the LP residual signal,* (e): *the synchronized dEGG with the GCI positions located by the SEDREAMS algorithm.*

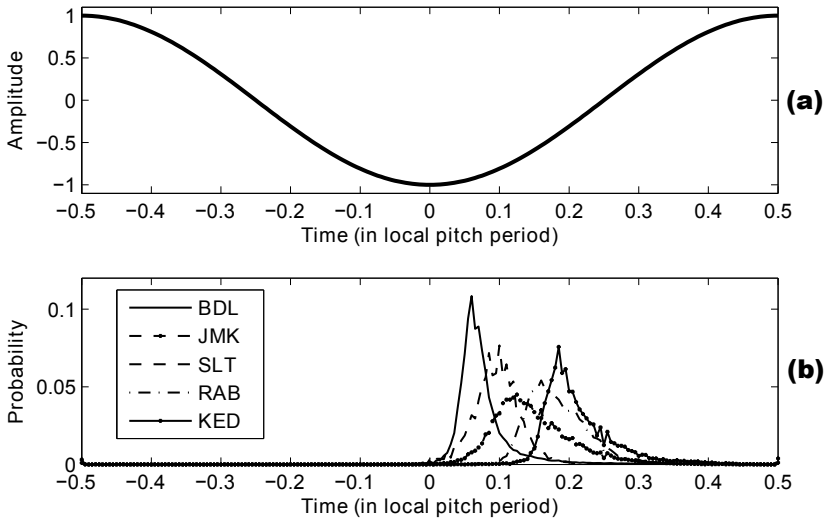

Figure 3.6 - *Distributions, for five speakers, of the actual GCI positions (plot* (b)*) within a normalized cycle of the mean-based signal (plot* (a)*).*

inside it. The LP residual is therefore inspected, assuming that the largest discontinuity of this signal within a given interval corresponds to the GCI location.

Figs. 3.5(d) and 3.5(e) show the LP residual and the time-aligned dEGG for our example. It is clearly noted that combining the intervals extracted from the mean-based signal with a peak picking method on the LP residue allows the accurate and unambiguous detection of GCIs (as indicated in Fig.3.5(e)).

It is worth noting that the advantage of using the mean-based signal is two-fold. First of all, since it oscillates at the local pitch period, this signal guarantees good performance in terms of reliability (i.e the risk of misses or false alarms is limited). Secondly, the intervals of presence that are derived from this signal imply that the GCI timing error is bounded by the depth of these intervals (i.e 0.35 times the local pitch period).

3.4 Assessment of GCI Extraction Techniques

3.4.1 Speech Material

The evaluation of the GCI detection methods relies on ground-truth obtained from EGG recordings. The methods are compared on six large corpora containing contemporaneous EGG recordings whose description is summarized in Table 3.1. The first three corpora come from the CMU ARCTIC databases [32]. They were collected at the Language Technologies Institute at Carnegie Mellon University with the goal of developing unit selection speech synthesizers. Each phonetically balanced dataset contains 1150 sentences uttered by a single speaker: BDL (US male), JMK (US male) and SLT (US female). The fourth corpus consists of a set of nonsense words containing all phone-phone transitions for English, uttered by the UK male speaker RAB. The fifth corpus is the KED Timit database and contains 453 utterances spoken by a US male speaker. These five first databases are freely available on the Festvox webpage [32]. The sixth corpus is the APLAWD dataset [33] which contains ten repetitions of five phonetically balanced English sentences spoken by each of five male and five female talkers. For each of these six corpora, the speech and EGG signals sampled at 16 kHz are considered. The APLAWD database contains a square wave calibration signal for correcting low-frequency phase distortion, introduced in the recording chain, with an allpass equalization filter [34]. While this is particularly important in the field of voice source estimation and modelling [35], we have found GCI detection to be relatively insensitive to such phase distortion. An intuitive explanation is that the glottal excitation at the GCI excites many high-frequency bins such that low-frequency distortion does not have a significant effect upon the timing of the estimated GCI.

Dataset	**Speaker(s)**	**Approximative duration**
BDL	1 male	54 min.
JMK	1 male	55 min.
SLT	1 female	54 min.
RAB	1 male	29 min.
KED	1 male	20 min.
APLAWD	5 males - 5 females	20 min.
Total	9 males - 6 females	232 min.

Table 3.1 - *Description of the databases.*

3.4.2 Objective Evaluation

The most common way to assess the performance of GCI detection techniques is to compare the estimates with the reference locations extracted from EGG signals (Section 3.4.2). Besides it is also proposed to evaluate also their efficiency on a specific application of speech processing: the causal-anticausal deconvolution (Section 3.4.2).

Comparison with Electroglottographic Signals

Electroglottography (EGG), also known as electrolaryngography, is a non-intrusive technique for measuring the time-varying impedance between the vocal folds. The EGG signal is obtained by passing a weak electrical current between a pair of electrodes placed in contact with the skin on both sides of the larynx. This measure is proportionate to the contact area of the vocal folds. As clearly seen in the explanatory figures of Section 3.2, true positions of GCIs can then be easily detected by locating the greatest positive peaks in the differenced EGG signal. Note that, for the automatic assessment, EGG signals need to be time-aligned with speech signals by compensating the delay between the EGG and the microphone. This was done in this work by a manual verification for each database (inside which the delay is assumed to remain constant).

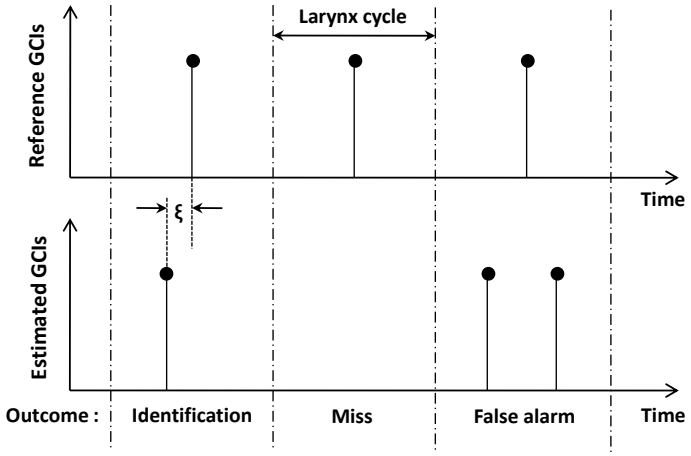

Figure 3.7 - *Characterization of GCI estimates showing three glottal cycles with examples of each possible outcome from GCI estimation [14]. Identification accuracy is characterized by ξ.*

Performance of a GCI detection method can be evaluated by comparing the locations that are estimated with the synchronized reference positions derived from the EGG recording. For this, we here make use of the performance measure defined in [14], presented with the help of Fig. 3.7. The first three measures describe how *reliable* the algorithm is in identifying GCIs:

- the Identification Rate (IDR): the proportion of glottal cycles for which exactly one GCI is detected,
- the Miss Rate (MR): the proportion of glottal cycles for which no GCI is detected,
- and the False Alarm Rate (FAR): the proportion of glottal cycles for which more than one GCI is detected.

For each correct GCI detection (i.e respecting the IDR criterion), a timing error ξ is made with reference to the EGG-derived GCI position. When analyzing a given dataset with a particular method of GCI detection, ξ has a probability density comparable to the histograms of Fig. 3.10 (which will be detailed later in this chapter). Such a distribution can be characterized by the following measures for quantifying the *accuracy* of the method [14]:

- the Identification Accuracy (IDA): the standard deviation of the distribution,
- the Accuracy to $\pm$ 0.25 ms: the proportion of detections for which the timing error is smaller than this bound.

A Speech Processing Application: the Causal-Anticausal Deconvolution

Causal-anticausal decomposition (also known as mixed-phase decomposition) is a non-parametric technique of source-tract deconvolution known to be highly sensitive to GCI location errors [9]. It can therefore be employed as a framework for assessing our methods of GCI extraction on a speech processing application. A complete study of the causal-anticausal decomposition is given in Chapter 5. Nonetheless, we here provide a short background on this concept, necessary to the good understanding of the results exhibited in Section 3.5.2.

The principle of the causal-anticausal decomposition relies on the mixed-phase model of speech [36], [9]. According to this model, voiced speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e anticausal) components. While the vocal tract response and the glottal *return phase* can be considered as minimum-phase signals, it has been shown [36] that the glottal *open phase* is a maximum-phase signal (see Section 5.2 for further explanations). The key idea of the causal-anticausal (or mixed-phase) decomposition is then to separate both minimum and maximum-phase components of speech, where the latter is only due to the glottal contribution. By isolating the anticausal component of speech, causal-anticausal separation allows to estimate the glottal open phase.

It is emphasized in Section 5.4 that windowing is crucial and dramatically conditions the efficiency of the causal-anticausal decomposition. It is indeed essential that the window applied to the segment of voiced speech respects some constraints in order to exhibit correct mixed-phase properties. Among these constraints, the window should be synchronized on a GCI, and have an appropriate shape and length (proportional to the pitch period). If the windowing is such that the speech segment respects the properties of the mixed-phase model, a correct deconvolution is achieved and the anticausal component gives a reliable estimate of the glottal flow (i.e which corroborates the models of the glottal source, such as the Liljencrants-Fant (LF) model [37]), as illustrated in Fig. 3.8(a). On the contrary, if this is not the case (possibly due to the fact that the window is not perfectly synchronized with the GCI), the causal-anticausal decomposition fails, and the resulting anticausal component generally contains an irrelevant high-frequency noise (see Fig.3.8(b)).

As a simple (but accurate) criterion for deciding whether a frame has been correctly decomposed or not, the spectral center of gravity of the anticausal component is investigated. For a given dataset, this feature has a distribution as the one displayed in Fig. 3.9. A principal mode around 2 kHz clearly emerges and corresponds to the majority of frames for which a correct decomposition is carried out (as in Fig.3.8(a)). A second mode at higher frequencies is also observed. It is related to the frames where the causal-anticausal decomposition fails, leading to a maximum-phase signal containing an irrelevant high-frequency noise (as in Fig.3.8(b)). It can be noticed from this histogram that fixing a threshold at around 2.7 kHz optimally discriminate frames that are correctly and incorrectly decomposed.

In conclusion, it is expected that the use of good GCI estimates reduces the proportion of frames that are incorrectly decomposed using the causal-anticausal separation.

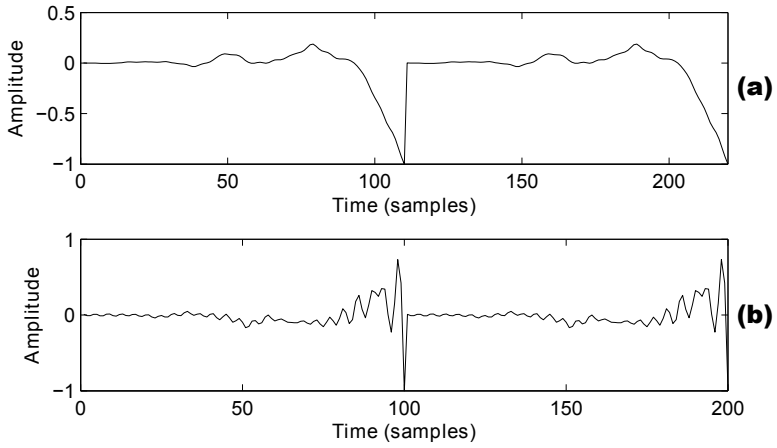

Figure 3.8 - *Two cycles of the anticausal component isolated by mixed-phase decomposition (a): when the speech segment exhibits characteristics of the mixed-phase model, (b): when this is not the case.*

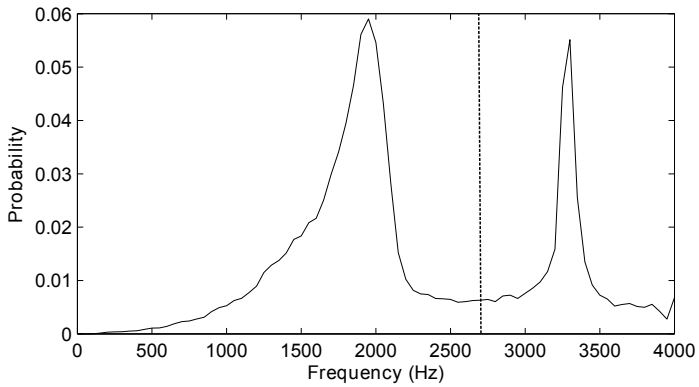

Figure 3.9 - *Example of distribution for the spectral center of gravity of the maximum-phase component. Fixing a threshold around 2.7kHz makes a good separation between correctly and incorrectly decomposed frames.*

3.5 Experiments on Clean Speech Data

Based on the experimental protocol described in Section 3.4, the performance of the four methods of GCI detection introduced in Section 3.2 and the SEDREAMS algorithm (Section 3.3) are now compared on the original clean speech utterances.

3.5.1 Comparison with Electroglottographic Signals

Results obtained from the comparison with electroglottographic recordings are presented in Table 3.2 for the various databases.

In terms of *reliability* performance, SEDREAMS and YAGA algorithms generally give the highest identification rates. Among others, it turns out that SEDREAMS correctly identifies more than 98% of GCIs for any dataset. This is also true for YAGA, except on the RAB database where it reaches 95.70%. Although the performance of ZFR is below these two techniques for JMK, RAB and KED speakers, its results are rather similar on other datasets, obtaining even the best reliability scores on

Database	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to ±0.25ms (%)
BDL	HE	97.04	1.93	1.03	0.58	46.24
	DYPSA	95.54	2.12	2.34	0.42	83.74
	ZFR	97.97	1.05	**0.98**	0.30	80.93
	SEDREAMS	98.08	0.77	1.15	0.31	89.35
	YAGA	**98.43**	**0.39**	1.18	**0.29**	**90.31**
JMK	HE	93.01	3.94	3.05	0.90	38.66
	DYPSA	98.26	0.88	0.86	0.46	77.26
	ZFR	96.17	3.43	**0.4**	0.60	41.62
	SEDREAMS	**99.29**	**0.25**	0.46	0.42	80.78
	YAGA	99.13	0.27	0.60	**0.40**	**81.05**
SLT	HE	96.16	2.83	1.01	0.56	52.46
	DYPSA	97.18	1.41	1.41	0.44	72.17
	ZFR	**99.26**	0.15	**0.59**	**0.22**	83.70
	SEDREAMS	99.15	**0.12**	0.73	0.30	81.35
	YAGA	98.90	0.20	0.90	0.28	**86.18**
RAB	HE	92.08	2.55	5.37	0.78	38.67
	DYPSA	82.33	1.87	15.80	0.46	86.76
	ZFR	92.94	6.31	0.75	0.56	55.87
	SEDREAMS	**98.87**	**0.63**	**0.50**	**0.37**	**91.26**
	YAGA	95.70	0.47	3.83	0.49	89.77
KED	HE	94.73	1.75	3.52	0.56	65.81
	DYPSA	97.24	1.56	1.20	0.34	89.46
	ZFR	87.36	7.90	4.74	0.63	46.82
	SEDREAMS	**98.65**	0.67	**0.68**	**0.33**	94.65
	YAGA	98.21	**0.63**	1.16	0.34	**95.14**
APLAWD	HE	91.74	5.64	2.62	0.73	54.20
	DYPSA	96.12	2.24	1.64	0.59	77.82
	ZFR	**98.89**	0.59	0.52	0.55	57.87
	SEDREAMS	98.67	0.82	**0.51**	**0.45**	85.15
	YAGA	98.88	**0.52**	0.60	0.49	**85.51**

Table 3.2 - *Summary of the performance of the five methods of GCI estimation for the six databases.*

SLT and APLAWD. As for the DYPSA method, its performance remains behind SEDREAMS and YAGA, albeit it reaches IDRs comprised between 95.54% and 98.26%, except for the RAB speaker where the technique fails, leading to an important amount of false alarms (15.80%). Finally the HE-based approach is outperformed by all other methods most of the time. However it achieves on all databases identification rates, comprised between 91.74% and 97.04%.

In terms of *accuracy*, it is observed on all the databases, except for the RAB speaker, that YAGA leads the highest rates of frames for which the timing error is lower than 0.25 ms. The SEDREAMS algorithm gives almost comparable accuracy performance, just below the accuracy of YAGA. The DYPSA and HE algorithms, are outperformed by YAGA and SEDREAMS on all datasets. As it was the case for the reliability results, the accuracy of ZFR strongly depends on the considered speaker. It achieves very good results on the BDL and SLT speakers even though the overall accuracy is rather low especially for the KED corpus.

The accuracy performance is illustrated in Fig. 3.10 for the five measures. The distributions of the GCI identification error ξ is averaged over all datasets. The histograms for the SEDREAMS and YAGA methods are the sharpest and are highly similar. It is worth pointing out that some discrepancy is expected even if the GCI methods identify the acoustic events with high accuracy, since the delay between the speech signal, recorded by the microphone, and the EGG does not remain constant during recordings.

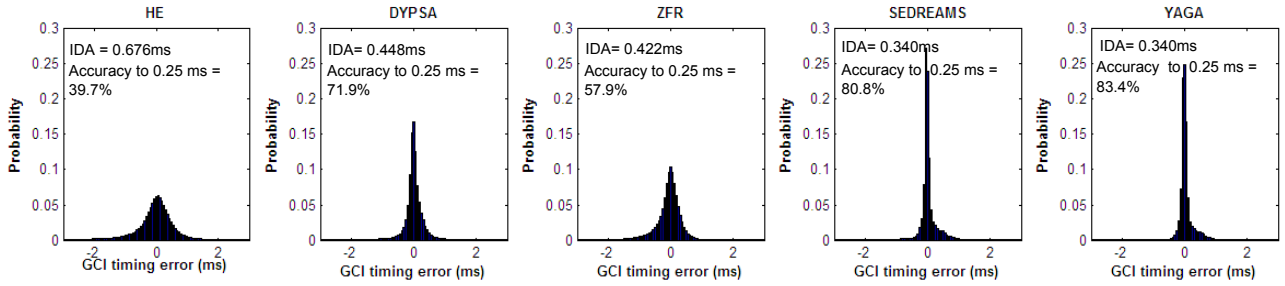

Figure 3.10 - *Histograms of the GCI timing error averaged over all databases for the five compared techniques.*

In conclusion from the results of Table 3.2, the SEDREAMS and YAGA techniques, with highly similar performance, generally outperform other methods of GCI detection on clean speech, both in terms of reliability and accuracy. The ZFR method can also reach comparable (or even slightly better) results for some databases, but its performance is observed to be strongly sensitive to the considered speaker. In general, these three approaches are respectively followed by the DYPSA algorithm and the HE-based method.

3.5.2 Performance based on Causal-Anticausal Deconvolution

As introduced in Section 3.4.2, the Causal-Anticausal deconvolution is a well-suited approach for evaluating our techniques of GCI determination on a concrete application of speech processing. It was indeed emphasized that this method of glottal flow estimation is highly sensitive to GCI location errors. Besides we presented in Section 3.4.2 an objective spectral criterion for deciding whether the mixed-phase separation fails or not. It is important to note at this point that the constraint of precise GCI-synchronization is a necessary, but not sufficient, condition for having a correct deconvolution.

Figure 3.11 displays, for all databases and GCI estimation techniques, the proportion of speech frames that are incorrectly decomposed via mixed-phase separation (achieved in this work by the complex cepstrum-based algorithm [38]). It can be observed that for all datasets (except for SLT), SEDREAMS and YAGA outperform other approaches and lead again to almost the same results. They are closely followed by the DYPSA algorithm whose accuracy was also shown to be quite high in the previous section. The ZFR method turns out to be generally outperformed by these three latter techniques, but still gives the best results on the SLT voice. Finally, it is seen that the HE-based approach leads to the highest rates of incorrectly decomposed frames. Interestingly, these results achieved in the applicative context of the mixed-phase deconvolution corroborate the conclusions drawn from the comparison with EGG signals, especially regarding their accuracy to ± 0.25 ms (see Section 3.5.1). This means that the choice of an efficient technique of GCI estimation, as those compared in this work, may significantly improve the performance of applications of speech processing for which a pitch-synchronous analysis or synthesis is required.

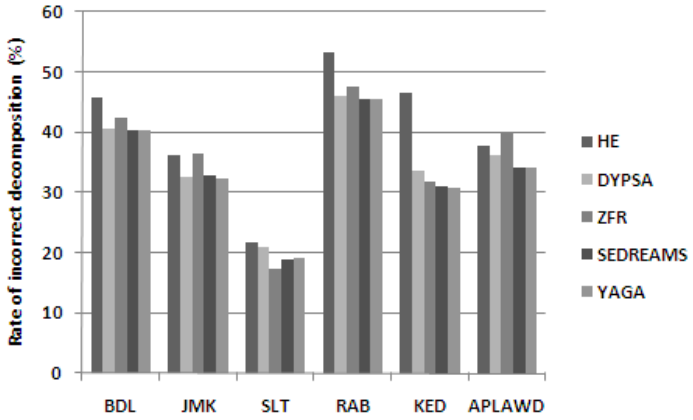

Figure 3.11 - *Proportion of speech frames leading to an incorrect mixed-phase deconvolution using all GCI estimation techniques on all databases.*

3.6 Robustness of GCI Extraction Methods

In some speech processing applications, such as speech synthesis, utterances are recorded in well controlled conditions. For such high-quality speech signals, the performance of GCI estimation techniques was studied in Section 3.5. For many other types of speech processing systems however, there is no other choice than capturing the speech signal in a *real world environment*, where noise and/or reverberation may dramatically degrade its quality. The goal of this section is to evaluate how GCI detection methods are affected by additive noise (Section 3.6.1) and by reverberation (Section 3.6.2). Note that results presented here below were averaged over the six databases.

3.6.1 Robustness to an Additive Noise

In a first experiment, noise was added to the original speech waveform at various Signal-to-Noise Ratio (SNR). Both a White Gaussian Noise (WGN) and a babble noise (also known as cocktail party noise) were considered. The noise signals were taken from the Noisex-92 database [39], and were added so as to control the segmental SNR without silence removal. Results for these two noise types are exhibited in Figs. 3.12 and 3.13 according to the measures detailed in Section 3.4.2. In these figures, miss rate and false alarm rate are in logarithmic scale for the sake of clarity. It is observed that, for both noise types, the general trends remain unchanged. However it turns out that the degradation of reliability is more severe with the white noise, while the accuracy is more affected by the babble noise.

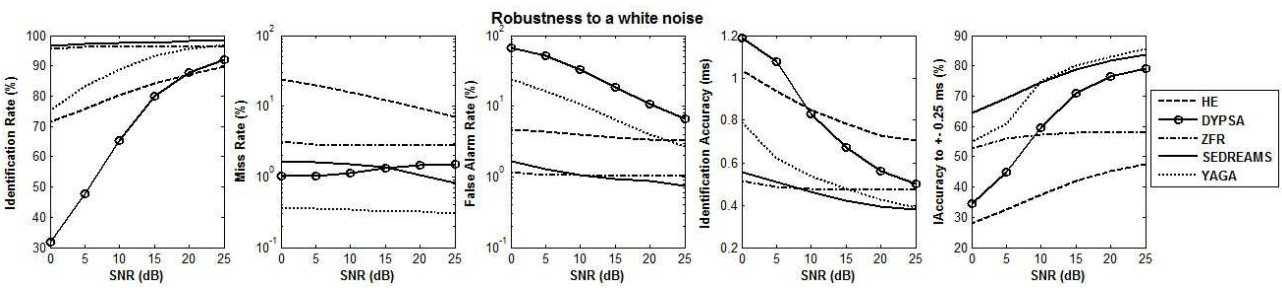

Figure 3.12 - *Robustness of GCI estimation methods to an additive white noise, according to the five measures of performance. Miss rate and false alarm rate are in logarithmic scale.*

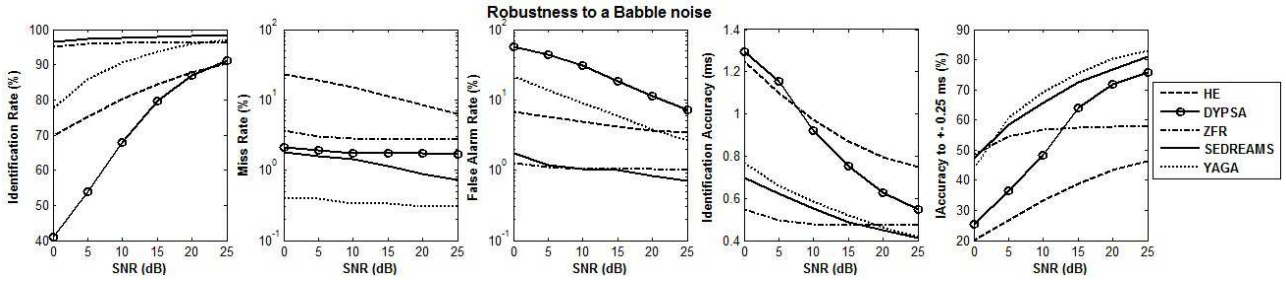

Figure 3.13 - *Robustness of GCI estimation methods to an additive babble noise, according to the five measures of performance. Miss rate and false alarm rate are in logarithmic scale.*

In terms of reliability, it is noticed that SEDREAMS and ZFR lead to the best robustness, since their performance is almost unchanged up to 0dB of SNR. Secondly, the degradation for YAGA and HE is almost equivalent, while it is noticed that DYPSA is strongly affected by additive noise. Among others, it is observed that HE is characterized by an increasing miss rate as the noise level increases, while the degradation is reflected by an increasing number of false alarms for DYPSA, and for YAGA in a lesser extent. This latter observation is probably due to the difficulty of the dynamic programing process to deal with spurious GCI candidates caused by the additive noise.

Regarding the accuracy capabilities, similar conclusions hold. Nevertheless the sensitivity of SEDREAMS is this time comparable to that of YAGA and HE. Again, the ZFR algorithm is found to be the most robust technique, while DYPSA is the one presenting the strongest degradation and HE displays the worst identification accuracy.

The good robustness of ZFR and SEDREAMS can be explained by the low sensitivity of respectively the zero-frequency resonators and the mean-based signal to an additive noise. In the case of ZFR, analysis is confined around 0 Hz, which tends to minimize not only the effect of the vocal tract, but of an additive noise as well. As for SEDREAMS, the mean-based signal is computed as in Equation 3.14, which is a linear relation. In other words, the mean-based signal of the noise is added to the mean-based signal of the speech signal. On a duration of $1.75 \cdot \bar{T}_{0,mean}$, the white noise is assumed to be almost zero-mean. A similar conclusion is observed for the babble noise, which is composed of several sources of speech talking at the same time. It can indeed be understood that the higher the number of sources in the babble noise, the lesser its degradation on the target mean-based signal. Finally, the strong sensitivity of DYPSA and YAGA might be explained, among others, by the fact that they rely on some thresholds, which have been optimized for clean speech.

3.6.2 Robustness to Reverberation

In many modern telecommunication applications, speech signals are obtained in enclosed spaces with the talker situated at a distance from the microphone. The received speech signal is distorted by reverberation, caused by reflected signals from walls and hard objects, diminishing intelligibility and perceived speech quality [40, 41]. It has been further observed that the performance of GCI identification algorithms is degraded when applied to reverberant signals [4].

The observation of reverberant speech at microphone m is:

$$x_m(n) = h_m(n) * s(n), \quad m = 1, 2, \ldots, M, \tag{3.15}$$

where $h_m(n)$ is the L-tap Room Impulse Response (RIR) of the acoustic channel between the source to the mth microphone. It has been shown that multiple time-aligned observations with a microphone array can be exploited for GCI estimation in reverberant environments [17]; in this chapter we only

consider the robustness of single-channel algorithms to the observation at channel $x_1(n)$. RIRs are characterised by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging {100, 200, ..., 500} ms was simulated using the source-image method [42] and the simulated impulse responses convolved with the clean speech signals described in Section 3.4.

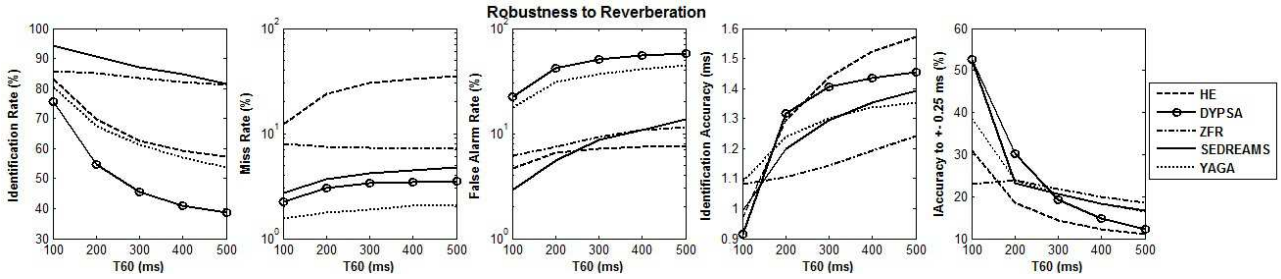

Figure 3.14 - *Robustness of GCI estimation methods to reverberation, according to the five measures of performance. Miss rate and false alarm rate are in logarithmic scale.*

The results in Figure 3.14 show that the performance of the algorithms monotonically reduces with increasing reverberation, with the most significant change in performance occurring between $T_{60} = 100$ and 200 ms. They also reveal that reverberation has a particularly detrimental effect upon identification rate of the LP-based approaches, namely HE, DYPSA and YAGA. This is consistent with previous studies which have shown that the RIR results in additional spurious peaks in the LP residual of similar amplitude to the voiced excitation [43, 44], generally increasing false alarm rate for DYPSA and YAGA but increasing miss rate for HE. Although spurious peaks result in increased false alarms, the identification accuracy of the hits is much less affected. The non-LP approaches generally exhibit better identification rates in reverberation, in particular SEDREAMS. The ZFR algorithm appears to be the least sensitive to reverberation while providing the best overall performance. However, the challenge of GCI detection from single-channel reverberant observations remains an ongoing research problem as no single algorithm consistently provides good results for all five measures.

3.7 Computational Complexity of GCI Extraction Methods

In the previous sections, methods of GCI estimation have been compared according to their reliability and accuracy both in clean conditions (Section 3.5) and noisy/reverberant environments (Section 3.6). In order to provide a complete comparison, an investigation into computational complexity is described in this section. The techniques described in Section 3.2, as well as the SEDREAMS algorithm proposed in Section 3.3, are relatively complex and their computational complexity is highly data-dependent; it is therefore difficult to find a closed-form expression for computational complexity. In this section we discuss those components that present a high computational load and provide a quantitative analysis based upon empirical measurements.

For HE, ZFR and SEDREAMS, the most time-consuming step is the computation of the oscillating signal which they rely on. For the HE method, the CoG-based signal is computed from Equation 3.1 and requires, for each sample, around $2.2 \cdot F_s/\bar{T}_{0,mean}$ multiplications and the same number of additions. For ZFR, the mean removal operation (Equation 3.10) is repeated three times, and thus requires about $4.5 \cdot F_s/\bar{T}_{0,mean}$ additions for each sample of the zero frequency-filtered signal. As for the SEDREAMS algorithm, the computation of each sample of the mean-based signal (Equation 3.14) requires $1.75 \cdot F_s/\bar{T}_{0,mean}$ multiplications and the same number of additions.

However, it is worth emphasizing that the computation time requested by HE and SEDREAMS can be significantly reduced. Indeed these methods only exploit some particular points of the oscillating signal which they rely on: the negative zero-crossings for HE, and the extrema for SEDREAMS. It is then not necessary to compute all the samples of these signals for finding these particular events. Based on this idea, a multiscale approach can be used. For example, the oscillating signals can be first calculated only for the samples multiple of 2^p. From this downsampled signal, a first approximation of the particular points is obtained. This approximation is then refined iteratively using the p successive smaller scales. The lower bounding value of p means there are, for the first approximation, at least two samples per cycle. In the following, we used $p = 4$ so that voices with pitch up to 570 Hz can be processed. The resulting methods are hereafter called *Fast HE* and *Fast SEDREAMS*. Notice that a similar acceleration cannot be transposed to ZFR as the operation of mean removal is applied 3 times successively.

In the case of DYPSA and YAGA, the signal conditioning stages present a relatively low computational load. The LPC residual, Group Delay Function and Multiscale Product scale approximately $\mathcal{O}(N^2)$, $\mathcal{O}(N \log_2 N)$ and $\mathcal{O}(N)$ respectively, where N is the total number of samples in the speech signal. Computational load is significantly heavier in the dynamic programming stages due to the large number of erroneous GCI candidates that must be removed. In particular, the waveform similarity measure, used to determine the similarity of two neighbouring cycles, presents a high computational load due to the large number of executions required to find the optimum path. At present this is calculated on full-band speech although it is expected that calculation of waveform similarity on a downsampled signal may yield similar results for a much-reduced computational load. A second optimization lies in the length of the group delay evaluation window, which is inversely proportional to the number of candidates generated. At present this takes a fixed value based upon the maximum expected F_0; far fewer erroneous candidates could be generated by dynamically varying the length based upon a crude initial estimate of F_0.

So as to compare their computational complexity, the *Relative Computation Time* (RCT) of each GCI estimation method is evaluated on all databases:

$$RCT(\%) = 100 \cdot \frac{\text{CPU time (s)}}{\text{Sound duration (s)}} \tag{3.16}$$

Table 3.3 shows, for both male and female speakers, the averaged RCT obtained for our Matlab implementations and with a Intel Core 2 Duo T7500 2.20 GHz CPU with 3GB of RAM. First of all, it is observed that results are ostensibly the same for both genders. Regarding the non-accelerated versions of the GCI detection methods, it turns out that DYPSA is the fastest (with a RCT around 20%), followed by SEDREAMS and YAGA, which both have a RCT of about 28%. The HE-based technique gives a RCT of around 33%, and ZFR, due to its operation of mean removal which has to be repeated three times, is the slowest method with a RCT of 75%. Interestingly, it is noticed that the accelerated versions of HE and SEDREAMS reduce the computation time by about 5 times on male voices, and by around 4 times for female speakers. This leads to the fastest GCI detection algorithms, reaching a RCT of around 6% for Fast SEDREAMS, and about 8% for Fast HE. Note finally that these results could be highly reduced by using, for example, a C-implementation of these techniques, albeit the conclusions remain identical.

3.8 Conclusion

This chapter proposed a new procedure, called the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm, for detecting the GCIs directly from speech

Method	Male	Female
HE	35.0	31.8
Fast HE	7.6	7.8
DYPSA	19.9	19.4
ZFR	75.7	74.9
SEDREAMS	27.8	27.1
Fast SEDREAMS	5.4	6.9
YAGA	28.6	28.3

Table 3.3 - *Relative Computation Time (RCT), in %, for all methods and for male and female speakers. Results have been averaged across all databases.*

signals. The procedure was divided into two successive steps. The first one computed a mean-based signal and extracted from it intervals where speech events were expected to occur. This step guaranteed good performance in terms of identification rate. The second one refined the location of the speech events within the intervals by inspecting the LP residual, which ensured good performance in terms of identification accuracy.

A major contribution of this chapter was also the comparative evaluation of SEDREAMS with four of the most effective methods for automatically determining GCIs from the speech waveform: Hilbert Envelope-based detection (HE), the Zero Frequency Resonator-based method (ZFR), DYPSA and YAGA. The performance of these methods was assessed on six databases containing several male and female speakers, for a total amount of data of approximately four hours. In our first experiments on clean speech, the SEDREAMS and YAGA algorithms gave the best results, with a comparable performance. For *any* database, they reached an identification rate greater than 98% and more than 80% of GCIs were located with an accuracy of 0.25 ms. Although the ZFR technique can lead to a similar performance, its efficiency can also be rather low in some cases. In general, these three approaches were shown to respectively outperform DYPSA and HE. In a second experiment on clean speech, the impact of the performance of these five methods was studied on a concrete application of speech processing: the causal-anticausal deconvolution. Results showed that adopting a GCI detection with high performance could significantly improve the proportion of correctly deconvolved frames. In the last experiment, the robustness of the five techniques to additive noise, as well as to reverberation was investigated. The ZFR and SEDREAMS algorithms were shown to have the highest robustness, with an almost unchanged reliability. DYPSA was observed to be especially affected, which was reflected by a high rate of false alarms. Although the degradation of accuracy was relatively slow with the level of additive noise, it was noticed that reverberation dramatically affects the precision GCI detection methods. In addition, the computational complexity of the algorithms was studied. A method for accelerating the GCI location using HE and SEDREAMS was proposed. This led, for our Matlab implementation, to a computation time about 6% real-time for the fast version of SEDREAMS.

Depending on the speech application to design, some GCI methods could be preferred to some others, based on their performance for the criteria studied in this chapter. However, if the application is placed in an unknown environment, we suggest the use of SEDREAMS for the following reasons: *i)* it gave the best results with YAGA on clean speech, *ii)* it was the best performing technique in noisy conditions, *iii)* it led with ZFR to the best robustness in a reverberant environment, and *iv)* it was the most suited method for a real-time implementation.

Finally note that we made the SEDREAMS algorithm freely available on the web at:
http://tcts.fpms.ac.be/∼drugman/.

Bibliography

[1] J. C. Catford. *Fundamental Problems in Phonetics*. Indiana University Press, 1977.

[2] Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001.

[3] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5–6):453–467, December 1990.

[4] N. D. Gaubitch and P. A. Naylor. Spatiotemporal averaging method for enhancement of reverberant speech. In *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Cardiff, UK, 2007.

[5] D. Y. Wong, J. D. Markel, and J. A. H Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(4):350–355, August 1979.

[6] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9:21–29, 2001.

[7] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech Conference*, 2009.

[8] M. R. P. Thomas, J. Gudnason, and P. A. Naylor. Data-driven voice source waveform modelling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

[9] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

[10] H. W. Strube. Determination of the instant of glottal closure from the speech wave. *J. Acoust. Soc. Am.*, 56(5):1625–1629, 1974.

[11] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1988.

[12] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580, April 1975.

[13] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech Audio Process.*, 7(5): 569–576, September 1999.

[14] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Trans. Speech Audio Process.*, 15(1):34–43, 2007.

[15] M. R. P. Thomas, J. Gudnason, and P. A. Naylor. Detection of glottal opening and closing instants in voiced speech using the YAGA algorithm. *Submitted for peer review*, 2010.

[16] Pavel Chytil and Misha Pavel. Variability of glottal pulse estimation using cepstral method. In *Proc. 7th Nordic Signal Processing Symposium (NORSIG)*, pages 314–317, 2006. doi: 10.1109/NORSIG.2006.275243.

[17] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor. Multichannel DYPSA for estimation of glottal closure instants in reverberant speech. In *Proc. European Signal Processing Conf. (EUSIPCO)*, 2007.

[18] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust., Speech, Signal Process.*, 27:309–319, 1979.

[19] C. Ma, Y. Kamp, and L. F. Willems. A Frobenius norm approach to glottal closure detection from the speech signal. *IEEE Trans. Speech Audio Process.*, 2:258–265, April 1994.

[20] K. S. R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(8):1602–1613, 2008.

[21] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech Conference*, 2009.

[22] A. Bouzid and N. Ellouze. Glottal opening instant detection from speech signal. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pages 729–732, Vienna, September 2004.

[23] V. N. Tuan and C. d'Alessandro. Robust glottal closure detection using the wavelet transform. In *Eurospeech*, pages 2805–2808, Budapest, September 1999.

[24] Wen Ding and N. Campbell. Determining polarity of speech signals based on gradient of spurious glottal waveforms. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2:857–860, 1998.

[25] T. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust. Speech Signal Proc.*, 27:309–319, 1979.

[26] Y. M. Cheng and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust., Speech, Signal Process.*, 37:1805–1815, December 1989.

[27] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana. Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. *IEEE Signal Process. Lett.*, 14(10):762–765, 2007.

[28] B. Yegnanarayana and R. Smits. A robust method for determining instants of major excitations in voiced speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–779, 1995.

[29] Mike Brookes, Patrick A. Naylor, and Jon Gudnason. A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Trans. Speech Audio Process.*, 14, 2006.

[30] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(7):710–732, 1992.

[31] B. M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. *IEEE Trans. Inf. Theory*, 45(3):1043–1051, 1999.

[32] Online. The festvox website. In http://festvox.org/, .

[33] G. Lindsey, A. Breen, and S. Nevard. SPAR's archivable actual-word databases. Technical report, University College London, 1987.

[34] M. J. Hunt. Automatic correction of low-frequency phase distortion in analogue magnetic recordings. *Acoustics Letters*, 2:6–10, 1978.

[35] K. Funaki and K. Tochinai Y. Miyanaga. Recursive ARMAX speech analysis based on a glottal source model with phase compensation. *Signal Processing*, 74(3):279–295, May 1999.

[36] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *ISCA ITRW VOQUAL03*, pages 15–19, 2003.

[37] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[38] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. Interspeech Conference*, 2009.

[39] Online. Noisex-92. In *http://www.speech.cs. cmu.edu/comp.speech/Sectionl/Data/noisex.html*, .

[40] R. H. Bolt and A. D. MacDonald. Theory of speech masking by reverberation. *J. Acoust. Soc. Am.*, 21:577–580, 1949.

[41] H. Kuttruff. *Room Acoustics*. Taylor & Frances, fourth edition, 2000.

[42] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, 1979.

[43] M. S. Brandstein and D. B. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[44] B. Yegnanarayana and P. Satyanarayana. Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio Process.*, 8(3):267–281, 2000.

Part II

Glottal Flow Estimation and Applications

Chapter 4

Introduction on the Glottal Flow Estimation

Contents

4.1 Glottal Flow Estimation: Problem Positioning

During the mechanism of phonation, an airflow is evicted from the lungs, arises in the trachea and is modulated by its passage through the space delimited by the vocal folds, called the glottis [1]. Speech then results from filtering this so-called glottal flow by the vocal tract cavities, and converting the resulting velocity flow into pressure at the lips [1]. In many speech processing applications, it is important to separate the contributions from the glottis and the vocal tract. Achieving such a *source-filter deconvolution* could lead to a distinct characterization and modeling of these two components, as well as to a better understanding of the human phonation process. Such a decomposition is thus a preliminary condition for the study of glottal-based vocal effects, which can be segmental (as for vocal fry), or be controlled by speakers on a separate, supra-segmental layer. Their dynamics is very different from that of the vocal tract contribution, and requires further investigation. Glottal source estimation is then a fundamental problem in speech processing, finding applications in speech synthesis [2], voice pathology detection [3], speaker recognition [4], emotive speech analysis/synthesis [5], etc.

Glottal flow estimation mainly refers to the estimation of the voiced excitation of the vocal tract. During the production of voiced sounds, the airflow arising from the trachea causes a quasi-periodic vibration of the vocal folds [1], organized into so-called opening/closure cycles. During the *open phase*, vocal folds are progressively displaced from their initial state due to the increasing subglottal pressure. When the elastic displacement limit is reached, they suddenly return to this position during the so-called *return phase*. Figure 4.1 displays the typical shape of one cycle of the glottal flow (Fig.4.1(a))

and its time derivative (Fig.4.1(b)) according to the Liljencrants-Fant (LF) model [6]. It is often preferred to gather the lip radiation effect (whose action is close to a differentiation operator) with the glottal component, and work in this way with the glottal flow derivative on the one hand, and with the vocal tract contribution on the other hand. It is seen in Figure 4.1 (bottom plot) that the boundary between open and return phases corresponds to a particular event called the Glottal Closure Instant (GCI). GCIs refer to the instants of significant excitation of the vocal tract [7]. Being able to determine their location, as done in Chapter 3, is of particular importance in so-called pitch-synchronous speech processing techniques, and in particular for a more accurate separation between vocal tract and glottal contributions.

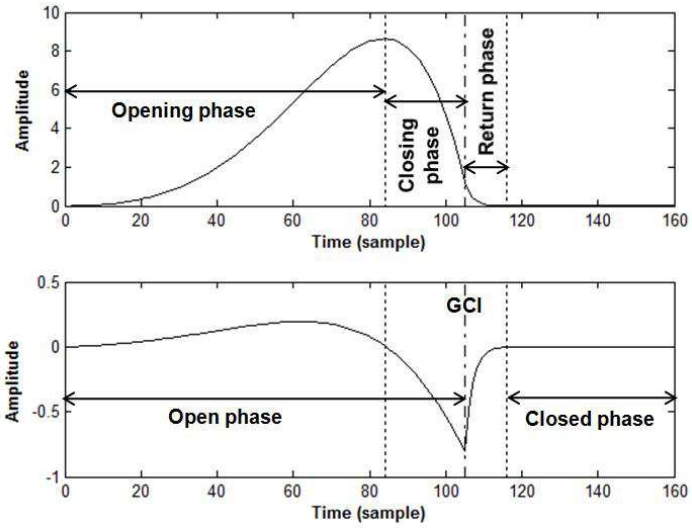

Figure 4.1 - *Typical waveforms, according to the Liljencrants-Fant (LF) model, of one cycle of: (top) the glottal flow, (bottom) the glottal flow derivative. The various phases of the glottal cycle, as well as the Glottal Closure Instant (GCI) are also indicated.*

Several models of the glottal flow have been proposed in the literature. Almost all works devoted to glottal flow modeling are expressed in the time domain. These are, for example, the Klatt [8], the R++ [9], the Rosenberg C [10] or the well-known LF model [6]. These models differ by the analytic expression they use for fitting the glottal waveform, whose shape is in any case close to the one illustrated in Figure 4.1 for the LF model. This latter model makes use of two parameters for describing the glottal open phase: *i)* the *Open Quotient* O_q which is the ratio between the open phase duration and the pitch period, and *ii)* the asymmetry coefficient α_m which is the ratio between the durations of the opening phase and the open phase. As for the return phase, it is generally characterized by its time constant. In contrast, the Causal-Anticausal Linear Model (CALM) proposed in [11] describes glottal signal in the frequency domain.

The glottal flow models and their spectra are compared in [12]. The glottal flow derivative has a magnitude spectrum as illustrated in Figure 4.2. This spectrum is characterized by a low-frequency resonance called *glottal formant*, which is due to the glottal open phase. After the glottal formant frequency, the spectrum of the glottal flow derivative goes down with an asymptotic behaviour of -20dB/decade (or -6dB/octave). The glottal return phase then plays on what is called the *spectral tilt* which makes this asymptotic slope, after a given cut-off frequency (related to the return phase time constant) more severe (to -40dB/decade).

In Part II, we limit our scope to methods which perform an estimation of the glottal source

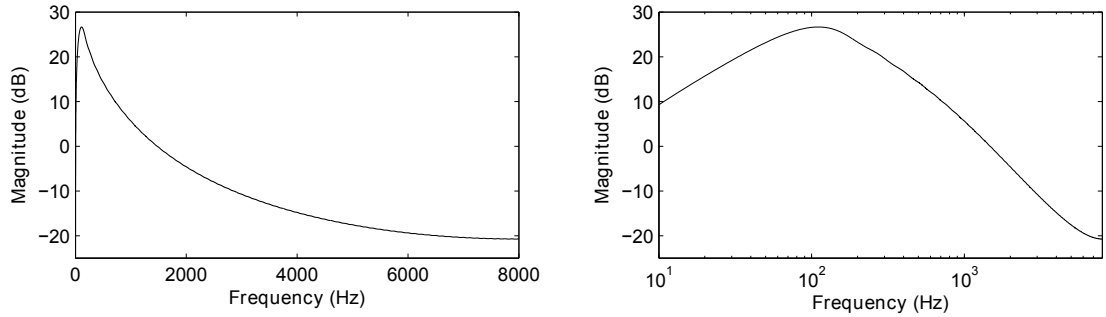

Figure 4.2 - *A typical magnitude spectrum of the glottal flow derivative.* Left plot: *in linear frequency scale,* right plot: *in logarithmic frequency scale. The resonance is called glottal formant.*

contribution directly from the speech waveform. Although some devices such as electroglottographs or laryngographs, which measure the impedance between the vocal folds (but not the glottal flow itself), are informative about the glottal behaviour [13], in most cases the use of such aparatus is inconvenient and only the speech signal is available for analysis. This problem is then a typical case of blind separation, since neither the vocal tract nor the glottal contribution are observable. This also implies that no quantitative assessment of the performance of glottal source estimation techniques is possible on natural speech, as no target reference signal is available.

In order to provide a necessary background for Part II, Sections 4.2 and 4.3 present a brief overview of the existing methods for respectively estimating and parameterizing the glottal source. Section 4.4 then describes the structure of Part II and highlights its main contributions.

4.2 Methods for Glottal Source Estimation

The main techniques for estimating the glottal source directly from the speech waveform are here reviewed. Relying on the speech signal alone, as it is generally the case in real applications, allows to avoid the use of intrusive (e.g video camera at the vocal folds) or inconvenient (e.g. laryngograph) device.

Such techniques can be separated into two classes, according to the way they perform the source-filter separation. The first category (Section 4.2.1) is based on inverse filtering, while the second one (Section 4.2.2) relies on the mixed-phase properties of speech.

4.2.1 Methods based on Inverse Filtering

Most glottal source estimation techniques are based on an inverse filtering process. These methods first estimate a parametric model of the vocal tract, and then obtain the glottal flow by removing the vocal tract contribution via inverse filtering. Methods in this category differ by the way the vocal tract is estimated. Some perform the estimation by focusing the analysis during the glottal closed phase, while others make use of an iterative and/or adaptive procedure. A more extended review of the inverse filtering-based process for glottal waveform analysis can be found in [14].

Closed Phase Inverse Filtering

Closed phase refers to the timespan during which the glottis is closed (see Figure 4.1). During this period, the effects of the subglottal cavities are minimized, providing a better way for estimating the

vocal tract transfer function. Therefore, methods based on a Closed Phase Inverse Filtering (CPIF) estimate a parametric model of the spectral envelope, computed during the estimated closed phase duration [15]. The main drawback of these techniques lies in the difficulty in obtaining an accurate determination of the closed phase. Several approaches have been proposed in the literature to solve this problem. In [16], authors use information from the electroglottographic signal (which is avoided in this study) to identify the period during which the glottis is closed. In [4], it was proposed to determine the closed phase by analyzing the formant frequency modulation between open and closed phases. In [17], the robustness of CPIF to the frame position was improved by imposing some dc gain constraints. Besides this problem of accurate determination of the closed phase, it may happen that this period is so short (for high-pitched voices) that not enough samples are available for a reliable filter estimation. It was therefore proposed in [18] to perform multicycle closed-phase LPC, where a small number of neighbouring glottal cycles are considered in order to have enough data for an accurate vocal tract estimation. Finally note that an approach allowing non-zero glottal wave to exist over closed glottal phases was proposed in [19].

Iterative and/or Adaptive Inverse Filtering

Some methods are based on iterative and/or adaptive procedures in order to improve the quality of the glottal flow estimation. In [20], Fu and Murphy proposed to integrate, within the AutoRegressive eXogenous (ARX) model of speech production, the LF model of the glottal source. The resulting ARXLF model is estimated via an adaptive and iterative optimization [21]. Both source and filter parameters are consequently jointly estimated. The method proposed by Moore in [22] iteratively finds the best candidate for a glottal waveform estimate within a speech frame, without requiring a precise location of the GCIs. Finally a popular approach was proposed by Alku in [23] and called Iterative Adaptive Inverse Filtering (IAIF). This method is based on an iterative refinement of both the vocal tract and the glottal components. In [24], the same authors proposed an improvement, in which the LPC analysis is replaced by the Discrete All Pole (DAP) modeling technique [25], shown to be more accurate for high-pitched voices.

4.2.2 Mixed-Phase Decomposition

A completely different category of glottal flow estimation methods relies on the mixed-phase model of speech [26]. According to this model, speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e anticausal) components. While the vocal tract impulse response and the *return phase* of the glottal component can be considered as minimum-phase signals, it has been shown in [11] that the *open phase* of the glottal flow is a maximum-phase signal. Besides it has been shown in [27] that mixed-phase models are appropriate for modeling voiced speech due to the maximum-phase nature of the glottal excitation. In [27] Garner *et al.* showed that the use of an anticausal all-pole filter for the glottal pulse is necessary to resolve magnitude and phase information correctly. The key idea of mixed-phase decomposition methods is then to separate minimum from maximum-phase components of speech, where the latter is only due to the glottal contribution. In [28], Bozkurt *et al.* presented an algorithm based on the Zeros of the Z-Transform (ZZT) to carry out this deconvolution. In Chapter 5, we will focus on the mixed-phase decomposition and propose a method using Complex Cesptrum (CC) which is functionally equivalent to ZZT, but is strongly advantageous in terms of computation time.

4.3 Glottal Source Parametrization

Once the glottal signal has been estimated by any of the aforementioned algorithms, it is interesting to derive a parametric representation of it, using a small number of features. Various approaches, both in the time and frequency domains, have been proposed to characterize the human voice source. This section gives a brief overview of the most commonly used parameters in the literature, since some of them are used throughout Part II. Some of these features are defined with the help of Figure 4.3. Note that two recent PhD theses have addressed the estimation of glottal parameters from the speech signal [29], [30].

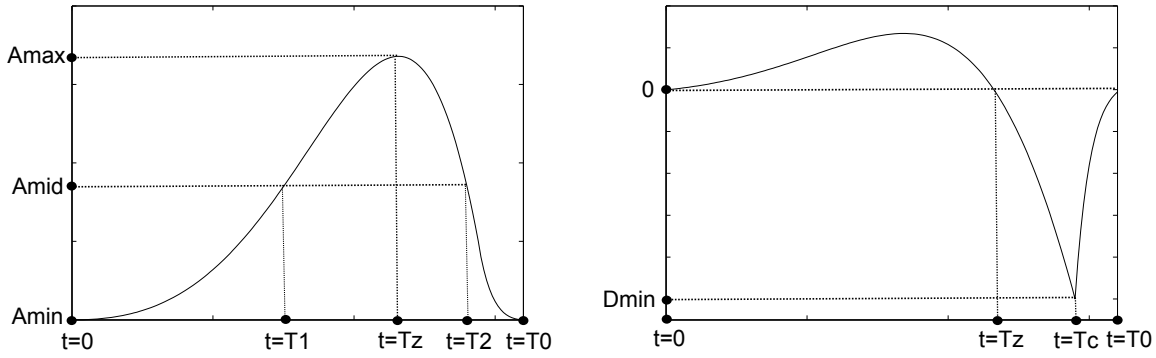

Figure 4.3 - *Particular instants and amplitudes of: (left) the glottal flow, (right) the glottal flow derivative. Amplitude A_{mid} is defined as $A_{mid} = \frac{A_{mid}+A_{min}}{2}$).*

4.3.1 Time-domain features

Several time-domain features can be expressed as a function of time intervals derived from the glottal waveform [31]. These are used to characterize the shape of the waveform, by capturing for example the location of the primary or secondary opening instant [32], of the glottal flow maximum, etc. The formulation of the source signal in the commonly used LF model [6] is based on time-domain parameters, such as the Open Quotient $O_q = \frac{T_c}{T_0}$, the Asymmetry coefficient $\alpha_m = \frac{T_z}{T_c}$, or the Voice Speed Quotient $S_q = \frac{T_z}{T_c - T_z}$ [12]. However in most cases these instants are difficult to locate with precision from the glottal flow estimation. Avoiding this problem and preferred to the traditional Open Quotient, the Quasi-Open Quotient (QOQ) was proposed as a parameter describing the relative open time of the glottis [33]. It is defined as the ratio between the quasi-open time and the quasi-closed time of the glottis, and corresponds to the timespan (normalized to the pitch period) during which the glottal flow is above 50% of the difference between the maximum and minimum flow ($QOQ = \frac{T_2 - T_1}{T_0}$). Note that QOQ was used in [32] for studying the physical variations of the glottal source related to the vocal expression of stress and emotion. In [34] several variants of Oq have been tested in terms of the degree by which they reflect phonation changes. QOQ was found to be the best for this task.

Another set of parameters is extracted from the amplitude of peaks in the glottal pulse or its derivative [35]. The Normalized Amplitude Quotient (NAQ) proposed by Alku in [36] turns out to be an essential glottal feature. NAQ is a parameter characterizing the glottal closing phase [36]. It is defined as the ratio between the maximum of the glottal flow and the minimum of its derivative, normalized with respect to the fundamental period ($NAQ = \frac{A_{max} - A_{min}}{D_{min} \cdot T_0}$). Its robustness and efficiency to separate different types of phonation was shown in [36], [34]. Note that a quasi-similar feature,

called *basic shape parameter*, was proposed by Fant in [37], where it was qualified as *"most effective single measure for describing voice qualities"*.

In [4], authors propose to use 7 LF parameters and 5 energy coefficients (defined in 5 subsegments of the glottal cycle) respectively for characterizing the coarse and fine structures of the glottal flow estimate. Finally some approaches aim at fitting a model on the glottal flow estimate by computing a distance in the time domain [4], [38].

4.3.2 Frequency-domain features

As mentioned in Section 4.1, the spectrum of the LF model presents a low-frequency resonance called the *glottal formant* [12]. Some approaches characterize the glottal formant both in terms of frequency and bandwidth [39]. By defining a spectral error measure, other studies try to match a model to the glottal flow estimation [40], [37], [38]. This is also the case for the Parabolic Spectrum Parameter (PSP) proposed in [41].

An extensively used measure is the $H1-H2$ parameter [37]. This parameter is defined as the ratio between the amplitudes of the magnitude spectrum of the glottal source at the fundamental frequency and at the second harmonic [8], [42]. It has been widely used as a measure characterizing voice quality [43], [37], [17].

For quantifying the amount of harmonics in the glottal source, the Harmonic to Noise Ratio (HNR) and the Harmonic Richness Factor (HRF) have been proposed in [44] and [45]. More precisely, HRF quantifies the amount of harmonics in the magnitude spectrum of the glottal source. It is defined as the ratio between the sum of the amplitudes of harmonics, and the amplitude at the fundamental frequency [46]. It was shown to be informative about the phonation type in [45] and [17].

4.4 Structure and Contributions of Part II

The remaining of Part II is structured as follows. Chapter 5 explains the principle of the mixed-phase decomposition of speech and proposes a new algorithm for achieving it. This method is based on the Complex Cepstrum and it is shown in Chapter 5 that it can efficiently be used for glottal flow estimation. Chapter 6 provides a comparative evaluation of glottal flow estimation methods. The performance of the Complex Cepstrum-based technique is assessed and compared to approaches relying on inverse filtering. In Chapter 7, we suggest to use a chirp z-transform for methods achieving the mixed-phase decomposition. An automatic way of carrying it out is proposed for both the Zeros of the Z-Transform technique, as well as the Complex Cepstrum-based method. The advantage of the chirp approach is that it removes the constraint of being synchronous on a Glottal Closure Instant. The resulting method is then evaluated with regard to its traditional non-chirp equivalent.

The two next chapters focus on the applicability of the methods of glottal flow estimation in two specific fields of speech processing. First, Chapter 8 investigates the use of glottal-based features for the automatic detection of voice disorders. The second application concerns the analysis of expressive speech and is detailed in Chapter 9.

Bibliography

[1] T. Quatieri. *Discrete-time speech signal processing*. Prentice-Hall, 2002.

[2] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi. Glottal spectral separation for parametric speech synthesis. In *Proc. Interspeech*, pages 1829–1832, 2008.

[3] T. Drugman, T. Dubuisson, and T. Dutoit. On the mutual information between source and filter contributions for voice pathology detection. In *Proc. Interspeech*, 2009.

[4] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–586, 1999.

[5] M. Airas and P. Alku. Emotions in vowel segments of continuous speech : Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, 63:26–46, 2006.

[6] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[7] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[8] D. Klatt and L. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[9] R. Veldhuis. A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103:566–571, 1998.

[10] A. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49:583–590, 1971.

[11] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *ISCA ITRW VOQUAL03*, pages 15–19, 2003.

[12] B. Doval and C. d'Alessandro. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006.

[13] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. *J. Acoust. Soc. Am.*, 115:1321–1332, 2004.

[14] J. Walker and P. Murphy. A review of glottal waveform analysis. In *Progress in Nonlinear Speech Processing*, pages 1–21, 2007.

[15] D. Wong, J. Markel, and A. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 27(4), 1979.

[16] D. Veeneman and S. Bement. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 33(2):369–377, 1985.

[17] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[18] D. Brookes and D. Chan. Speaker characteristics from a glottal airflow model using glottal inverse filtering. *Institue of Acoust.*, 15:501–508, 1994.

[19] H. Deng, R. Ward, M. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 14(2):445–455, 2006.

[20] Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(2):492–501, 2006.

[21] D. Vincent, O. Rosec, and T. Chovanel. Estimation of lf glottal source parameters based on an arx model. In *Proc. Interspeech*, pages 333–336, 2005.

[22] E. Moore and M. Clements. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In *Proc. ICASSP*, 2004.

[23] P. Alku, J. Svec, E. Vilkman, and F. Sram. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

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

[25] A. El Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans. on Signal Processing*, 39 (2):411–423, 1991.

[26] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

[27] W. Gardner and B. Rao. Noncausal all-pole modeling of voiced speech. *IEEE Trans. Speech and Audio Processing*, 5(1):1–10, 1997.

[28] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12, 2005.

[29] G. Degottex. *Glottal source and vocal-tract separation.* PhD thesis, UPMC-Ircam, France, 2010.

[30] N. Sturmel. *Analyse de la qualité vocale appliquée à la parole expressive.* PhD thesis, Université Paris Sud 11, Faculté des Sciences d'Orsay, France, 2011.

[31] P. Alku. An automatic method to estimate the time-based parameters of the glottal pulseform. In *Proc. ICASSP*, volume 2, pages 29–32, 1992.

[32] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen. Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics*, 24:313–335, 1996.

[33] T. Hacki. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica*, 41:43–48, 1989.

[34] M. Airas and P. Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. Interspeech*, pages 1410–1413, 2007.

[35] C. Gobl and A. Chasaide. Amplitude-based source parameters for measuring voice quality. In *VOQUAL03*, pages 151–156, 2003.

[36] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112:701–710, 2002.

[37] G. Fant. The lf-model revisited. transformations and frequency domain analysis. *STL-QPSR*, 36 (2-3):119–156, 1995.

[38] T. Drugman, T. Dubuisson, N. d'Alessandro, A. Moinet, and T. Dutoit. Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase. In *16th European Signal Processing Conference*, 2008.

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

[40] Z. Ling, Yu Hu, and R. Wang. A novel source analysis method by matching spectral characters of lf model with straight spectrum. *Lecture Notes in Computer Science*, 3784:441–448, 2005.

[41] P. Alku, H. Strik, and E. Vilkman. Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

[42] I. Titze and J. Sundberg. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.

[43] H. Hanson. Individual variations in glottal characteristics of female speakers. In *Proc. ICASSP*, pages 772–775, 1995.

[44] P. Murphy and O. Akande. Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation. In *Nonlinear Speech Processing Workshop*, pages 224–232, 2005.

[45] D. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:2394–2410, 1991.

[46] D. Childers. *Speech Processing and Synthesis Toolboxes*. Wiley and Sons, Inc., 1999.

Chapter 5

Mixed-Phase Decomposition of Speech using Complex Cepstrum for Glottal Source Estimation

Contents

Abstract

This chapter investigates the possibility of using complex cepstrum for glottal flow estimation. Via a systematic study of the windowing effects on the deconvolution quality, we show that the complex cepstrum causal-anticausal decomposition can be effectively used for glottal flow estimation when specific windowing criteria are met. It is also shown that this complex cepstral decomposition gives similar glottal estimates as obtained with the Zeros of the Z-Transform (ZZT) technique, but uses operations based on the Fast Fourier Transform (FFT) instead of requiring the factoring of high-degree polynomials. The resulting method is consequently much faster for achieving the same decomposition quality. Finally in our tests on a large corpus of real expressive speech, we show that the proposed method has the potential to be used for voice quality analysis.

This chapter is based upon the following publications:

- Thomas Drugman, Baris Bozkurt, Thierry Dutoit, *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*, Interspeech Conference, Brighton, United Kingdom, 2009.
- Thomas Drugman, Baris Bozkurt, Thierry Dutoit, *Causal-Anticausal Decomposition of Speech using Complex Cepstrum for Glottal Source Estimation*, Speech Communication, Volume 53, Issue 6, July 2011, Pages 855-866, 2011.

Many thanks to Dr. Baris Bozkurt (Izmir Institute of Technology) for his helpful guidance.

5.1 Introduction

As mentioned in Chapter 4, glottal source estimation aims at isolating the glottal flow contribution directly from the speech waveform. For this, most of the methods proposed in the literature are based on an inverse filtering process (see Section 4.2.1). These methods first estimate a parametric model of the vocal tract, and then obtain the glottal flow by removing the vocal tract contribution via inverse filtering. The methods in this category differ by the way the vocal tract is estimated. In some approaches [1], [2], this estimation is computed during the glottal closed phase, as the effects of the subglottal cavities are minimized during this period, providing a better way for estimating the vocal tract transfer function. Some other methods (such as [3]) are based on iterative and/or adaptive procedures in order to improve the quality of the glottal flow estimation.

In this chapter we consider a non-parametric decomposition of the speech signal based on the mixed-phase model [4],[5]. According to this model, speech contains a maximum-phase (i.e anticausal) component corresponding to the glottal open phase. In [6], Bozkurt *et al.* proposed an algorithm based on the Zeros of the Z-Transform (ZTT) which has the ability to achieve such a deconvolution. However, the ZZT method suffers from high computational load due to the necessity of factorizing large degree polynomials. It has also been discussed in previous studies that the complex cepstrum had the potential to be used for excitation analysis ([7],[8]) but no technique is yet available for reliable glottal flow estimation. This chapter discusses a complex cepstrum-based method that performs the same operation as the ZZT (i.e. the estimation of the glottal open phase from the speech signal) in a much faster way.

The goal of this chapter is two-fold. First we explain in which conditions complex cepstrum can be used for glottal source estimation. The link with the ZZT-based technique is emphasized and both methods are shown to be two means of achieving the same operation: the causal-anticausal decomposition. However it is shown that the complex cepstrum performs it in a much faster way. Secondly the effects of windowing are studied in a systematic framework. This leads to a set of constraints on the window so that the resulting windowed speech segment exhibits properties described by the mixed-phase model of speech. It should be emphasized that no method is here proposed for estimating the return phase component of the glottal flow signal. As the glottal return phase has a causal character [5], its contribution is mixed in the also causal vocal tract filter contribution of the speech signal.

The chapter is structured as follows. Section 5.2 presents the theoretical framework for the causal-anticausal decomposition of voiced speech signals. Two algorithms achieving this deconvolution, namely the Zeros of the Z-Transform (ZZT) and the Complex Cepstrum (CC) based techniques, are described in Section 5.3. The influence of windowing on the causal-anticausal decomposition is investigated in Section 5.4 by a systematic study on synthetic signals. Relying on the conclusions of this study, it is shown in Section 5.5 that the complex cepstrum can be efficiently used for glottal source estimation on real speech. Among others we demonstrate the potential of this method for voice quality analysis on an expressive speech corpus. Finally Section 5.6 concludes and summarizes the contributions of this chapter.

5.2 Causal-Anticausal Decomposition of Voiced Speech

5.2.1 Mixed-Phase Model of Voiced Speech

It is generally accepted that voiced speech results from the excitation of a linear time-invariant system with impulse response $h(n)$, by a periodic pulse train $p(n)$ [8]:

$$x(n) = p(n) \star h(n), \tag{5.1}$$

where $\star$ denotes the convolution operation. According to the mechanism of voice production, speech is considered as the result of a glottal flow signal filtered by the vocal tract cavities and radiated by the lips. The system transfer function $H(z)$ then consists of the three following contributions:

$$H(z) = A \cdot G(z)V(z)R(z) \tag{5.2}$$

where A is the source gain, $G(z)$ the glottal flow over a single cycle, $V(z)$ the vocal tract transmittance and $R(z)$ the radiation load. The resonant vocal tract contribution is generally represented for "pure" vowels by a set of minimum-phase poles ($|v_{2,k}| < 1$), while modeling nasalized sounds requires to also consider minimum-phase (i.e causal) zeros ($|v_{1,k}| < 1$). $V(z)$ can then be written as the rational form:

$$V(z) = \frac{\prod_{k=1}^{M}(1 - v_{1,k}z^{-1})}{\prod_{k=1}^{N}(1 - v_{2,k}z^{-1})} \tag{5.3}$$

During the production of voiced sounds, airflow arising from the trachea causes a quasi-periodic vibration of the vocal folds [8]. These latter are then subject to quasi-periodic opening/closure cycles. During the *open phase*, vocal folds are progressively displaced from their initial state because of the increasing subglottal pressure [9]. When the elastic displacement limit is reached, they suddenly return to this position during the so-called *return phase*. It has been shown in [10], [5] that the glottal open phase can be modeled by a pair of maximum-phase (i.e anticausal) poles ($|g_2| > 1$) producing the so-called *glottal formant*, while the return phase can be assumed to be a first order causal filter response ($|g_1| < 1$) resulting in a *spectral tilt*:

$$G(z) = \frac{1}{(1 - g_1z^{-1})(1 - g_2z^{-1})(1 - g_2^*z^{-1})} \tag{5.4}$$

As for lip radiation, its action is generally assumed as a differential operator:

$$R(z) = 1 - rz^{-1} \tag{5.5}$$

with r close to 1. For this reason, it is generally preferred to consider $G(z)R(z)$ in combination, and consequently to study the *glottal flow derivative* or *differentiated glottal flow* instead of the glottal flow itself.

Gathering the previous equations, the system z-transform $H(z)$ can be expressed as a rational fraction with general form [7]:

$$H(z) = A\frac{\prod_{k=1}^{M_i}(1 - a_kz^{-1})}{\prod_{k=1}^{N_i}(1 - b_kz^{-1})\prod_{k=1}^{N_o}(1 - c_kz^{-1})} \tag{5.6}$$

where a_k and b_k respectively denote the zeros and poles inside the unit circle ($|a_k|$ and $|b_k| < 1$), while c_k are the poles outside the unit circle ($|c_k| > 1$). The basic idea behind using causal-anticausal decomposition for glottal flow estimation is the following: *since c_k are only related to the glottal flow, isolating the maximum-phase (i.e anticausal) component of voiced speech should then give an estimation of the glottal open phase.* Besides, if the glottal return phase can be considered as abrupt and if the glottal closure is complete, the anticausal contribution of speech corresponds to the glottal flow. If this is not the case [11], these latter components are causal (given their damped nature) and the anticausal contribution of voiced speech still gives an estimation of the glottal open phase.

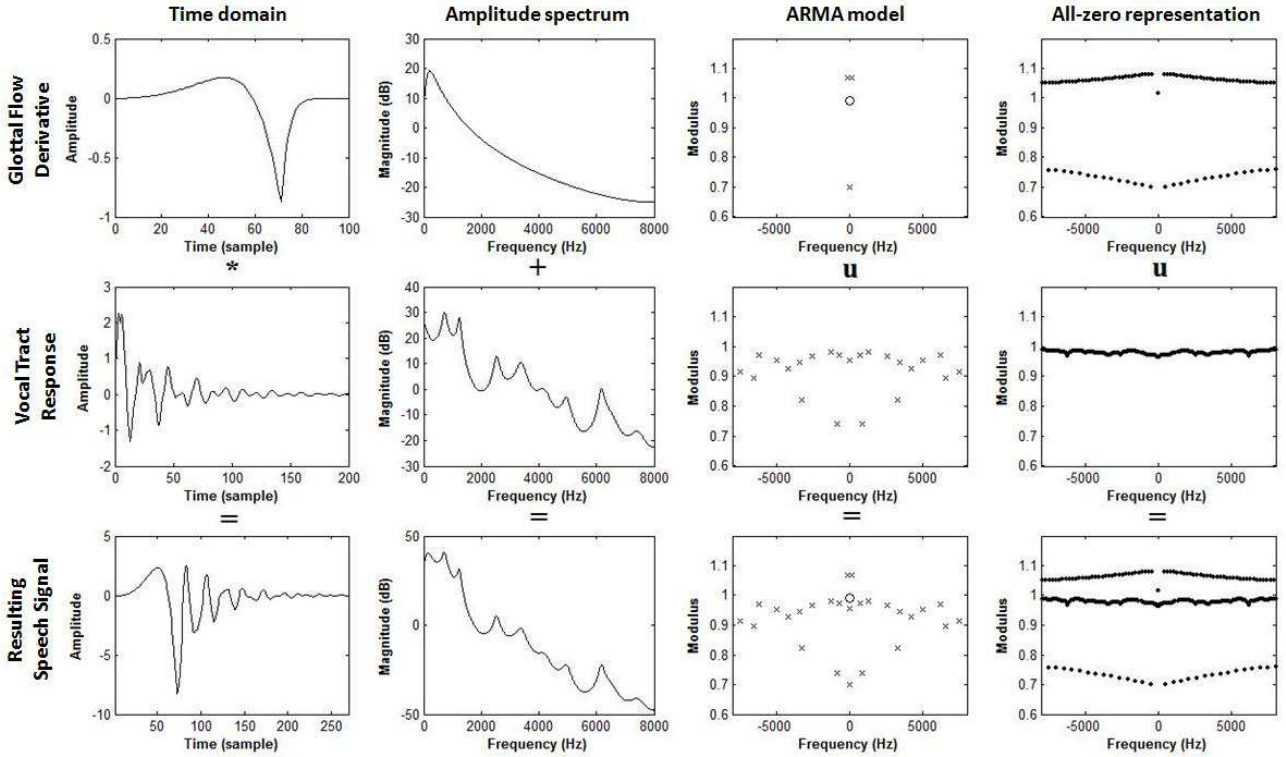

Figure 5.1 - *Illustration of the mixed-phase model. The three rows respectively correspond to the glottal flow derivative, the vocal tract response, and the resulting voiced speech. These three signals are all represented in four domains (from the left to the right): waveform, amplitude spectrum, pole-zero modeling, and all-zero (or ZZT) representation. Each column shows how voiced speech is obtained, in each of the four domains.*

Figure 5.1 illustrates the mixed-phase model on a single frame of synthetic vowel. In each row the glottal flow and vocal tract contributions, as well as the resulting speech signal, are shown in a different representation space. It should be emphasized here that the all-zero representation (later referred to as the Zeros of Z-Transform (ZZT) representation, and shown in the last column) is obtained by a root finding operation (i.e. a finite(n)-length signal frame is represented with only zeros in the z-domain). There exists $n-1$ zeros (of the z-transform) for a signal frame with n samples. However the zero in the third column comes from the AutoRegressive Moving Average (ARMA) model and hence should not be confused with the ZZT. The first row shows a typical glottal flow derivative signal. From the ZZT representation (last column, in polar coordinates), it can be noticed that some zeros lie outside the unit circle while others are located inside it. The outside zeros correspond to the maximum-phase glottal opening, while the others come from the minimum-phase glottal closure [6]. The vocal tract response is displayed in the second row. All its zeros are inside the unit circle due to its damped exponential character. Finally the last row is related to the resulting voiced speech. Interestingly its set of zeros is simply the union of the zeros of the two previous components. This is due to the fact that the convolution operation in the time domain corresponds to the multiplication of the z-transform polynomials in the z-domain. For a detailed study of ZZT representation and the mixed-phase speech model, the reader is referred to [6].

5.2.2 Short-Time Analysis of Voiced Speech

For real speech data, Equation (5.1) is only valid for a short-time signal [12], [13]. Most practical applications therefore require processing of windowed (i.e short-time) speech segments:

$$s(n) = w(n)x(n) \tag{5.7}$$
$$= w(n)(A \cdot p(n) \star g(n) \star v(n) \star r(n)) \tag{5.8}$$

and the goal of the decomposition is to extract the glottal source component $g(n)$ from $s(n)$. As it will be discussed throughout this chapter, windowing is of crucial importance in order to achieve a correct deconvolution. Indeed, the z-transform of $s(n)$ can be written as:

$$S(z) = W(z) \star X(z) \tag{5.9}$$
$$= \sum_{n=0}^{N-1} w(n)x(n)z^{-n} \tag{5.10}$$
$$= s(0)z^{-N+1} \prod_{k=1}^{M_i}(z - Z_{C,k}) \prod_{k=1}^{M_o}(z - Z_{AC,k}) \tag{5.11}$$

where Z_C and Z_{AC} are respectively a set of M_i causal ($|Z_{C,k}| < 1$) and M_o anticausal ($|Z_{AC,k}| > 1$) zeros (with $M_o + M_i = N - 1$). As it will be underlined in Section 5.3.1, Equation (5.11) corresponds to the ZZT representation.

From these expressions, two important considerations have now to be taken into account:

- Since $s(n)$ is finite length, $S(z)$ is a polynomial in z (see Eq. (5.11)). This means that the poles of $H(z)$ are now embedded under an all-zero form. Indeed let us consider a single real pole a. The z-transform of the related impulse response $y(n)$ limited to N poins is [14]:

$$Y(z) = \sum_{n=0}^{N-1} a^n z^{-n} = \frac{1 - (az^{-1})^N}{1 - az^{-1}} \tag{5.12}$$

  which is an all-zero form, since the root of the denominator is also a root of the numerator (and the pole is consequently cancelled).

- It can be seen from Equations (5.9) and (5.10) that the window $w(n)$ may have a dramatic influence on $S(z)$ [13], [8]. As windowing in the time domain results in a convolution of the window spectrum with the speech spectrum, the resulting change in the ZZT is a highly complex issue to study [15]. Indeed the multiplication by the windowing function (as in Equation (5.10)) modifies the root distribution of $X(z)$ in a complex way that cannot be studied analytically. For this reason, the impact of the windowing effects on the mixed-phase model is studied in this chapter in an empirical way, as it was done in [13] and [8] for the convolutional model.

To emphasize the crucial role of windowing, Figures 5.2 and 5.3 respectively display a case of correct and erroneous glottal flow estimation via causal-anticausal decomposition on a real speech segment. In these figures, the top-left panel (a) contains the speech signal together with the applied window and the synchronized differenced ElectroGlottoGraph *dEGG* (after compensation of the delay between the laryngograph and the microphone). Peaks in the dEGG signal are informative about the location of

the Glottal Closure Instant (GCI). The top-right panel (b) plots the roots of the windowed signal ($Z_{C,k}$ and $Z_{AC,k}$) in polar coordinates. The bottom panels (c) and (d) correspond to the time waveform and amplitude spectrum of the maximum-phase (i.e anticausal) component which is expected to correspond to the glottal flow open phase.

In Figure 5.2, an appropriate window respecting the conditions we will derive in Section 5.4 is used. This results in a good separation between the zeros inside and outside the unit circle (see Fig. 5.2(b)). The windowed signal then exhibits good mixed-phase properties and the resulting maximum and minimum-phase components corroborate the model exposed in Section 5.2.1. On the contrary, a 25 ms long Hanning window is employed in Figure 5.3, as widely used in speech processing. It can be seen that even when this window is centered on a GCI, the resulting causal-anticausal decomposition is erroneous. Zeros on each side of the unit circle are not well separated: the windowed signal does not exhibit characteristics of the mixed-phase model. This simple comparison highlights the dramatic influence of windowing on the deconvolution. In Section 5.4, we discuss in detail the set of properties the window should convey so as to yield a good decomposition.

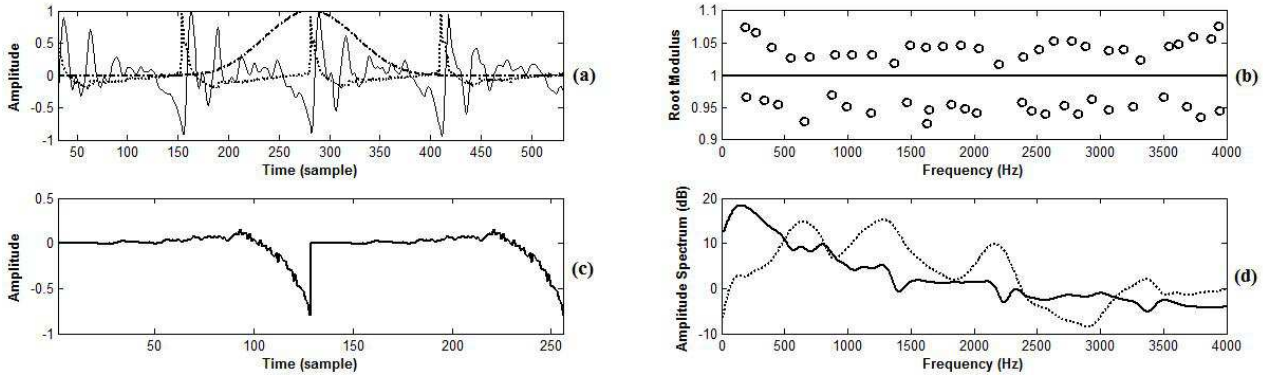

Figure 5.2 - *Example of decomposition on a real speech segment using an appropriate window.* (a): *The speech signal (solid line) with the synchronized dEGG (dotted line) and the applied window (dash-dotted line).* (b): *The zero distribution in polar coordinates.* (c): *Two cycles of the maximum-phase component (corresponding to the glottal flow open phase).* (d): *Amplitude spectra of the minimum (dotted line) and maximum-phase (solid line) components of the speech signal. It can be observed that the windowed signal respects the mixed-phase model since the zeros on each side of the unit circle are well separated.*

5.3 Algorithms for Causal-Anticausal Decomposition of Voiced Speech

For a segment $s(n)$ resulting from an appropriate windowing of a voiced speech signal $x(n)$, two algorithms are compared for achieving causal-anticausal decomposition, thereby leading to an estimate $\tilde{g}(n)$ of the real glottal source $g(n)$. The first one relies on the Zeros of the Z-Transform (ZZT, [6]) and is summarized in Section 5.3.1. The second technique is based on the Complex Cepstrum (CC) and is described in Section 5.3.2. It is important to note that both methods are functionally equivalent to each other, in the sense that they take the same input $s(n)$ and should give the same output $\tilde{g}(n)$. As emphasized in Section 5.2.2, the quality of the decomposition then only depends on the applied windowing, i.e whether $s(n) = w(n)x(n)$ exhibits expected mixed-phase properties or not. It will then be shown that both methods lead to similar results (see Section 5.5.2). However, on a practical

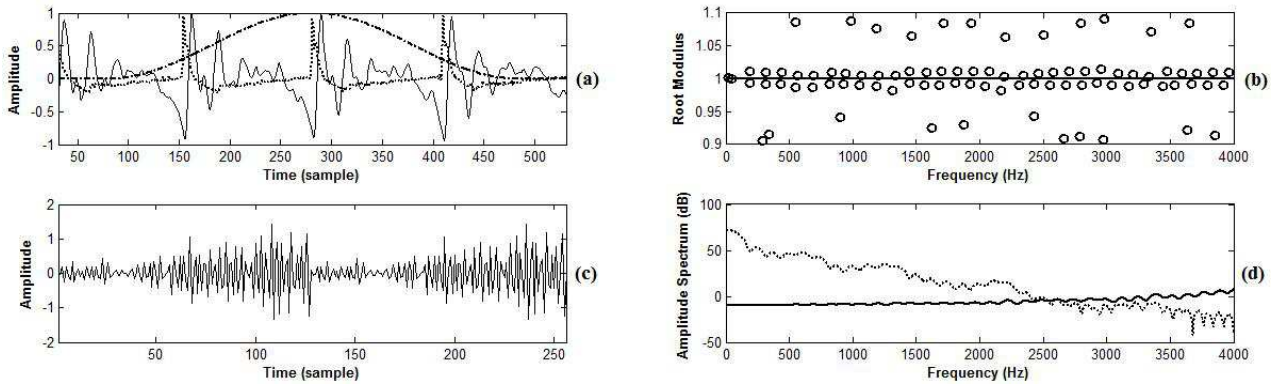

Figure 5.3 - *Example of decomposition on a real speech segment using a 25 ms long Hanning window.* (a): *The speech signal (solid line) with the synchronized dEGG (dotted line) and the applied window (dash-dotted line).* (b): *The zero distribution in polar coordinates.* (c): *Two cycles of the maximum-phase component.* (d): *Amplitude spectra of the minimum (dotted line) and maximum-phase (solid line) components of the speech signal. The zeros on each side of the unit circle are not well separated and the windowed signal does not respect the mixed-phase model. The resulting deconvolved components are irrelevant (while their convolution still gives the input speech signal).*

point of view, the use of the complex cepstrum is advantageous since it will be shown that it is much faster than ZZT. Note that we made a Matlab toolbox containing these two methods freely available in http://tcts.fpms.ac.be/∼drugman/.

5.3.1 Zeros of the Z-Transform-based Decomposition

According to Equation (5.11), $S(z)$ is a polynomial in z with zeros inside and outside the unit circle. The idea of the ZZT-based decomposition is to isolate the roots Z_{AC} and to reconstruct from them the anticausal component. The algorithm can then be summarized as follows [6]:

1. *Window the signal with guidelines provided in Section 5.4,*
2. *Compute the roots of the polynomial $S(z)$,*
3. *Isolate the roots with a modulus greater than 1,*
4. *Compute $\tilde{G}(z)$ from these roots.*

A workflow summarizing the ZZT-based technique is given in Figure 5.4. Although very simple, this technique requires the factorization of a polynomial whose order is generally high (depending on the sampling rate and window length). Even though current factoring algorithms are accurate, the computational load still remains high [16].

In addition to [6] where the ZZT algorithm is introduced, some recent studies [17], [18] have shown that ZZT outperforms other well-known methods of glottal flow estimation in clean recordings. Its main disadvantages are reported as sensitivity to noise and high computational load.

5.3.2 Complex Cepstrum-based Decomposition

Homomorphic systems have been developed in order to separate non-linearly combined signals [7]. As a particular example, the case where inputs are convolved is especially important in speech processing.

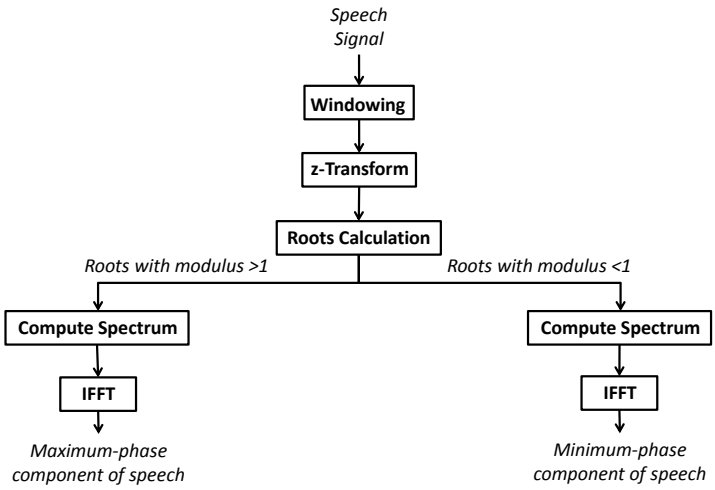

Figure 5.4 - *Block diagram of the ZZT-based decomposition.*

Separation can then be achieved by a linear homomorphic filtering in the complex cepstrum domain, which interestingly presents the property to map time-domain convolution into addition. In speech analysis, complex cepstrum is usually employed to deconvolve the speech signal into a periodic pulse train and the vocal system impulse response [8], [13]. It finds applications such as pitch detection [19], vocoding [20], etc.

We show here how to use the complex cepstrum in order to estimate the glottal flow by achieving the causal-anticausal decomposition introduced in Section 5.2.2. To our knowledge, no complex cepstrum-based glottal flow estimation method is available in the literature. Hence it is one of the novel contributions of this chapter to introduce one and to test it on a large real speech database.

The complex cepstrum (CC) $\hat{s}(n)$ of a discrete signal $s(n)$ is defined by the following equations [7]:

$$S(\omega) = \sum_{n=-\infty}^{\infty} s(n) e^{-j\omega n} \tag{5.13}$$

$$\log[S(\omega)] = \log(|S(\omega)|) + j\angle S(\omega) \tag{5.14}$$

$$\hat{s}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[S(\omega)] e^{j\omega n} d\omega \tag{5.15}$$

where Equations (5.13), (5.14) and (5.15) are respectively the Discrete-Time Fourier Transform (DTFT), the complex logarithm and the inverse DTFT (IDTFT). A workflow summarizing the CCD-based method is presented in Figure 5.5. One difficulty when computing the CC lies in the estimation of $\angle S(\omega)$, which requires an efficient phase unwrapping algorithm. In this work, we computed the Fast Fourier Transform (FFT) on a sufficiently large number of points (typically 4096) such that the grid on the unit circle is sufficiently fine to facilitate in this way the phase evaluation.

If $S(z)$ is written as in Equation (5.11), it can be easily shown [7] that the corresponding complex cepstrum can be expressed as:

$$\hat{s}(n) = \begin{cases} |s(0)| & for\ n = 0 \\ \sum_{k=1}^{M_o} \frac{{Z_{AC,k}}^n}{n} & for\ n < 0 \\ \sum_{k=1}^{M_i} \frac{{Z_{C,k}}^n}{n} & for\ n > 0 \end{cases} \tag{5.16}$$

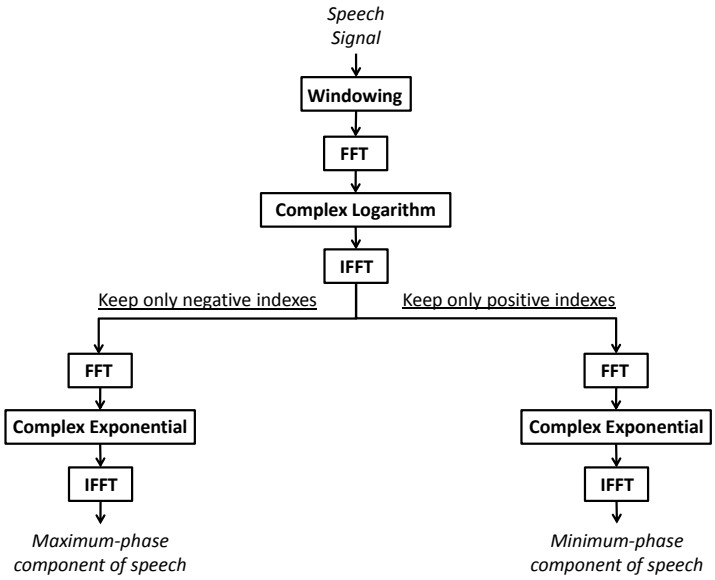

Figure 5.5 - *Block diagram of the Complex Cepstrum-based decomposition.*

This equation shows the close link between ZZT and CC-based techniques. Relying on this equation, Steiglitz and Dickinson demonstrated the possibility of computing the complex cepstrum and unwrapped phase by factoring the z-transform [21], [22]. The approach we propose is just the inverse thought process in the sense that our goal is precisely to use the complex cepstrum in order to avoid any factorization. In this way we show that the complex cepstrum can be used as an efficient means to estimate the glottal flow, while circumventing the requirement of factoring polynomials (as it is the case for the ZZT). Indeed it will be shown in Section 5.4.2 that optimal windows have their length proportional to the pitch period. The ZZT-based technique then requires to compute the roots of generally high-order polynomials (depending on the sampling rate and on the pitch). Although current polynomial factoring algorithms are accurate, the computational load still remains high, with a complexity order of $O(n^2)$ for the fastest algorithms [16], where n denotes the number of samples in the considered frame. On the other hand, the CC-based method just relies on FFT and IFFT operations which can be fast computed, and whose order is $O(N_{FFT}\log(N_{FFT}))$, where N_{FFT} is fixed to 4096 in this work for facilitating phase unwrapping, as mentioned above. For this reason a change in the frame length has little influence on the computation time for the CC-based method. Table 5.1 compares both methods in terms of computation time. The use of the complex cepstrum now offers the possibility of integrating a causal-anticausal decomposition module into a real-time application, which was previously almost impossible with the ZZT-based technique.

Pitch	ZZT-based decomposition	CC-based decompostion
60 Hz	111.4	1.038
180 Hz	11.2	1

Table 5.1 - *Comparison of the relative computation time (for our Matlab implementation with $F_s = 16kHz$) required for decomposing a two pitch period long speech frame. Durations were normalized according to the time needed by the complex cepstrum-based deconvolution for $F_0 = 180Hz$.*

Regarding Equation (5.16), it is obvious that causal-anticausal decomposition can be performed

using the complex cepstrum, as follows [23]:

1. *Window the signal with guidelines provided in Section 5.4,*
2. *Compute the complex cepstrum $\hat{s}(n)$ using Equations (5.13), (5.14) and (5.15),*
3. *Set $\hat{s}(n)$ to zero for $n > 0$,*
4. *Compute $\tilde{g}(n)$ by applying the inverse operations of Equations (5.13), (5.14) and (5.15) on the resulting complex cepstrum.*

Figure 5.6 illustrates the complex cepstrum-based decomposition for the example shown in Figure 5.2. A simple linear liftering keeping only the negative (positive) indexes of the complex cepstrum allows to isolate the maximum and minimum phase components of voiced speech. It should be emphasized that windowing is still very critical, as it is the case for the ZZT decomposition. The example in Figure 5.3 (where a $25ms$ long Hanning window is used) would lead to an unsuccessful decomposition. We think that this critical dependence on the window function, length and location was the main hindrance in developing a complex cepstrum-based glottal flow estimation method, although its potential was known earlier in the literature [8].

It is also worth noting that since the CC method is an alternative means of achieving the mixed-phase decomposition, it suffers from the same noise sensitivity as the ZZT does.

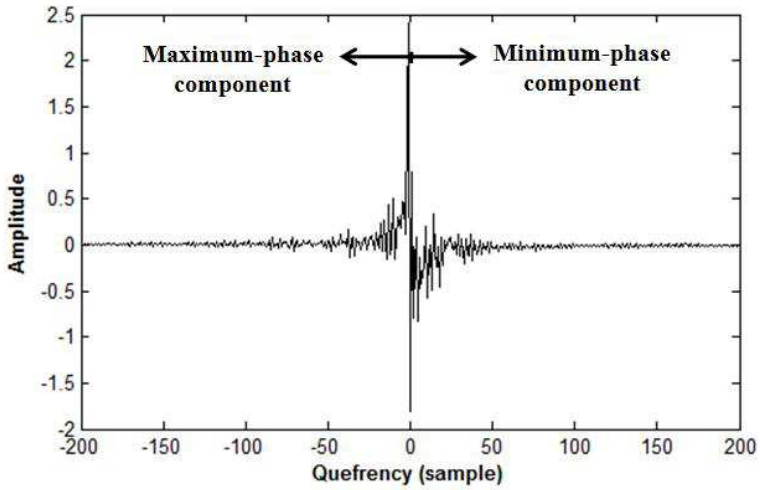

Figure 5.6 - *The complex cepstrum $\hat{s}(n)$ of the windowed speech segment $s(n)$ presented in Figure 5.2(a). The maximum- (minimum-) phase component can be isolated by only considering the negative (positive) indexes of the complex cepstrum.*

5.4 Experiments on Synthetic Speech

The goal of this Section is to study, on synthetic speech signals, the impact of the windowing effects on the causal-anticausal decomposition. It is one of the main contributions of this study to provide a parametric analysis of the windowing problem and provide guidelines for reliable complex cepstrum-based glottal flow estimation. The experimental protocol we opted for is close to the one presented in [17]. Synthetic speech signals (sampled at 16 kHz) are generated for a wide range of test conditions [23]. The idea is to cover the diversity of configurations one could find in natural speech by varying all

parameters over their whole range. Synthetic speech is produced according to the source-filter model by passing a synthetic train of Liljencrants-Fant (LF) glottal waves [24] through an auto-regressive filter extracted by LPC analysis (with an order of 18) of real sustained vowels uttered by a male speaker. As the mean pitch in these utterances is about 100 Hz, it is reasonable to consider that the fundamental frequency should not exceed 60 and 180 Hz in continuous speech. Experiments in this section can then be seen as a proof of concept on synthetic male speech. Table 5.2 summarizes all test conditions.

Pitch	60:20:180 Hz
Open quotient	0.4:0.05:0.9
Asymmetry coefficient	0.6:0.05:0.9
Vowel	/a/, /@/, /i/, /y/

Table 5.2 - *Table of synthesis parameter variation range.*

Decomposition quality is assessed through two objective measures [23]:

- **Spectral distortion** : Many frequency-domain measures for quantifying the distance between two speech frames have been proposed in the speech coding litterature [25]. Ideally the subjective ear sensitivity should be formalised by incorporating psychoacoustic effects such as masking or isosonic curves. A simple relevant measure between the estimated $\hat{g}(n)$ and the real glottal pulse $g(n)$ is the spectral distortion (SD) defined as [25]:

$$SD(g,\hat{g}) = \sqrt{\int_{-\pi}^{\pi} (20\log_{10}|\frac{G(\omega)}{\hat{G}(\omega)}|)^2 \frac{d\omega}{2\pi}} \tag{5.17}$$

  where $G(\omega)$ and $\hat{G}(\omega)$ denote the DTFT of the original target glottal pulse $g(n)$ and of the estimate $\hat{g}(n)$. To give an idea, it is argued in [26] that a difference of about 1dB (with a sampling rate of 8kHz) is rather imperceptible.

- **Glottal formant determination rate** : The amplitude spectrum for a voiced source generally presents a resonance called the *glottal formant* ([27], see also Section 5.2.1). As this parameter is an essential feature of the glottal open phase, an error on its determination after decomposition should be penalized. For this, we define the *glottal formant determination rate* as the proportion of frames for which the relative error on the glottal formant frequency is lower than 10%.

This formal experimental protocol allows us to reliably assess our technique and to test its sensivity to various factors influencing the decomposition, such as the window location, function and length. Indeed, Tribolet *et al.* already observed in 1977 that the window shape and onset may lead to zeros whose topology can be detrimental for accurate pulse estimation [12]. The goal of this empirical study on synthetic signals is precisely to handle these zeros close to the unit circle, so that the applied window leads to a correct causal-anticausal separation.

5.4.1 Influence of the window location

In [8] the need of aligning the center of the window with the system response is highlighted. Analysis is then performed on windows centered on GCIs, as these particular events demarcate the boundary between the causal and anticausal responses, and the linear phase contribution is removed. Figure 5.7 illustrates the sensitivity of the causal-anticausal decomposition to the window position. It can

be noticed that the performance rapidly degrades, especially if the window is centered on the left of the GCI. It is then recommended to apply a GCI-centered windowing. For this, the performance of methods for automatically detecting GCIs from the speech waveform has been studied in Chapter 3. The impact of the accuracy of these algorithms on the mixed-phase decomposition has even been investigated in Section 3.5.2. For cases for which GCI information is not available or unreliable, Chapter 7 will extend the formalism of the mixed-phase separation to a chirp analysis, allowing the deconvolution to be achieved in an asynchronous way.

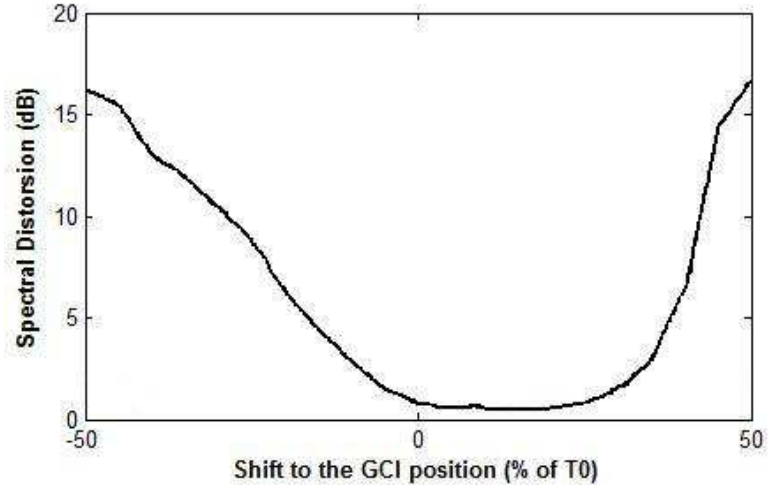

Figure 5.7 - *Sensitivity of the causal-anticausal decomposition to a GCI location error. The spectral distortion dramatically increases if a non GCI-centered windowing is applied (particularly on the left of the GCI).*

5.4.2 Influence of the window shape and length

In Section 5.2.2, Figures 5.2 and 5.3 showed an example of correct and erroneous decomposition respectively. The only difference between these figures was the length and shape of the applied windowing. To study this effect let us consider a particular family of windows $w(n)$ of N points satisfying the form [7]:

$$w(n) = \frac{\alpha}{2} - \frac{1}{2}\cos(\frac{2\pi n}{N-1}) + \frac{1-\alpha}{2}\cos(\frac{4\pi n}{N-1}) \tag{5.18}$$

where α is a parameter comprised between 0.7 and 1 (for α below 0.7, the window includes negative values which should be avoided). The widely used Hanning and Blackman windows are particular cases of this family for $\alpha = 1$ and $\alpha = 0.84$ respectively. Figure 5.8 displays the evolution of the decomposition quality when α and the window length vary. It turns out that a good deconvolution can be achieved as long as the window length is adapted to its shape (or vice versa). For example, the optimal length is about 1.5 T_0 for a Hanning window and 1.75 T_0 for a Blackman window. A similar observation can be drawn from Figure 5.9 according to the spectral distortion criterion. Note that we displayed the inverse spectral distortion $1/SD$ instead of SD only for better viewing purposes. At this point it is interesting to notice that these constraints on the window aiming at respecting the mixed-phase model are sensibly different from those imposed to respect the so-called *convolutional model* [13], [8]. For this latter case, it was indeed recommended to use windows such as Hanning or Hamming with a duration of about 2 to 3 pitch periods. It can be seen from Figure 5.8 that this

would lead to poor causal-anticausal decomposition results. Finally note that it was proposed in [28] to analytically derive the optimal frame length for the causal-anticausal decomposition, by satisfying an immiscibility criterion based on a Cauchy bound.

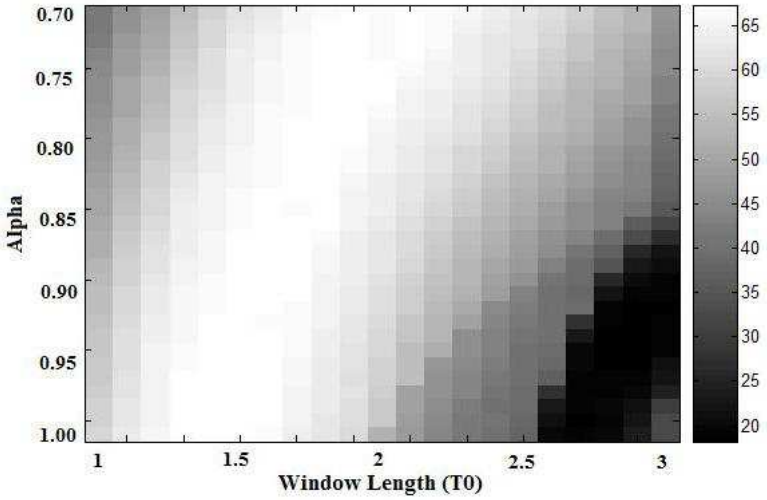

Figure 5.8 - *Evolution of the glottal formant determination rate according the window length and shape. Note that the Hanning and Blackman windows respectively correspond to* $\alpha = 1$ *and* $\alpha = 0.84$.

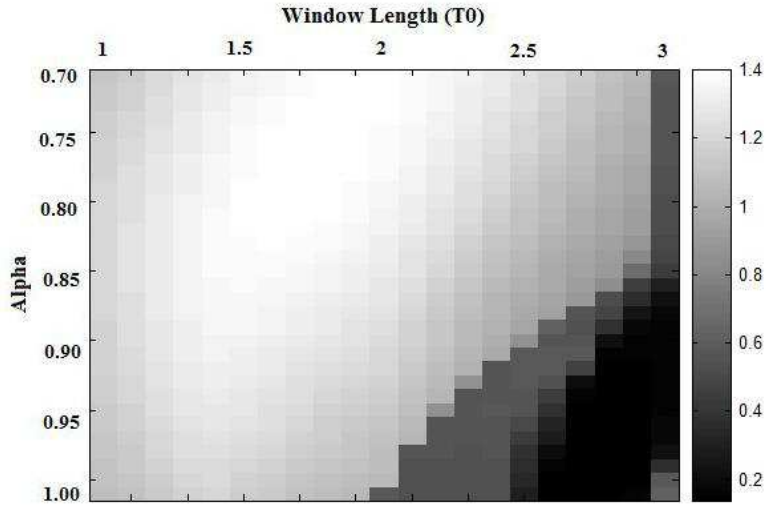

Figure 5.9 - *Evolution of the inverse spectral distortion* $1/SD$ *according the window length and shape. Note that the Hanning and Blackman windows respectively correspond to* $\alpha = 1$ *and* $\alpha = 0.84$. *The inverse SD is plotted instead of the SD itself only for clarity purpose.*

5.5 Experiments on Real Speech

The goal of this section is to show that a reliable glottal flow estimation is possible on real speech using the complex cesptrum. The efficiency of this method will be confirmed in Sections 5.5.1 and 5.5.2 by

analyzing short segments of real speech. Besides we demonstrate in Section 5.5.3 the potential of using complex cepstrum for voice quality analysis on a large expressive speech corpus.

For these experiments, speech signals sampled at $16kHz$ are considered. The pitch contours are extracted using the Snack library [29] and the GCIs are located directly from the speech waveforms using the SEDREAMS algorithm proposed in Section 3.3 (or [30]). Speech frames are then obtained by applying a GCI-centered windowing. The window we use satisfies Equation (5.18) for $\alpha = 0.7$ and is two pitch period-long so as to respect the conditions derived in Section 5.4. Causal-anticausal decomposition is then achieved by the complex cepstrum-based method.

5.5.1 Example of Decomposition

Figure 5.10 illustrates a concrete case of decomposition on a voiced speech segment (diphone */am/*) uttered by a female speaker. The top plot displays the speech signal together with its corresponding estimated glottal flow derivative (bottom plot). It can be seen that even on a nasalized phoneme the glottal source estimation seems to be correctly carried out for most speech frames (i.e the obtained waveforms turn out to corroborate the model of the glottal pulse described in Section 5.2.1). For some rare cases the causal-anticausal decomposition is erroneous and the maximum-phase component contains a high-frequency irrelevant noise. Nevertheless the spectrum of this maximum-phase contribution almost always presents a low-frequency resonance due to the glottal formant. As an illustration, Figure 5.11 shows how the spectrum is altered when estimates include a high-frequency irrelevant noise. It is observed that even for these cases, the low-frequency contents is maintained and contains information about the glottal formant.

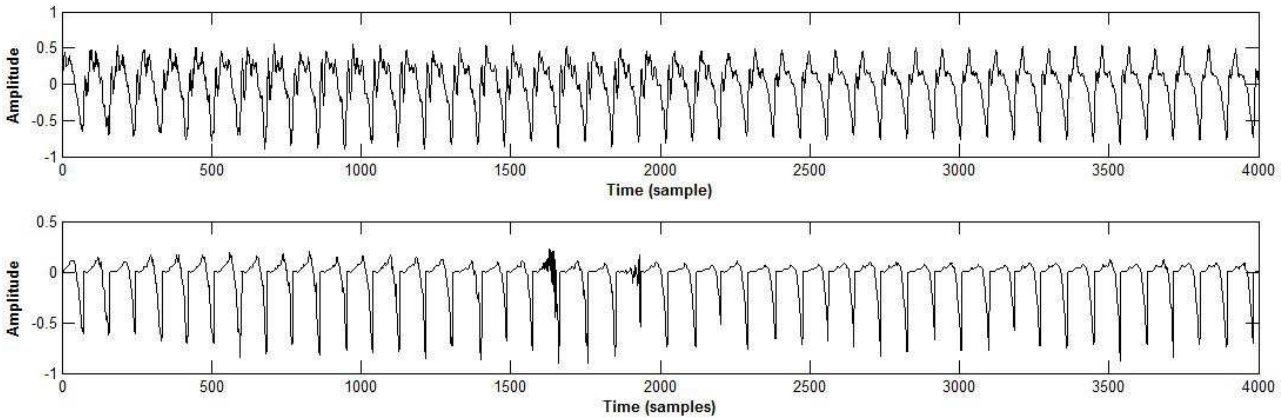

Figure 5.10 - Top panel: *A segment of voiced speech (diphone /*am*/) uttered by a female speaker.* Bottom panel: *Its corresponding glottal source estimation obtained using the complex cepstrum-based decomposition. It turns out that a plausible estimation can be achieved for most of the speech frames.*

5.5.2 Analysis of sustained vowels

In this experiment, we consider a sustained vowel */a/* with a flat pitch which was voluntarily produced with an increasing pressed vocal effort [1]. Here the aim is to show that voice quality variation is reflected as expected on the glottal flow estimates obtained using the causal-anticausal decomposition. Figure 5.12 plots the evolution of the glottal formant frequency Fg and bandwidth Bw during the phonation

[1] Many thanks to N. Henrich and B. Doval for providing the recording of the sustained vowel.

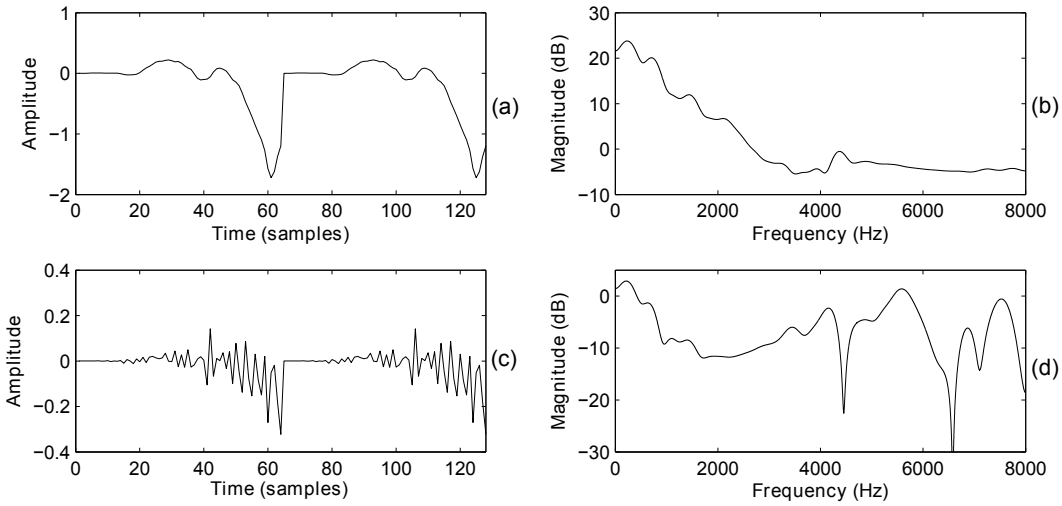

Figure 5.11 - *Examples of spectrum obtained with the complex cepstrum-based decomposition. (a): Two cycles of the glottal flow derivative for a plausible estimate, (b): the corresponding magnitude spectrum, (c): Two cycles of the glottal flow derivative when the estimate contains an irrelevant high-frequency noise, (d): the corresponding magnitude spectrum. It is observed that even for the worst case (second row), the low-frequency contents is not altered and contains information about the glottal formant.*

[23]. These features were estimated with both ZZT and CC-based methods. It can be observed that these techniques lead to similar results. The very slight differences may be due to the fact that, for the complex cepstrum, Equation (5.16) is realized on a finite number n of points. Another possible explanation is the precision problem in root computation for the ZZT-based technique. In any case, it can be noticed that the increasing vocal effort can be characterized by increasing values of Fg and Bw.

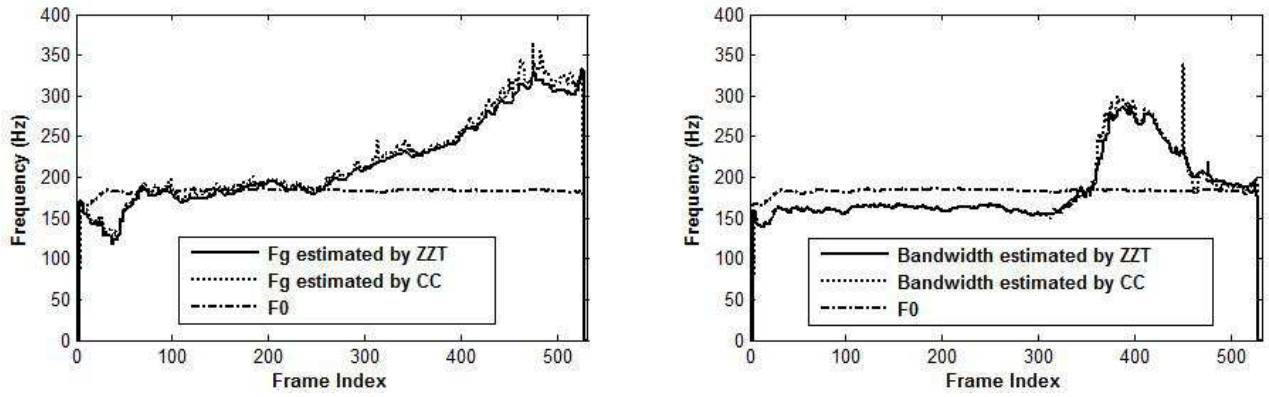

Figure 5.12 - *Glottal formant characteristics estimated by both ZZT and CC-based techniques on a real sustained vowel with an increasing pressed effort [23].* Left panel: *Evolution of the glottal formant frequency.* Right panel: *Evolution of the glottal formant 3dB bandwidth.*

5.5.3 Analysis of an Expressive Speech Corpus

The goal of this section is to show that the differences in the glottal source when a speaker produces various voice qualities can be tracked using causal-anticausal decomposition. For this, the De7 database is used [2]. This database was designed by Marc Schroeder as one of the first attempts of creating diphone databases for expressive speech synthesis [31]. The database contains three voice qualities (modal, soft and loud) uttered by a German female speaker, with about 50 minutes of speech available for each voice quality.

For each voiced speech frame, the complex cepstrum-based decomposition is performed. The resulting maximum-phase component is then downsampled at 8kHz and is assumed to give an estimate of the glottal flow derivative for the considered frame. For each segment of voiced speech, a signal similar to the one illustrated in Figure 5.10 is consequently obtained. In this figure, it was observed that an erroneous decomposition might appear for some frames, leading to an irrelevant high-frequency noise in the estimated anticausal contribution (also observed in Figure 5.3). One first thing one could wonder is how large is the proportion of such frames over the whole database.

As a criterion deciding whether a frame is considered as correctly decomposed or not, we inspect the spectral center of gravity. The distribution of this feature is displayed in Figure 5.13 for the loud voice. A principal mode at around 2kHz clearly emerges and corresponds to the majority of frames for which a correct decomposition is carried out. A second minor mode at higher frequencies is also observed. It is related to the frames where the causal-anticausal decomposition fails, leading to a maximum-phase signal containing an irrelevant high-frequency noise (as explained above). It can be noticed from this histogram (and it was confirmed by a manual verification of numerous frames) that fixing a threshold at around 2.7 kHz makes a good distinction between frames that are correctly and incorrectly decomposed.

According to this criterion, Table 5.3 summarizes for the whole database the percentage of frames leading to a correct estimation of the glottal flow. It turns out that a high proportion of frames (around 85% for each dataset) are correctly decomposed. For the remaining frames, the windowed signal does not match the mixed-phase model. This might be explained by a non-suited windowing, or by the fact that during the production of these particular sounds the mixed-phase model does not hold. Indeed, it is important to highlight that applying an appropriate windowing is a necessary but not sufficient condition for achieving a correct deconvolution.

Voice Quality	% of frames correctly decomposed
Loud	87.22%
Modal	84.41%
Soft	83.69%

Table 5.3 - *Proportion of frames leading to a correct causal-anticausal decomposition for the three voice qualities.*

For each frame correctly deconvolved, the glottal flow is then characterized by the 3 following common features described in Section 4.3:

- the Normalized Amplitude Quotient(NAQ) characterizing the glottal closing phase [32],
- the $H1 - H2$ ratio widely used as a measure characterizing voice quality [33], [34], [35],

[2] Many thanks to M. Schroeder for providing the De7 database.

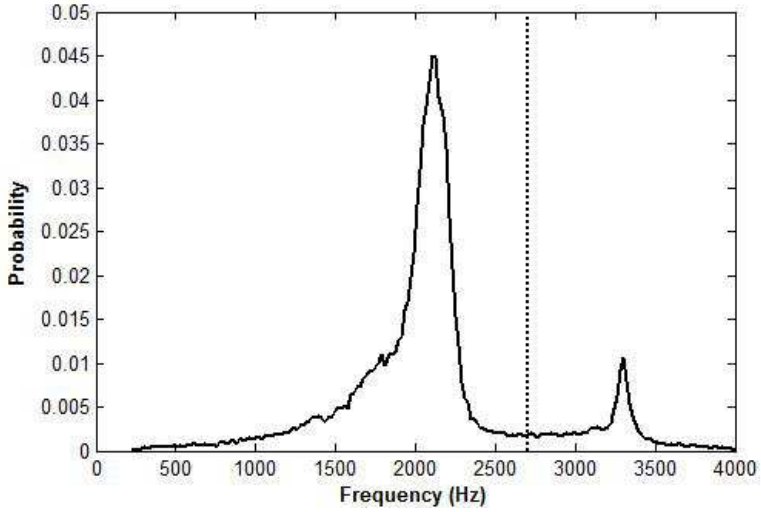

Figure 5.13 - *Distribution of the spectral center of gravity of the maximum-phase component, computed for the whole dataset of loud samples. Fixing a threshold around 2.7kHz makes a good separation between correctly and incorrectly decomposed frames.*

- and the Harmonic Richness Factor (HRF) quantifying the amount of harmonics in the magnitude spectrum of the glottal source and shown to be informative about the phonation type in [36] and [35].

Figure 5.14 shows the histograms of these 3 parameters for the three voice qualities. Significant differences between the distributions are observed. Among others it turns out that the production of a louder (softer) voice results in lower (higher) NAQ and $H1-H2$ values, and of a higher (lower) Harmonic Richness Factor (HRF). These conclusions corroborate the results recently obtained on sustained vowels by Alku in [35] and [32]. Another observation that can be drawn from the histogram of $H1-H2$ is the presence of two modes for the modal and loud voices. This may be explained by the fact that the estimated glottal source sometimes comprises a ripple both in the time and frequency domains [37]. Indeed consider Figure 5.15 where two typical cycles of the glottal source are presented for both the soft and loud voice. Two conclusions can be drawn from it. First of all, it is clearly seen that the glottal open phase response for the soft voice is slower than for the loud voice. As it was underlined in the experiment of Section 5.5.2, this confirms the fact Fg/F_0 increases with the vocal effort. Secondly the presence of a ripple in the loud glottal waveform is highlighted. This has two possible origins: an incomplete separation between Fg and the first formant F_1 [38], and/or a non-linear interaction between the vocal tract and the glottis [37], [39]. This ripple affects the low-frequency contents of the glottal source spectrum, and may consequently perturb the estimation of the $H1-H2$ feature. This may therefore explain the second mode in the $H1-H2$ histogram for the modal and loud voices (where ripple was observed).

5.6 Conclusion

This chapter explained the causal-anticausal decomposition principles in order to estimate the glottal source directly from the speech waveform. We showed that the complex cepstrum can be effectively used for this purpose as an alternative to the Zeros of the Z-Transform (ZZT) algorithm. Both techniques were shown to be functionally equivalent to each other, while the complex cepstrum is advantageous for

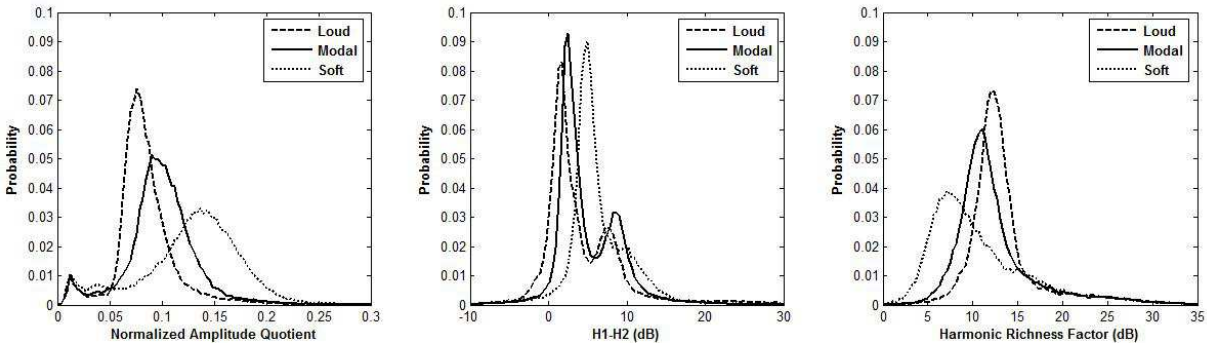

Figure 5.14 - *Distributions, computed on a large expressive speech corpus, of glottal source parameters for three voice qualities:* (left) *the Normalized Amplitude Quotient (NAQ),* (middle) *the $H1-H2$ ratio, and* (right) *the Harmonic Richness Factor (HRF).*

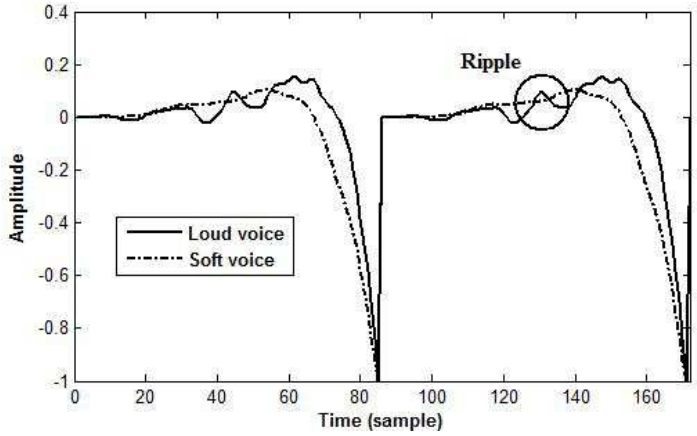

Figure 5.15 - *Comparison between two cycles of typical glottal source for both soft (dash-dotted line) and loud voice (solid line). The presence of a ripple in the loud excitation can be observed.*

its much higher speed, making it suitable for real-time applications. Windowing effects were studied in a systematic way on synthetic signals. It was emphasized that windowing plays a crucial role. More particularly we derived a set of constraints the window should respect so that the windowed signal matches the mixed-phase model. Finally, results on a real speech database (logatoms recorded for the design of an unlimited domain expressive speech synthesizer) were presented for voice quality analysis. The glottal flow was estimated on a large database containing various voice qualities. Interestingly some significant differences between the voice qualities were observed in the excitation. The methods proposed in this chapter may be used in several potential applications of speech processing such as emotion detection, speaker recognition, expressive speech synthesis, automatic voice pathology detection and various other applications where real-time glottal source estimation may be useful. Finally note that a Matlab toolbox containing these algorithms is freely available from http://tcts.fpms.ac.be/∼drugman/.

Bibliography

[1] D. Veeneman and S. Bement. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 33(2):369–377, 1985.

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

[3] P. Alku, J. Svec, E. Vilkman, and F. Sram. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[4] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

[5] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *ISCA ITRW VOQUAL03*, pages 15–19, 2003.

[6] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12, 2005.

[7] A. Oppenheim and R. Schafer. *Discrete-time signal processing.* Prentice-Hall, 1989.

[8] T. Quatieri. *Discrete-time speech signal processing.* Prentice-Hall, 2002.

[9] D. Childers. *Speech Processing and Synthesis Toolboxes.* Wiley and Sons, Inc., 1999.

[10] W. Gardner and B. Rao. Noncausal all-pole modeling of voiced speech. *IEEE Trans. Speech and Audio Processing*, 5(1):1–10, 1997.

[11] H. Deng, R. Ward, M. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Trans. ASSP*, 14:445–455, 2006.

[12] J. Tribolet, T. Quatieri, and A. Oppenheim. Short-time homomorphic analysis. In *Proc. ICASSP*, volume 2, pages 716–72, 1977.

[13] W. Verhelst and O. Steenhaut. A new model for the short-time complex cepstrum of voiced speech. *IEEE Trans. ASSP*, 34:43–51, 1986.

[14] A. Oppenheim, A. Willsky, and I. Young. Signals and systems. *Prentice Hall International Editions*, 1983.

[15] B. Bozkurt, L. Couvreur, and T. Dutoit. Chirp group delay analysis of speech signals. *Speech Comm.*, 49:159–176, 2007.

[16] G. Sitton, C. Burrus, J. Fox, and S. Treitel. Factoring very-high degree polynomials. *IEEE Signal Processing Magazine*, pages 27–42, 2003.

[17] N. Sturmel, C. d'Alessandro, and B. Doval. A comparative evaluation of the zeros of z transform representation for voice source estimation. In *Proc. Interspeech*, pages 558–561, 2007.

[18] C. D'Alessandro, B. Bozkurt, B. Doval, T. Dutoit, N. Henrich, V. Tuan, and N. Sturmel. Phase-based methods for voice source analysis. *Advances in Nonlinear Speech Processing, LNCS 4885*, pages 1–27, 2008.

[19] J. Wangrae, K. Jongkuk, and B. Myung Jin. A study on pitch detection in time-frequency hybrid domain. *Lecture Notes in Computer Science, Springer Berlin*, pages 337–340, 2005.

[20] T. Quatieri. Minimum- and mixed-phase speech analysis/synthesis by adaptive homomorphic deconvolution. *IEEE Trans. ASSP*, 27(4):328–335, 1979.

[21] K. Steiglitz and B. Dickinson. Computation of the complex cepstrum by factorization of the z-transform. In *Proc. ICASSP*, volume 2, pages 723–726, 1977.

[22] K. Steiglitz and B. Dickinson. Phase unwrapping by factorization. *IEEE Trans. ASSP*, 30(6): 984–991, 1982.

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

[24] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[25] F. Nordin and T. Eriksson. A speech spectrum distortion measure with interframe memory. In *Proc. ICASSP*, volume 2, pages 717–720, 2001.

[26] B. Atal K. Paliwal. Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Trans. Speech Audio Processing*, 1:3–14, 1993.

[27] B. Doval and C. d'Alessandro. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006.

[28] C. Pedersen, O. Andersen, and P. Dalsgaard. Separation of mixed-phase signals by zeros of the z-transform - a reformulation of complex cepstrum-based separation by causality. In *Proc. ICASSP*, 2010.

[29] Online. The snack sound toolkit. In *http://www.speech.kth.se/snack/*.

[30] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[31] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[32] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112:701–710, 2002.

[33] H. Hanson. Individual variations in glottal characteristics of female speakers. In *Proc. ICASSP*, pages 772–775, 1995.

[34] G. Fant. The lf-model revisited. transformations and frequency domain analysis. *STL-QPSR*, 36 (2-3):119–156, 1995.

[35] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[36] D. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:2394–2410, 1991.

[37] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–586, 1999.

[38] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. A method for glottal formant frequency estimation. In *Proc. Interspeech*, 2004.

[39] T. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Comm.*, pages 167–184, 1982.

Chapter 6

A Comparative Study of Glottal Source Estimation Techniques

Contents

Abstract

Source-tract decomposition (or glottal flow estimation) is one of the basic problems of speech processing. For this, several techniques have been proposed in the literature. However studies comparing different approaches are almost nonexistent. Besides, experiments have been systematically performed either on synthetic speech or on sustained vowels. In this study we compare three of the main representative state-of-the-art methods of glottal flow estimation: closed-phase inverse filtering, iterative and adaptive inverse filtering, and mixed-phase decomposition. These techniques are first submitted to an objective assessment test on synthetic speech signals. Their sensitivity to various factors affecting the estimation quality, as well as their robustness to noise are studied. In a second experiment, their ability to label voice quality (tensed, modal, soft) is studied on a large corpus of real connected speech. It is shown that changes of voice quality are reflected by significant modifications in glottal feature distributions. Techniques based on the mixed-phase decomposition and on a closed-phase inverse filtering process turn out to give the best results on both clean synthetic and real speech signals. On the other hand, iterative and adaptive inverse filtering is recommended in noisy environments for its high robustness.

This chapter is based upon the following publication:

- Thomas Drugman, Baris Bozkurt, Thierry Dutoit, *A Comparative Study of Glottal Source Estimation Techniques*, Computer, Speech and Language, Elsevier, *Accepted for publication.*

Many thanks to Dr. Baris Bozkurt (Izmir Institute of Technology) for his helpful guidance.

6.1 Introduction

Chapter 4 briefly introduced the issue of source-tract (or source-filter) deconvolution (i.e the separation of the glottal source and the vocal tract contributions) directly from the speech signal. As one of the basic problems and challenges of speech processing research, glottal flow estimation has been studied by many researchers and various techniques are available in the literature [1]. However the diversity of algorithms and the fact that the reference for the actual glottal flow is not available often leads to the questionability about relative effectiveness of the methods in real life applications. In most of studies, tests are performed either on synthetic speech or on a few recorded sustained vowels. In addition, very few comparative studies exist (such as [2]). In this chapter, we compare three of the main representative state-of-the-art methods: closed-phase inverse filtering, iterative and adaptive inverse filtering, and mixed-phase decomposition. For testing, we first follow the common approach of using a large set of synthetic speech signals (by varying synthesis parameters independently), and then we examine how these techniques perform on a large real speech corpus. In the synthetic speech tests, the original glottal flow is available, so that objective measures of decomposition quality can be computed. In real speech tests the ability of the methods to discriminate different voice qualities (tensed, modal and soft) is studied on a large database (without limiting data to sustained vowels).

This chapter is structured as follows. In Section 6.2, the three techniques that are compared in this study are detailed. These methods are representatives of the three appproaches for glottal source estimation introduced in Section 4.2. They are evaluated in Section 6.3 through a wide systematic study on synthetic signals. Their robustness to noise, as well as the impact of the various factors that may affect source-tract separation, are investigated. Section 6.4 presents decomposition results on a real speech database containing various voice qualities, and shows that the glottal source estimated by the techniques considered in this work conveys relevant information about the phonation type. Finally Section 6.5 draws the conclusions of this study.

6.2 Methods Compared in this Chapter

This section describes the three methods of glottal flow estimation that are evaluated and compared in this chapter. These techniques were chosen as they are representatives of the three main state-of-the-art categories of glottal source estimation methods (as presented in Section 4.2), and as they rely on completely different perspectives.

6.2.1 Closed Phase Inverse Filtering

The Closed Phase Inverse Filtering (CPIF) technique aims at estimating a parametric model of the spectral envelope, computed during the estimated closed phase duration, as the effects of the subglottal cavities are minimized during this period, providing a better way for estimating the vocal tract transfer function. In this study, the CPIF technique that is used is based on a Discrete All Pole (DAP, [3]) inverse filtering process estimated during the closed phase. In order to provide a better fitting of spectral envelopes from discrete spectra [3], the DAP technique aims at computing the parameters of an autoregressive model by minimizing a discrete version of the Itakura-Saito distance [4], instead of the time squared error used by the traditional LPC. The use of the Itakura-Saito distance is justified as it is a spectral distortion measure arising from the human hearing perception. The closed phase period is determined using the Glottal Opening and Closure Instants (GOIs and GCIs) located by the SEDREAMS algorithm detailed in Section 3.3 (or [5]). This algorithm has been shown to be effective for reliably and accurately determining the position of these events on a large corpus containing several

speakers. For tests with synthetic speech, the exact closed phase period is known and is used for CPIF. Note that for high-pitched voices, two analysis windows were used as suggested in [6], [7] and [8]. In the rest of the chapter, speech signals sampled at 16 kHz are considered, and the order for DAP analysis is fixed to 18 ($=F_s/1000+2$, as commonly used in the literature). Through our experiments, we reported that the choice of the DAP order is not critical in the usual range, and that working with an order comprised between 12 and 18 leads to sensibly similar results. A workflow summarizing the CPIF technique is given in Figure 6.1.

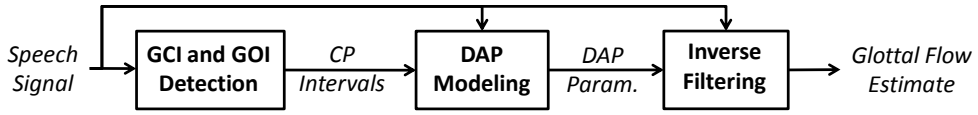

Figure 6.1 - *Block diagram of the Closed Phase Inverse Filtering (CPIF) method for glottal flow estimation.*

6.2.2 Iterative Adaptive Inverse Filtering

The Iterative Adaptive Inverse Filtering (IAIF) method is a popular approach proposed by Alku in [9] for improving the quality of the glottal flow estimation. It is based on an iterative refinement of both the vocal tract and the glottal components. In [10], the same authors proposed an improvement, in which the LPC analysis is replaced by the Discrete All Pole (DAP) modeling technique [3], shown to be more accurate for high-pitched voices. In this study, we used the implementation of the IAIF method [11] from the toolbox available on the TKK Aparat website [12], with its default options. A workflow summarizing the IAIF method is presented in Figure 6.2.

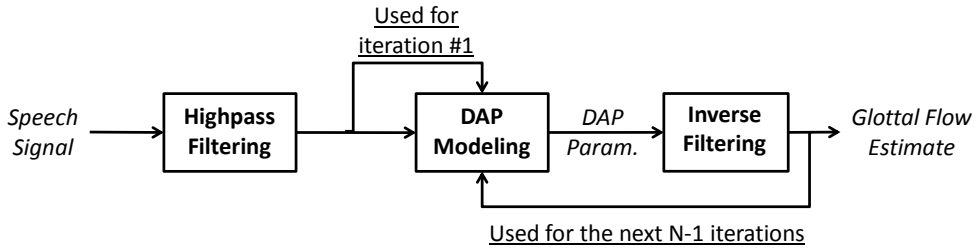

Figure 6.2 - *Block diagram of the Iterative Adaptive Inverse Filtering (IAIF) method for glottal flow estimation.*

6.2.3 Complex Cepstrum-based Decomposition

As highlighted in Chapter 5, the Complex Cesptrum-based Decomposition (CCD) and the Zeros of the Z-Transform (ZZT) techniques are two functionally equivalent algorithms for achieving the mixed-phase separation. In the rest of this chapter, CCD is considered for its higher computational speed. To guarantee good mixed-phase properties [13], GCI-centered two pitch period-long Blackman windows are used. For this, GCIs were located on real speech using the SEDREAMS technique described in Section 3.3 (or [5]). CC is calculated as explained in Section 5.3.2 and the Fast Fourier Transform (FFT) is computed on a sufficiently large number of points (typically 4096), which facilitates phase unwrapping.

6.3 Experiments on Synthetic Speech

The experimental protocol on synthetic speech signals is similar to the one presented in Section 5.4. However we here focus on a larger panel of sustained vowels which were uttered by a female speaker. As the mean pitch during these utterances was about 180 Hz, it can be considered that fundamental frequency should not exceed 100 and 240 Hz in continuous speech. For the LF parameters, the Open Quotient Oq and Asymmetry coefficient α_m are varied through their common range (see Table 6.1). For the filter, 14 types of typical vowels are considered. Noisy conditions are modeled by adding a white Gaussian noise to the speech signal at various Signal-to-Noise Ratios (SNRs), from almost clean conditions ($SNR = 80dB$) to strongly adverse environments ($SNR = 10dB$). Table 6.1 summarizes all test conditions, which makes a total of slightly more than 250,000 experiments. It is worth mentioning that the synthetic tests presented in this section focus on the study of non-pathological voices with a regular phonation. Although the glottal analysis of less regular voices (e.g presenting a jitter or a shimmer; or containing an additive noise component during the glottal production, as it is the case for a breathy voice) is a challenging issue, this latter problem is not addressed in the present study. Nonetheless Chapter 8 will investigate the use of the glottal contribution for detecting voice disorders.

Source			Filter	Noise
Pitch (Hz)	Oq	α_m	Vowel type	SNR (dB)
100:5:240	0.3:0.05:0.9	0.55:0.05:0.8	14 vowels	10:10:80

Table 6.1 - *Table of synthesis parameter variation range.*

The three source estimation techniques described in Section 4.2 (CPIF, IAIF and CCD) are compared. In order to assess their decomposition quality, two objective quantitative measures are used (and the effect of noise, fundamental frequency and vocal tract variations to these measures are studied in detail in the next subsections):

- **Error rate on NAQ and QOQ** : As metioned in Section 4.3, the Normalized Amplitude Quotient (NAQ) and the Quasi Open Quotient (QOQ) are two important features characterizing the glottal flow. An error on their estimation after source-tract decomposition should therefore be penalized. An example of distribution for the relative error on QOQ in clean conditions is displayed in Figure 6.3. Many attributes characterizing such a histogram can be proposed to evaluate the performance of an algorithm. The one we used in our experiments is defined as the proportion of frames for which the relative error is higher than a given threshold of $\pm 20\%$. The lower the error rate on the estimation of a given glottal parameter, the better the glottal flow estimation method.

- **Spectral distortion** : As introduced in the beginning of Section 5.4 (Equation (5.17)), the Spectral Distortion (SD) is a simple and relevant measure, in the frequency domain, between the estimated and the real glottal pulse.

An efficient technique of glottal flow estimation is then reflected by low spectral distortion values. Based on this experimental framework, we now study how the glottal source estimation techniques behave in noisy conditions, or with regard to some factors affecting the decomposition quality, such as the fundamental frequency or the vocal tract transfert function.

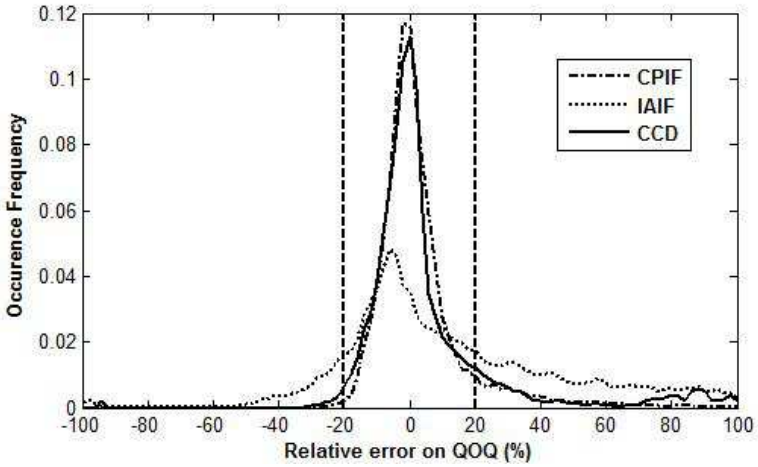

Figure 6.3 - *Distribution of the relative error on QOQ for the three methods in clean conditions ($SNR = 80dB$). The* error rate *is defined as the percentage of frames for which the relative error is higher than a given threshold of 20% (indicated on the plot).*

6.3.1 Robustness to Additive Noise

As mentioned above, white Gaussian noise has been added to the speech signal, with various SNR levels. This noise is used as a (weak) substitute for recording or production noise but also for every little deviation from the theoretical framework which distinguishes real and synthetic speech. Results according to our three performance measures are displayed in Figure 6.4. As expected, all techniques degrade as the noise power increases. More precisely, CCD turns out to be particularly sensitive. This can be explained by the fact that a weak presence of noise may dramatically affect the phase information, and consequently the decomposition quality. The performance of CPIF is also observed to strongly degrade as the noise level increases. This is probably due to the fact that noise may dramatically modify the spectral envelope estimated during the closed phase, and the resulting estimate of the vocal tract contribution becomes erroneous. On the contrary, even though IAIF is, in average, the less efficient on clean synthetic speech, it outperforms other techniques in adverse conditions (below 40 dB of SNR). One possible explanation of its robustness is the iterative process which it relies on. It can be indeed expected that, although the first iteration may be highly affected by noise (as it is the case for CPIF), the severity of the perturbation becomes weaker as the iterative procedure converges.

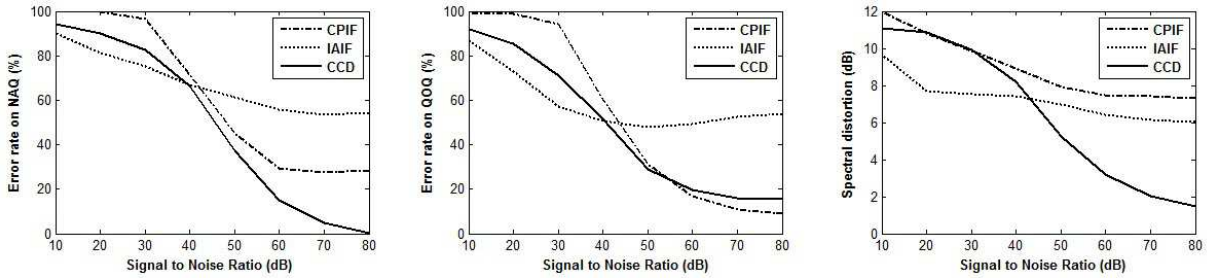

Figure 6.4 - *Evolution of the three performance measures (error rate on NAQ and QOQ, and spectral distortion) as a function of the Signal to Noise Ratio for the three glottal source estimation methods.*

6.3.2 Sensitivity to Fundamental Frequency

Female voices are known to be especially difficult to analyze and synthesize. The main reason for this is their high fundamental frequency which implies to process shorter glottal cycles. As a matter of fact the vocal tract response has not the time to freely return to its initial state between two glottal sollication periods (i.e. the duration of the vocal tract response can be much longer than that of the glottal closed phase). Figure 6.5 shows the evolution of our three performance measures with respect to the fundamental frequency in clean conditions. Interestingly, all methods maintain almost the same efficiency for high-pitched voices. Nonetheless an increase of the error rate on QOQ for CPIF, and an increase of the spectral distortion for CCD can be noticed. It can be also observed that, for clean synthetic speech, CCD gives the best results with an excellent determination of NAQ and a very low spectral distortion. Secondly, despite its high spectral errors, CPIF leads to an efficient parametrization of the glottal shape (with notably the best results for the determination of QOQ).

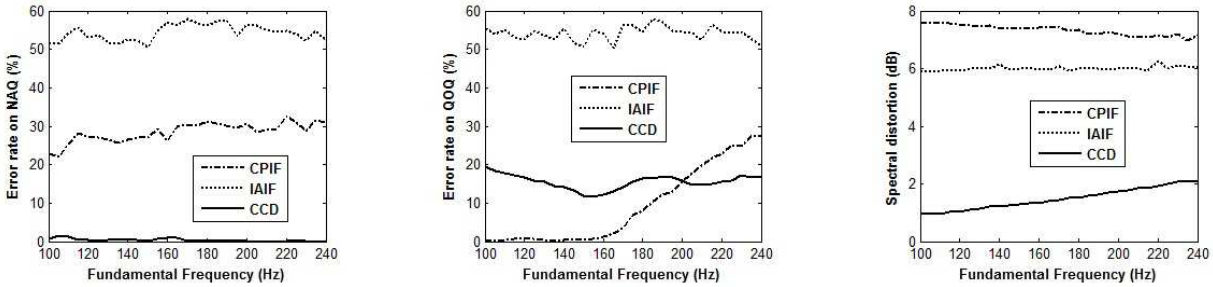

Figure 6.5 - *Evolution of the three performance measures as a function of F_0.*

6.3.3 Sensitivity to Vocal Tract

In our experiments, filter coefficients were extracted by LPC analysis on sustained vowels. Even though the whole vocal tract spectrum may affect the decomposition, the first formant, which corresponds to the dominant poles, generally imposes the longest contribution of its time response. To give an idea of its impact, Figure 6.6 exhibits, for the 14 vowels, the evolution of the spectral distortion as a function of the first formant frequency F_1. A general trend can be noticed from this graph: it is observed for all methods that the performance of the glottal flow estimation degrades as F_1 decreases. This will be explained in the next section by an increasing overlap between source and filter components, as the vocal tract impulse response gets longer. It is also noticed that this degradation is particularly important for both CPIF and IAIF methods, while the quality of CCD (which does not rely on a parametric modeling) is only slightly altered.

6.3.4 Conclusions on Synthetic Speech

Many factors may affect the quality of the source-tract separation. Intuitively, one can think about the *time interference* between minimum and maximum-phase contributions, respectively related to the vocal tract and to the glottal open phase. The stronger this interference, the more important the time overlap between the minimum-phase component and the maximum-phase response of the next glottal cycle, and consequently the more difficult the decomposition. Basically, this interference is conditioned by three main parameters:

- the pitch F_0, which imposes the spacing between two successive vocal system responses,

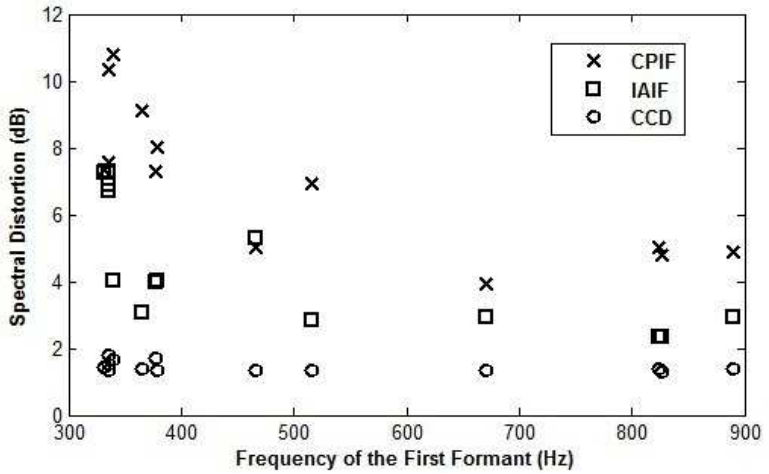

Figure 6.6 - *Evolution, for the 14 vowels, of the spectral distortion with the first formant frequency* F_1.

- the first formant F_1, which influences the length of the minimum-phase contribution of speech,
- and the glottal formant F_g, which controls the length of the maximum-phase contribution of speech. Indeed, the glottal formant is the most important spectral feature of the glottal open phase (see the low-frequency resonance in the amplitude spectrum of the glottal flow derivative in Figure 4.2). It is worth noting that F_g is known [14] to be a function of the time-domain characteristics of the glottal open phase (i.e of the maximum-phase component of speech): the open quotient O_q, and the asymmetry coefficient α_m.

A strong interference then appears with high pitch, and with low F_1 and F_g values. The previous experiments confirmed for all glottal source estimation techniques the performance degradation as a function of F_0 and F_1. Although we did not explicitly measure the sensitivity of these techniques to F_g in this chapter, it was confirmed in other informal experiments we performed.

It can be also observed from Figures 6.4 and 6.5 that the overall performance through an objective study on synthetic signals is the highest for the complex cepstrum-based technique. This method leads to the lowest values of spectral distortion and gives relatively high rates for the determination of both NAQ and QOQ parameters. The CPIF technique exhibits better performance in the determination of QOQ in clean conditions and especially for low-pitched speech. As for the IAIF technique, it turns out that it gives the worst results in clean synthetic speech but outperforms other approaches in adverse noisy conditions. Note that our results corroborate the conclusions drawn in [2] where the mixed-phase deconvolution (achieved in that study by the ZZT method) was shown to outperform other state-of-the-art approaches of glottal flow estimation.

6.4 Experiments on Real Speech

Reviewing the glottal flow estimation literature, one can easily notice that testing with natural speech is a real challenge. Even in very recent published works, all tests are performed only on sustained vowels. In addition, due to the unavailability of a reference for the real glottal flow (see Section 6.1), the procedure of evaluation is generally limited to providing plots of glottal flow estimates, and checking visually if they are consistent with expected glottal flow models. For real speech experiments,

here we will first follow this state-of-the-art experimentation (of presenting plots of estimates for a real speech example), and then extend it considerably both by extending the content of the data to a large connected speech database (including non-vowels), and extending the method to a comparative parametric analysis approach.

In this study, experiments on real speech are carried out on the De7 corpus[1], a diphone database designed for expressive speech synthesis [15]. The database contains three voice qualities (modal, soft and loud) uttered by a German female speaker, with about 50 minutes of speech available for each voice quality (leading to a total of around 2h30). Recordings sampled at 16 kHz are considered. Locations of both GCIs and GOIs are precisely determined from these signals using the SEDREAMS algorithm described in Section 3.3 (or [5]). As mentioned in Section 4.2, an accurate position of both events is required for an efficient CPIF technique, while the mixed-phase decomposition (as achieved by CCD) requires, among others, GCI-centered windows to exhibit correct phase properties.

Let us first consider in Figure 6.7 a concrete example of glottal source estimation on a given voiced segment (*/aI/* as in "*ice*") for the three techniques and for the three voice qualities. In the IAIF estimate, some ripples are observed as if some part of the vocal tract filter contribution could not be removed. On the other hand, it can be noticed that the estimates from CPIF and CCD are highly similar and are very close to the shape expected by the glottal flow models, such as the LF model [16]. It can be also observed that the abruptness of the glottal open phase around the GCI is stronger for the loud voice, while the excitation for the softer voice is smoother.

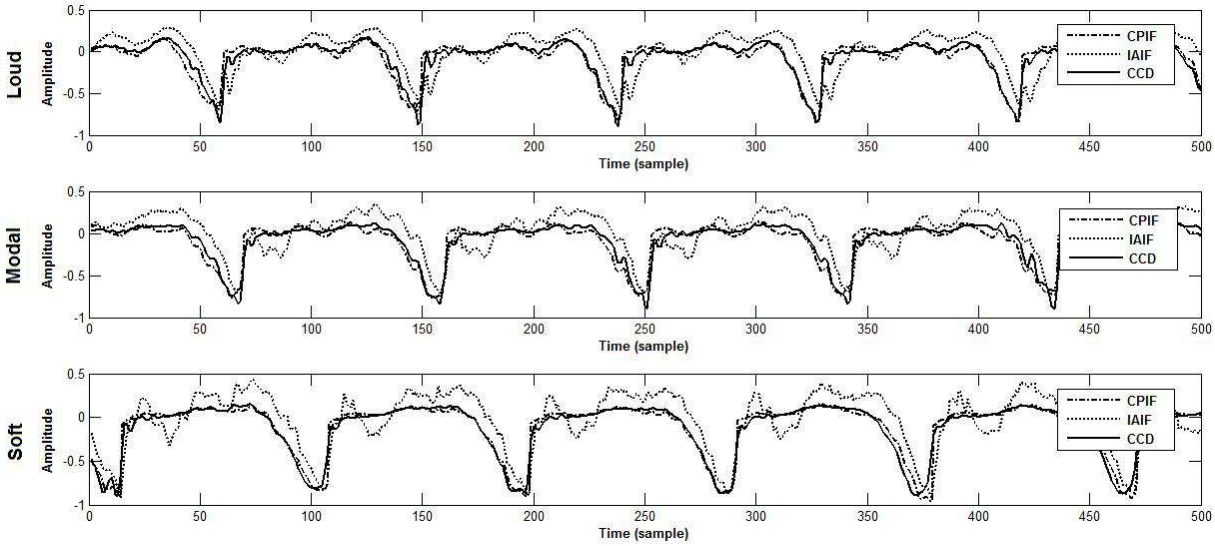

Figure 6.7 - *Example of glottal flow derivative estimation on a given segment of vowel (*/aI/ *as in "*ice*") for the three techniques and for the three voice qualities:* (top) *loud voice,* (middle) *modal voice,* (bottom) *soft voice.*

We now investigate whether the glottal source estimated by these techniques conveys information about voice quality. Indeed the glottis is assumed to play an important part for the production of such expressive speech [17]. As a matter of fact some differences between the glottal features are found in our experiments on the De7 database. In this experiment, the NAQ, H1-H2 and HRF parameters described in Section 4.3 are used. Figure 6.8 illustrates the distributions of these features estimated by CPIF, IAIF and CCD for the three voice qualities. This figure can be considered as a summary of the

[1]Many thanks to M. Schroeder for providing the De7 database.

voice quality analysis using three state-of-the-art methods on a large speech database. The parameters NAQ, H1-H2 and HRF have been used frequently in the literature to label phonation types [18], [19], [20]. Hence the separability of the phonation types based on these parameters can be considered as a measure of effectiveness for a particular glottal flow estimation method.

For the three methods, significant differences between the histograms of the different phonation types can be noted. This supports the claim that, by applying one of the given glottal flow estimation methods and by parametrizing the estimate with one or more of the given parameters, one can perform automatic voice quality/phonation type labeling with a much higher success rate than by random labeling. It is noticed from Figure 6.8 that parameter distributions are convincingly distinct, except for the IAIF and H1-H2 combination. The sorting of the distributions with respect to vocal effort are consistent and in line with results of other works ([18] and [21]). Among other things, strong similarities between histograms obtained by CPIF and CCD can be observed. In all cases, it turns out that the stronger the vocal effort, the lower NAQ and H1-H2, and the higher HRF.

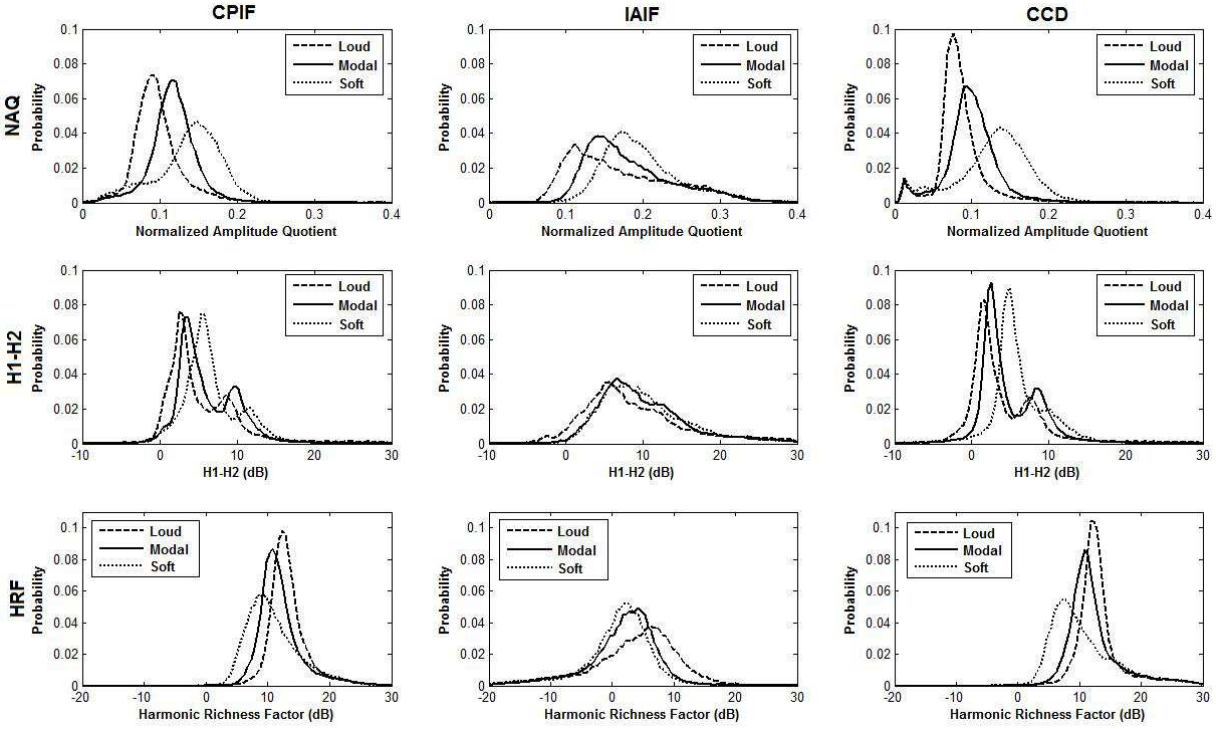

Figure 6.8 - *Distributions, for various voice qualities, of three glottal features (from top to bottom: NAQ, H1-H2 and HRF) estimated by three glottal source estimation techniques (from left to right: CPIF, IAIF and CCD). The voice qualities are shown as dashed (loud voice), solid (modal voice) and dotted (soft voice) lines.*

Although some significant differences in glottal feature distributions have been visually observed, it is interesting to quantify the discrimination between the voice qualities enabled by these features. For this, the Kullback-Leibler (KL) divergence, known to measure the separability between two discrete density functions A and B, can be used [22]:

$$D_{KL}(A,B) = \sum_i A(i) \log_2 \frac{A(i)}{B(i)} \tag{6.1}$$

Since this measure is non-symmetric (and consequently is not a true distance), its symmetrised

version, called Jensen-Shannon divergence, is often prefered. It is defined as a sum of two KL measures [22]:

$$D_{JS}(A,B) = \frac{1}{2}D_{KL}(A,M) + \frac{1}{2}D_{KL}(B,M) \tag{6.2}$$

where M is the average of the two distributions ($M = 0.5*(A+B)$). Figure 6.9 displays the values of the Jensen-Shannon distances between two types of voice quality, for the three considered features estimated by the three techniques.

From this figure, it can be noted that NAQ is the best discriminative feature (i.e. has the highest Jensen-Shannon distance between distributions), while H1-H2 and HRF convey a comparable amount of information for discriminating voice quality. As expected, the loud-soft distribution distances are highest compared to loud-modal and modal-soft distances. In seven cases out of nine (three different parameters and three different phonation type couples), CCD leads to the most relevant separation and in two cases (loud-modal separation with NAQ, loud-modal separation with HRF) CPIF provides a better separation. Both Figures 6.8 and 6.9 show that the effectiveness of CCD and CPIF is similar, with slightly better results for CCD, while IAIF exhibits clearly lower performance (except for one case: loud-modal separation with HRF).

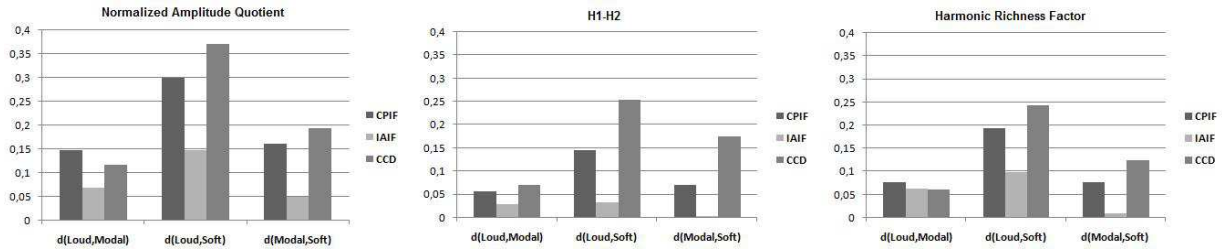

Figure 6.9 - *Jensen-Shannon distances between two types of voice quality using (from left to right) the NAQ, H1-H2 and HRF parameters. For each feature and pair of phonation types, the three techniques of glottal source estimation are compared.*

6.5 Conclusion

This study aimed at comparing the effectiveness of the main state-of-the-art glottal flow estimation techniques. For this, detailed tests on both synthetic and real speech were performed. For real speech, a large corpus was used for testing, without limiting analysis to sustained vowels. Due to the unavailability of the reference glottal flow signals for real speech examples, the separability of three voice qualities was considered as a measure of the ability of the methods to discriminate different phonation types. In synthetic speech tests, objective measures were used since the original glottal flow signals were available. Our first conclusion is that the usefulness of NAQ, H1-H2 and HRF for parameterizing the glottal flow is confirmed. We also confirmed other works in the literature (such as [18] and [21]) showing that these parameters can be effectively used as measures for discriminating different voice qualities. Our results show that the effectiveness of CPIF and CCD appears to be similar and rather high, with a slight preference towards CCD. However, it should be emphasized here that in our real speech tests, clean signals recorded for Text-To-Speech (TTS) synthesis were used. We can thus confirm the effectiveness of CCD for TTS applications (such as emotional/expressive TTS). However, for applications which require the analysis of noisy signals (such as telephone applications) further testing is needed. We observed that in the synthetic speech tests, the ranking dramatically

changed depending on the SNR and the robustness of CCD was observed to be rather low. IAIF has lower performance in most tests (both in synthetic and real speech tests) but shows up to be comparatively more effective in very low SNR values.

Bibliography

[1] J. Walker and P. Murphy. A review of glottal waveform analysis. In *Progress in Nonlinear Speech Processing*, pages 1–21, 2007.

[2] N. Sturmel, C. d'Alessandro, and B. Doval. A comparative evaluation of the zeros of z transform representation for voice source estimation. In *Proc. Interspeech*, pages 558–561, 2007.

[3] A. El Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans. on Signal Processing*, 39 (2):411–423, 1991.

[4] S. Saito F. Itakura. A statistical method for estimation of speech spectral density and formant frequencies. *Electron. Commun. Japan*, 53-A:36–43, 1970.

[5] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[6] D. Brookes and D. Chan. Speaker characteristics from a glottal airflow model using glottal inverse filtering. *Institue of Acoust.*, 15:501–508, 1994.

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

[8] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–586, 1999.

[9] P. Alku, J. Svec, E. Vilkman, and F. Sram. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

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

[11] M. Airas. Tkk aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

[12] Online. http://aparat.sourceforge.net/index.php/main_page. *TKK Aparat Main Page*, 2008.

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

[14] B. Doval and C. d'Alessandro. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006.

[15] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[16] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[17] C. d'Alessandro. Voice source parameters and prosodic analysis. In *Method in empirical prosody research*, pages 63–87, 2006.

[18] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112:701–710, 2002.

[19] H. Hanson. Individual variations in glottal characteristics of female speakers. In *Proc. ICASSP*, pages 772–775, 1995.

[20] D. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:2394–2410, 1991.

[21] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[22] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. on Information Theory*, 37:145–151, 1991.

Chapter 7

Glottal Source Estimation using an Automatic Chirp Decomposition

Contents

Abstract

In Chapter 5, it was shown that the glottal source can be effectively estimated by separating the causal and anticausal components of speech. Two algorithms were proposed for achieving the mixed-phase separation: the Zeros of the Z-Transform (ZZT) and the Complex Cesptrum-based Decomposition (CCD). In order to guarantee a correct estimation, some constraints on the window have been derived. Among these, the window has to be synchronized on a Glottal Closure Instant (GCI). This chapter extends the formalism of both ZZT and CCD methods by evaluating the z-transform on a contour (called *chirp* contour) possibly different from the unit circle. For each method a technique is proposed for the automatic determination of the optimal contour. The resulting method is shown to give a reliable estimation of the glottal flow wherever the window is located. This technique is then suited for its integration in usual speech processing systems, which generally operate in an asynchronous way.

This chapter is based upon the following publications:

- Thomas Drugman, Thierry Dutoit, *Chirp Complex Cepstrum-based Decomposition for Asynchronous Glottal Analysis*, Interspeech Conference, Makuhari, Japan, 2010.
- Thomas Drugman, Baris Bozkurt, Thierry Dutoit, *Glottal Source Estimation Using an Automatic Chirp Decomposition*, Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing, volume 5933, pp. 35-42, 2010.

Many thanks to Dr. Baris Bozkurt (Izmir Institute of Technology) for his helpful guidance.

7.1 Introduction

It has been demonstrated in Chapter 5 that the Zeros of the Z-Transform (ZZT) and the Complex Cesptrum-based Decomposition (CCD) techniques are functionally equivalent, and that they can be efficiently used for source-tract separation. CCD was also compared in Chapter 6 to other state-of-the-art methods for glottal source estimation, where it was shown to provide the best results on both synthetic and real speech. However, as highlighted in Section 5.4.1, an essential constraint for leading to a correct source-tract separation with these techniques is the condition of being synchronized on a Glottal Closure Instant (GCI). Although some works aim at estimating the GCI positions directly from the speech signal (see Chapter 3 or [1]), or use ElectroGlottoGraphs (EGGs, [2]), the large majority of current speech processing systems do not have this information available and consequently operate in an asynchronous way, i.e use a constant frame shift. This chapter proposes a modification of the formalism of both ZZT and CCD so that they can be integrated within asynchronous systems. For this, the z-transform is evaluated on a contour (called chirp contour) possibly different from the unit circle. A way to automatically determine the optimal contour is also proposed for both ZZT and CCD. As a result, it will be shown that the estimation is much less sensitive to GCI detection errors.

The chapter is structured as follows. Section 7.2.1 investigates the extension of the ZZT formalism by making use of a chirp analysis. The resulting method is then evaluated on both synthetic and real speech signals in Section 7.2.2. The same issue is adressed for the CCD approach in Section 7.3. The performance of the chirp CCD method is then assessed on a large expressive speech corpus in Section 7.3.2. Finally Section 7.4 concludes the chapter.

7.2 Extension of the ZZT Method to Chirp Decomposition

7.2.1 Theoretical Framework

As introduced in Section 5.3.1, the ZZT technique achieves mixed-phase decomposition by calculating the zeros of the z-transform. Some of these zeros lie inside the unit circle, while others are located outside. The firsts are due to the causal (or minimum-phase) component of speech, which is related to the vocal tract response and the glottal return phase. On the contrary, zeros outside the unit circle are due to the anticausal (or maximum-phase) component of speech, which is related to the glottal open phase. The mixed-phase decomposition can then be achieved by separating the ZZT using the unit circle in the z-plane as a discriminant boundary. Nevertheless, to obtain such a separation, the effects of the windowing are known to play a crucial role, as emphasized in Section 5.4. In particular, it was notably shown that a Blackman window centered on the Glottal Closure Instant (GCI) and whose length is twice the pitch period is appropriate in order to achieve a good decomposition.

The Chirp Z-Transform (CZT), as introduced by Rabiner *et al* [3] in 1969, allows the evaluation of the z-transform on a spiral contour in the z-plane. Its first application aimed at separating too close formants by reducing their bandwidth. Nowadays CZT has been applied to several fields of signal processing such as time interpolation, homomorphic filtering, pole enhancement, narrow-band analysis,...

As previously mentioned, the ZZT-based decomposition is strongly dependent on the applied windowing. This sensitivity may be explained by the fact that ZZT implicitly conveys phase information, for which time alignment is known to be crucial. In [4], it is observed that the window shape and onset may lead to zeros whose topology can be detrimental for accurate pulse estimation. The subject of this work is precisely to handle these zeros close to the unit circle, such that the ZZT-based technique correctly separates the causal (i.e minimum-phase) and anticausal (i.e maximum-phase) components.

For this, we evaluate the CZT on a circle whose radius R is chosen so as to split the root distribution into two well-separated groups. More precisely, it is observed that the significant impulse present in the excitation at the GCI results in a gap in the root distribution. When analysis is exactly GCI-synchronous, the unit circle perfectly separates causal and anticausal roots. On the opposite, when the window moves off from the GCI, the root distribution is transformed. Such a decomposition is then not guaranteed for the unit circle and another boundary is generally required. Figure 7.1 gives an example of root distribution for a natural voiced speech frame for which an error of 0.6 ms is made on the real GCI position. It is clearly seen that using the traditional ZZT-based decomposition ($R = 1$) for this frame will lead to erroneous results. In contrast, it is possible to find an optimal radius leading to a correct separation (around 0.96 in this case).

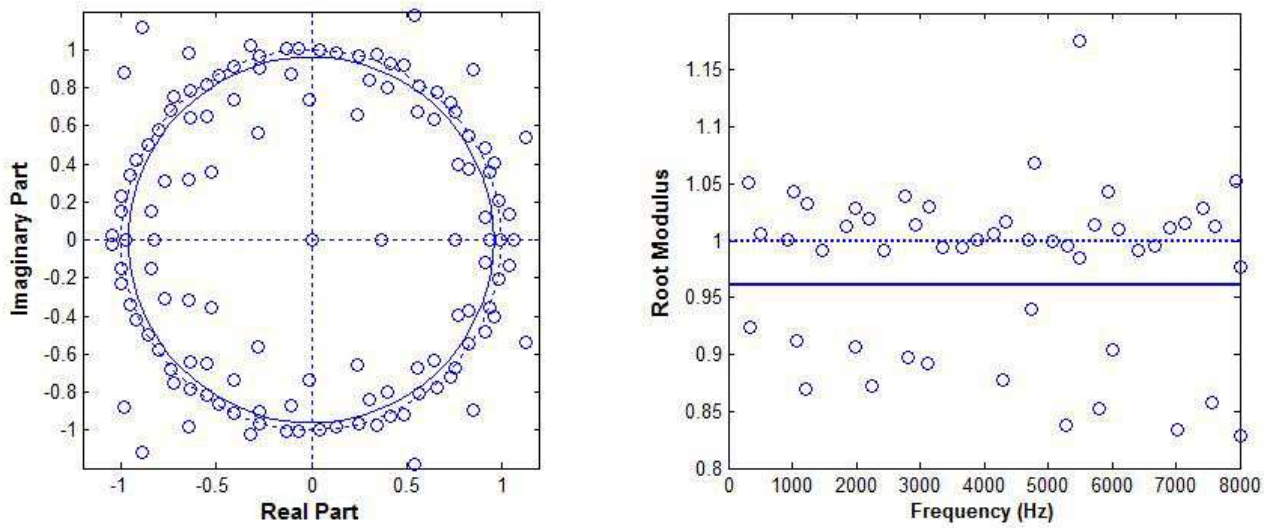

Figure 7.1 - *Example of root distribution for a natural speech frame.* Left panel*: representation in the z-plane,* Right panel*: representation in polar coordinates. The chirp circle (solid line) allows a correct decomposition, contrarily to unit circle (dotted line).*

In order to automatically determine such a radius, let us have the following thought process. We know that ideally the analysis should be GCI-synchronous. When this is not the case, the chirp analysis tends to modify the window such that its center coincides with the nearest GCI (to ensure a reliable phase information). Indeed, evaluating the chirp z-transform of a signal $x(t)$ on a circle of radius R is equivalent to evaluating the z-transform of $x(t) \cdot exp(log(1/R) \cdot t)$ on the unit circle (see Equation (7.3) as a proof). It can be demonstrated (see Appendix A) that for a Blackman window $w(t)$ of length L starting in $t = 0$:

$$w(t) = 0.42 - 0.5 \cdot \cos(\frac{2\pi t}{L}) + 0.08 \cdot \cos(\frac{4\pi t}{L}), \tag{7.1}$$

the radius R necessary to modify its shape so that its new maximum lies in position t^* ($< L$) is expressed as:

$$R = exp[\frac{2\pi}{L} \cdot \frac{41 \tan^2(\frac{\pi t^*}{L}) + 9}{25 \tan^3(\frac{\pi t^*}{L}) + 9 \tan(\frac{\pi t^*}{L}))}]. \tag{7.2}$$

In particular, we verify that $R = 1$ is optimal when the window is GCI-centered ($t^* = \frac{L}{2}$) and, since we are working with two-period long windows, the optimal radius does not exceed $exp(\pm\frac{50\pi}{17L})$ in

the worst cases (the nearest GCI is then positioned in $t^* = \frac{L}{4}$ or $t^* = \frac{3L}{4}$).

As a means for automatically determining the radius allowing an efficient separation, the sorted root moduli are inspected and the greatest discontinuity in the interval $[exp(-\frac{50\pi}{17L}), exp(\frac{50\pi}{17L})]$ is detected. Radius R is then chosen as the middle of this discontinuity, and is assumed to optimally split the roots into minimum and maximum-phase contributions.

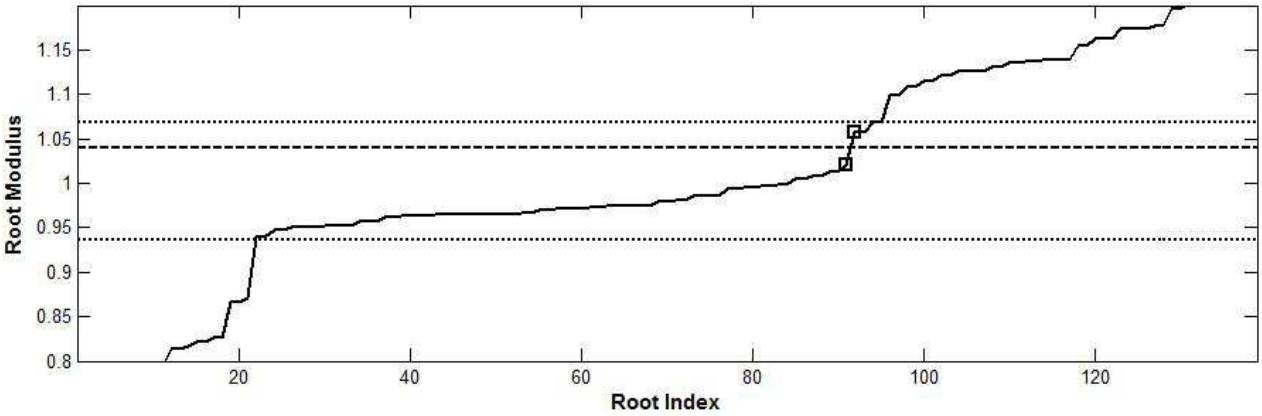

Figure 7.2 - *Determination of radius R (dashed line) for ZCZT computation by detecting, within the bounds $exp(\pm\frac{50\pi}{17L})$ (dotted lines), a discontinuity (indicated by rectangles) in the sorted root moduli (solid line).*

7.2.2 Evaluation

This section gives a comparative evaluation of the following methods:

- *the traditional ZZT-based technique*: $R = 1$,
- *the proposed ZCZT-based technique*: R is computed as explained at the end of Section 7.2.1 (see Fig. 7.2),
- *the ideal ZCZT-based technique*: R is computed from Equation 7.2 where the real GCI location t^* is known. This can be seen as the ultimate performance one can expect from the ZCZT-based technique.

Among others, it is emphasized how the proposed technique is advantageous in case of GCI location errors.

Results on Synthetic Speech

The experimental protocol on synthetic speech signals adopted in this section is identical to the one described in Section 5.4. In addition, a perturbation is taken into account by considering a possible error on the GCI location. This may vary between -50% and 50% of T_0, with a step of 5%. To evaluate the performance of our methods, the two same objective measures as in Section 5.4 are used, namely the determination rate on the glottal formant frequency F_g at 20% (i.e the percentage of frames for which the relative error made on F_g is lower than 20%), and the spectral distortion.

Figure 7.3 compares the results obtained for the three methods according to their sensitivity to GCI location. The high sensitivity of the traditional ZZT to GCI synchronization is clearly confirmed,

as the performance severely degrades as the window center moves off from the GCI. Interestingly, the proposed ZCZT-based technique is clearly seen as an enhancement of the traditional ZZT approach when an error on the exact GCI position is made. Besides the proposed approach is also observed to give a performance almost equivalent to the ideal ZCZT-based method.

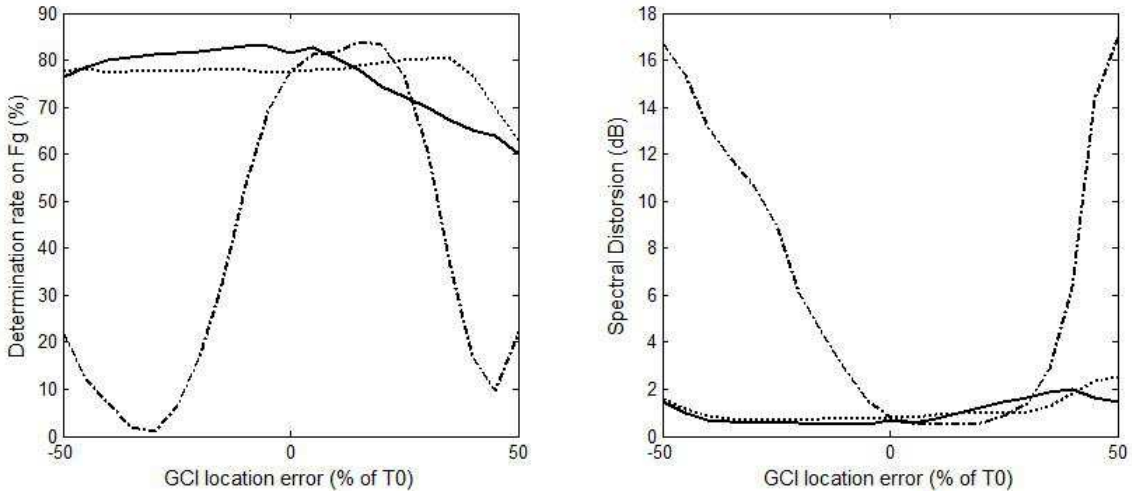

Figure 7.3 - *Comparison of the traditional ZZT (dashdotted line), proposed ZCZT (solid line) and ideal ZCZT (dotted line) based methods on synthetic signals according to their sensitivity to an error on the GCI location.* Left panel: *Influence on the determination rate on the glottal formant frequency.* Right panel: *Influence on the spectral distortion.*

Results on Real Speech

Figure 7.4 displays an example of decomposition on a real voiced speech segment (vowel /e/ from *BrianLou4.wav* of the Voqual03 database, $F_s = 16kHz$). The top panel exhibits the speech waveform together with the synchronized (compensation of the delay between the laryngograph and the microphone) differenced Electroglottograph (EGG) informative about the GCI positions. The center and bottom panels compare respectively the detected glottal formant frequency F_g and the radius for the three techniques. In the middle panel, deviations from the constant F_g can be considered as errors since F_g is expected to be almost constant during three pitch periods. It may be noticed that the traditional ZZT-based method degrades if analysis is not achieved in the GCI close vicinity. Contrarily, the proposed ZCZT-based technique gives a reliable estimation of the glottal source on a large segment around the GCI. Besides the obtained performance is comparable to what is carried out by the ideal ZCZT.

In Figure 7.5 the glottal source estimated by the traditional ZZT and the proposed ZCZT-based method are displayed for four different positions of the window (for the vowel /a/ from the same file). It can be observed that the proposed technique (solid line) gives a reliable estimation of the glottal flow wherever the window is located. On the contrary the sensivity of the traditional approach can be clearly noticed since its glottal source estimation turns out to be irrelevant when the analysis is not performed in a GCI-synchronous way.

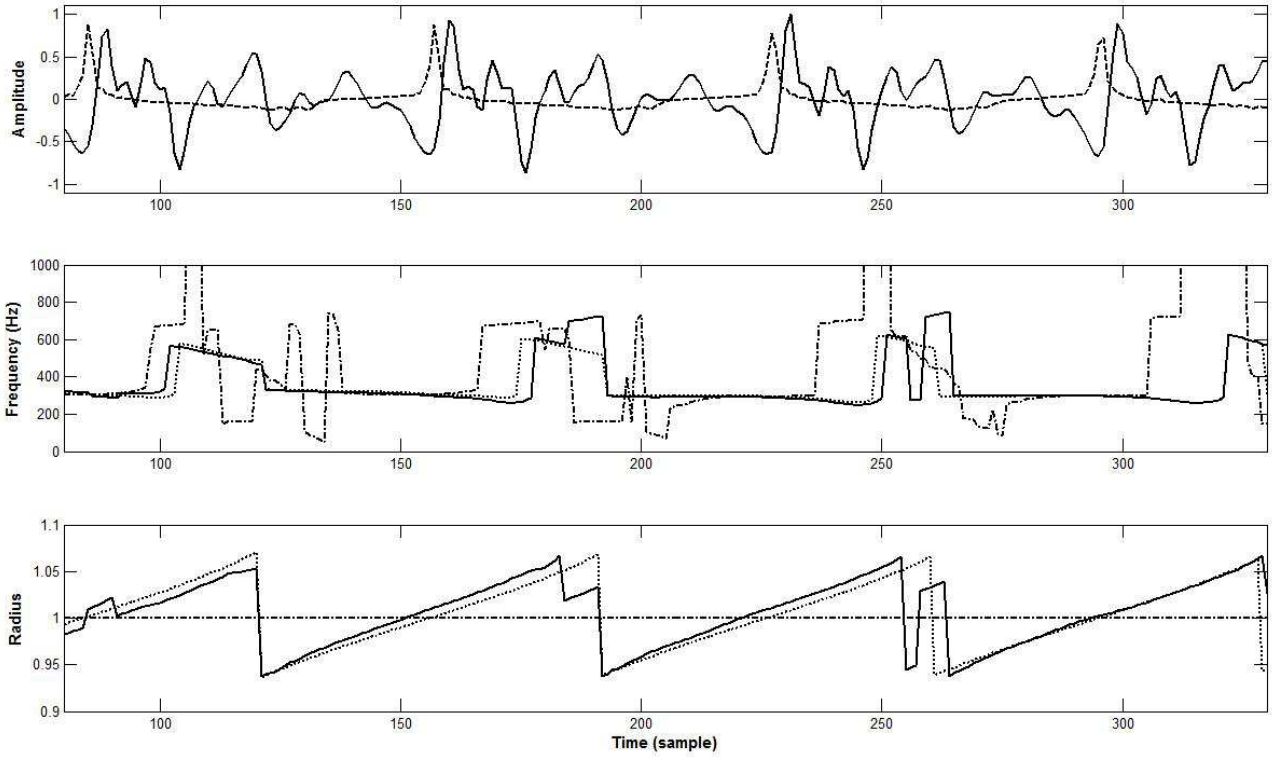

Figure 7.4 - *Comparison of ZZT and ZCZT-based methods on a real voiced speech segment.* Top panel: *the speech signal (solid line) with the synchronized differenced EGG (dashed line).* Middle panel: *the glottal formant frequency estimated by the traditional ZZT (dashdotted line), the proposed ZCZT (solid line) and the ideal ZCZT (dotted line) based techniques.* Bottom panel: *Their corresponding radius used to compute the chirp Z-transform.*

7.3 Extension of the Complex Cepstrum-based Method to Chirp Decomposition

7.3.1 Theoretical Framework

As it is the case for the ZZT technique, the principle of the Complex Cepstrum-based Decomposition (CCD) relies on the mixed-phase model of speech [5]. Although both techniques are functionnaly equivalent (see Section 5.3.2), CCD was shown in Chapter 5 to be much faster than ZZT. For this reason, this section only focuses on the use of the Complex Cepstrum.

In the chirp ZZT technique introduced in Section 7.2, the whole root distribution is calculated and, relying on this distribution, a chirp contour in the z-plane is found so as to optimally split the minimum and maximum-phase contributions. On the contrary, the CCD method aims at avoiding the root computation by making use of the Complex Cepstrum. Based on the conclusions of Section 7.2, it is here proposed to integrate the chirp analysis whitin the Complex Cepstrum-based Decomposition and to automatically find the optimal circle *without requiring the computation of the root distribution* (so as to still benefit from the speed of the CCD algorithm).

Achieving a chirp ZZT-based decomposition is straightforward since it is only necessary to modify the radius used to isolate the maximum-phase component. In order to integrate the chirp analysis for the CCD technique, let us consider the signal $x(n)$. Its CZT evaluated on a circle of radius R can be

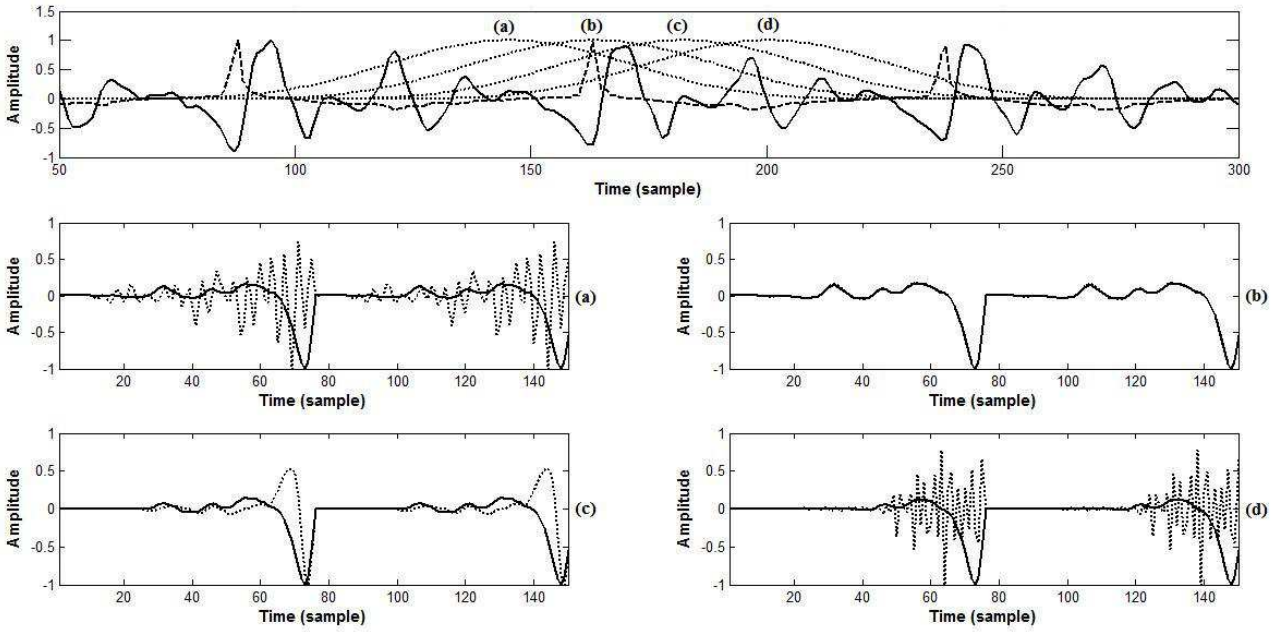

Figure 7.5 - *Examples of glottal source estimation using either the traditional ZZT or the proposed ZCZT-based method.* Top panel: *a voiced speech segment (solid line) with the synchronized differenced EGG (dashed line) and four different positions of the window (dotted line).* Panels (a) to (d): *for the corresponding window location, two cycles of the glottal source estimation achieved by the traditional ZZT (dotted line) and by the proposed ZCZT-based technique (solid line).*

written as [3]:

$$X(Rz) = \sum_{n=0}^{L-1} x(n)(Rz)^{-n} = \sum_{n=0}^{L-1} (x(n)R^{-n})z^{-n} \tag{7.3}$$

and is consequently equivalent to evaluating the z-transform of a signal $x'_R(n) = x(n)R^{-n}$ on the unit circle. The chirp CCD computed on a circle of radius R can therefore be achieved by applying the traditional CCD framework described in Section 5.3.2 to $x'_R(n)$ instead of $x(n)$.

In order to automatically estimate the radius giving an optimal separation between minimum and maximum-phase contributions, the unwrapped phase $\phi'_R(\omega)$ of $x'_R(n)$ is inspected. More precisely, the radius axis is uniformly discretized in N values ($N = 60$ in our experiments) between the bounds $exp(\pm\frac{50\pi}{17L})$. For each radius value R, $\phi'_R(\omega)$ is computed and the linear phase component is characterized by $\phi'_R(\pi)$ (with $\phi'_R(0) = 0$ by definition). From this, we define the variable $n_d(R)$:

$$n_d(R) = \frac{\phi'_R(\pi)}{\pi} \tag{7.4}$$

as the number of samples of circular delay, i.e the number of samples that $x'_R(n)$ shoud be circularly shifted so as to remove its linear phase component.

Figure 7.6 shows the evolution of $n_d(R)$ for the same signal as used in Figure 7.1. $n_d(R)$ is actually a step function where gaps are due to the passage of some roots from the inside to the outside of the considered chirp circle. These phase discontinuities are illustrated in Figure 7.7. Indeed consider a

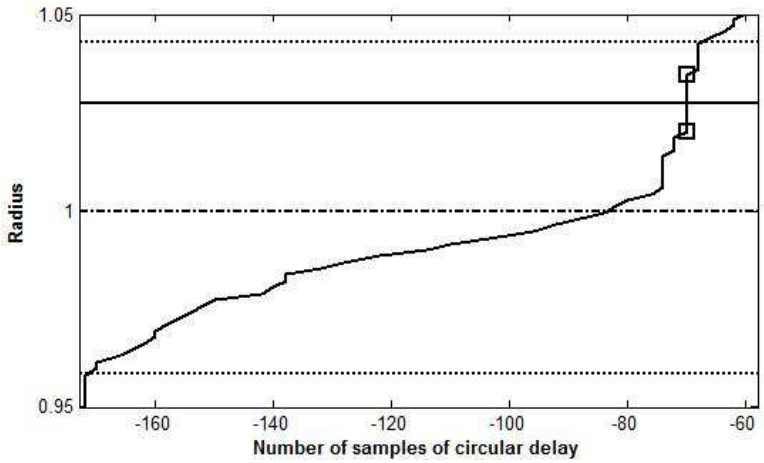

Figure 7.6 - *Evolution of $n_d(R)$ for the same signal as in Figure 7.1. The optimal radius (solid line) is defined as the middle of the largest interval (indicated by squares) for which $n_d(R)$ stays constant, within the bounds $exp(\pm\frac{50\pi}{17L})$ (dotted lines). The unit circle used in the traditional CCD is represented in dashdotted line.*

zero which, initially located inside the circle of radius R_1 used for the evaluation of the CZT, is now passed outside of the circle of radius R_2 (with $R_1 > R_2$). When the CZT is evaluated on a point close to this zero in the z-plane, this results in a phase jump of $-\pi$ (see angles α_1 and α_2 in Fig. 7.7) which is then reflected in $\phi'_R(\pi)$. The difference $n_d(R_1) - n_d(R_2)$ is consequently interpreted as the number of zeros which, initially inside the circle of radius R_1, have passed the boundary to be now located outside the circle of radius R_2. In other words, inspecting the variable $n_d(R)$ allows us to detect the discontinuities in the root distribution (without requiring its whole computation). Similarly to what is done in Section 7.2.1 for the chirp ZZT, the optimal radius used for the chirp CCD is then defined as the middle of the largest interval for which $n_d(R)$ stays constant, within the bounds $exp(\pm\frac{50\pi}{17L})$.

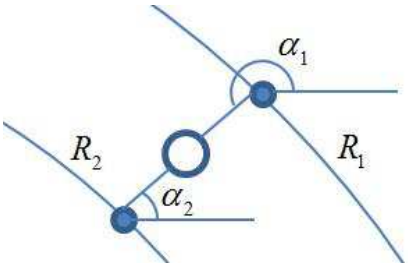

Figure 7.7 - *Illustration of a phase jump of $-\pi$ due to the passage of a zero from the inside of the circle of radius R_1 to the outside of the circle of radius R_2.*

7.3.2 Evaluation

Experiments are carried out on the De7 corpus [1]. The database contains three voice qualities (modal, soft and loud) uttered by a German female speaker, with about 50 minutes of speech available for each voice quality [6]. Besides, GCI positions are estimated by the SEDREAMS algorithm described in Section 3.3.

[1] Many thanks to Marc Schroeder for providing the De7 database.

The goal of this section is to compare the traditional and the proposed chirp CCD techniques by studying their efficiency for glottal source estimation. Experiments are divided into two parts. In the first one, the sensitivity of both methods to GCI location errors is investigated. In the second part, the whole expressive speech database is analyzed by the two techniques and it is shown that chirp CCD leads to results similar to the traditional CCD, but without the requirement of operating in a GCI-synchronous way.

Robustness to GCI location errors

When performing mixed-phase separation, it may appear for some frames that the decomposition is erroneous, leading to an irrelevant high-frequency noise in the estimated glottal source (see Section 5.5.3). A criterion based on the spectral center of gravity and deciding whether a frame is considered as correctly decomposed or not, has been proposed in Section 5.5.3. Relying on Figure 5.13, it has been shown that fixing a threshold for the spectral center of gravity of the estimated glottal source at around 2.7kHz makes a good distinction between frames that are correctly and incorrectly decomposed.

Given this criterion, the sensitivity of both traditional and chirp CCD techniques to a GCI location error (assuming the GCI detection by SEDREAMS as a reference) is displayed in Figure 7.8 for the loud dataset of the De7 corpus. The constraint of being GCI-synchronous for the traditional CCD is clearly confirmed on this graph. It is indeed seen that the performance dramatically degrades for this technique as the window center moves off from the GCI. On the contrary, the chirp CCD method gives a high rate of correctly decomposed frames (however slightly below the performance of the GCI-centered traditional CCD) wherever the window is located.

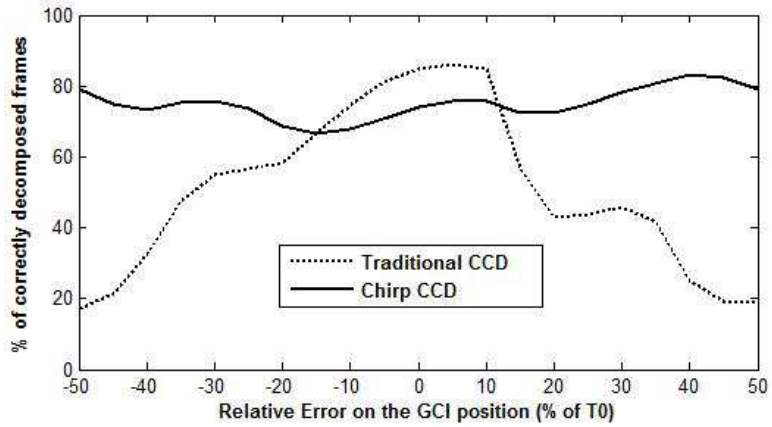

Figure 7.8 - *Robustness of both traditional and chirp CCD methods to a timing error on the GCI location.*

Asynchronous glottal analysis of emotional speech

In this section, we confirm the potential of the chirp CCD technique for asynchronously estimating the glottal flow on a large speech corpus. For this, the whole De7 database with its 3 voice qualities is analyzed. The glottal flow is estimated by 2 techniques:

- *the traditional CCD:* analysis is GCI-synchronous (the GCI positions being determined by the SEDREAMS algorithm),

- *the chirp CCD:* analysis is asynchronous. A constant frame shift of $10ms$ is considered, as widely used in many speech processing systems. Note however that a two pitch period-long Blackman window is applied, as this is essential for achieving a correct mixed-phase decomposition (see Section 5.4.2).

In a first time, we evaluate the proportion of frames that are correctly decomposed by these two techniques using the spectral center of gravity criterion. Overall results for the three voice qualities are summarized in Table 7.1. The traditional CCD gives relatively high rates of correct decomposition with around 85% for the three datasets. It can also be observed that the chirp CCD method makes double the erroneous decompositions than the traditional approach. Nevertheless a correct estimation of the glottal source is carried out by the chirp CCD for around 70% of speech frames, which is rather high for real connected speech.

Method	Loud	Modal	Soft
traditional CCD	87.22	84.41	83.69
chirp CCD	76.43	68.07	67.48

Table 7.1 - *Proportion (%) of correctly decomposed frames using the traditional and the chirp CCD techniques for the three voice qualities of the De7 database.*

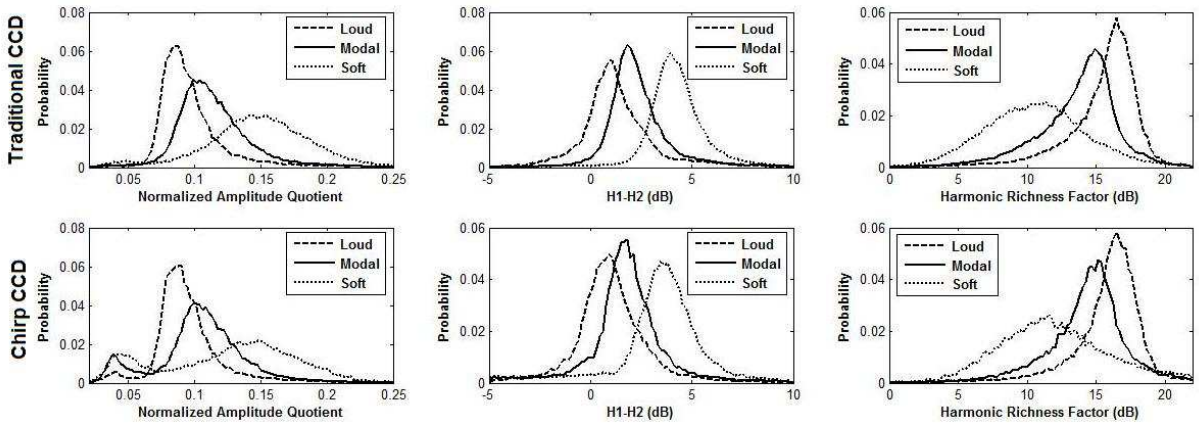

Figure 7.9 - *Distributions of glottal parameters estimated by (from top to bottom) the traditional and chirp CCD techniques, for three voice qualities. The considered glottal features are (from left to right): the Normalized Amplitude Quotient (NAQ), the H1-H2 ratio and the Harmonic Richness Factor (HRF).*

In a second time, frames of the glottal flow that were correctly estimated are characterized by the three following features (see Section 4.3 for their definition): the Normalized Amplitude Quotient (NAQ), the H1-H2 ratio between the two first harmonics and the Harmonic Richness Factor (HRF). These glottal parameters were shown in [7] and [8] to lead to a good separation between different types of phonation. The histograms of these parameters estimated by both traditional and chirp CCD methods are displayed in Figure 7.9 for the three voice qualities. Two main conclusions can be drawn from this figure. First, it turns out that the distributions obtained by both techniques are strongly similar. A minor difference can however be noticed for NAQ histograms, where the distributions obtained by the chirp method contain a weak irrelevant peak at low NAQ values. The second important conclusion is

that both techniques can be efficiently used for glottal-based voice quality analysis, leading to a clear discrimination between various phonation types.

7.4 Conclusion

This chapter proposed an extension of both the traditional Zeros of the Z-Transform (ZZT) and the Complex Cepstrum-based Decomposition (CCD) techniques. For this, the z-transform was evaluated on a contour in the z-plane possibly different from the unit circle. Circular contours were considered and an automatic way to find the optimal radius leading to well-separated groups of zeros was proposed. The resulting methods were shown to be much more robust to Glottal Closure Instant location errors than their traditional (non chirp) equivalent. Interestingly a reliable estimation of the glottal flow was obtained in an asynchronous way on real connected speech. Besides the proposed chirp CCD technique showed its potential to be used for automatic voice quality analysis. Thanks to its low computational load, the chirp CCD method is then suited for being incorporated within a real-time asynchronous speech processing application.

In the current version of the chirp CCD algorithm, the calculation of the linear phase $\phi'_R(\pi)$ is achieved using operations of FFT and phase unwrapping, which may be suboptimal in terms of computation time. Further work could address the acceleration of this process, for example relying on the time center of gravity [9], or using properties of the root distribution with respect to the unit circle [10].

Bibliography

[1] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[2] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. *J. Acoust. Soc. Am.*, 115:1321–1332, 2004.

[3] L. Rabiner, R. Schafer, and C. Rader. The chirp-z transform algorithm and its application. *Bell System Technical Journal*, 48(5):1249–1292, 1969.

[4] J. Tribolet, T. Quatieri, and A. Oppenheim. Short-time homomorphic analysis. In *Proc. ICASSP*, volume 2, pages 716–72, 1977.

[5] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

[6] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[7] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112:701–710, 2002.

[8] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[9] Y. Stylianou. Removing phase mismatches in concatenative speech synthesis. In *Speech Synthesis Workshop 3*, pages 267–272, 1998.

[10] M. Benidir. On the root distribution of general polynomials with respect to the unit circle. *Signal Processing*, 53(1):75–82, 1996.

Chapter 8

Using Glottal-based Features for the Automatic Detection of Voice Pathologies

Contents

Abstract

This chapter addresses the problem of automatic detection of voice pathologies directly from the speech signal. More precisely, we investigate the use of the glottal source estimation and phase-based features as a means to discriminate voice disorders. First we analyze the complementarity of characteristics derived from the glottal flow with features related to the speech signal, or to prosody. The relevance of these features is assessed through mutual information-based measures. This allows an intuitive interpretation in terms of discrimation power and redundancy between the features, independently of any subsequent classifier. We analyze which characteristics are interestingly informative or complementary for detecting voice pathologies. In a second time, we explore the potential of using phase-based features for automatically detecting voice disorders. It is shown that group delay functions are appropriate for characterizing irregularities in the phonation. Besides the adequacy of the mixed-phase model of speech is discussed. The proposed phase-based features are evaluated and compared to other parameters derived from the magnitude spectrum. They turn out to convey a great amount of relevant information, leading to high discrimination performance. The detection performance is also assessed via the use of an Artificial Neural Network.

This chapter is based upon the following publications:

- Thomas Drugman, Thomas Dubuisson, Thierry Dutoit, *On the Mutual Information between Source and Filter Contributions for Voice Pathology Detection*, Interspeech Conference, Brighton, United Kingdom, 2009.
- Thomas Drugman, Thomas Dubuisson, Thierry Dutoit, *Phase-based Information for Voice Pathology Detection*, IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 2011.

Many thanks to Thomas Dubuisson for his collaboration.

8.1 Introduction

Voice pathology refers to the diagnosis and treatment of functional and organic speech defects and disorders [1]. These may have several origins, such as vocal cords nodule or polyps, Reinke's edema, etc [1]. The acoustic evaluation of voice disorders is an essential tool for clinicians and is performed in a perceptive and objective way. On the one hand, the perceptive evaluation consists in qualifying and quantifying the voice disorder by listening to the production of a patient. This evaluation is performed by trained professionals who rate the phonation, e.g. using the Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS) scale [2]. This approach suffers from the dependency on the listener experience, as well as the inter- and intra-judges variability. On the other hand, the objective evaluation aims at qualifying and quantifying the voice disorder by acoustical, aerodynamic and physiological measures. Compared to methods based on electroglottography or high-speed imaging, the objective evaluation presents the advantage of being quantitative, cheaper, faster and more comfortable for the patient.

In order to provide objective tools to clinicians, a part of research in speech processing has focused on the detection of speech pathologies from audio recordings. Indeed it could be useful to detect disorders when the perturbations are still weak, to prevent the degradation of the pathology, or to quantify the voice quality before and after surgery in case of stronger disorders [3].

Traditional methods of voice pathology detection generally rely on the computation of acoustic features extracted from the speech signal. From another point of view, video recordings of the vocal folds show that the behaviour of the vocal folds is linked to the perception of different kinds of voice qualities, including pathologies (for instance, the incomplete closure of the folds can imply the perception of a *breathy* voice [4]). Isolating and parametrizing the glottal excitation should therefore lead to a better discrimination between normophonic and dysphonic voices. Such parametrizations of the glottal pulse have already been proposed both in time and frequency domains (see Chapter 4, or [5]). It is for example used to characterize different types of phonations [6] or to derive biomechanical parameters for voice pathology detection [3].

In addition several studies on speech perception, such as [7], have highlighted the importance of phase information. In this way, phase-based features have recently shown their efficiency in various fields of speech processing, such as automatic speaker [8] or speech [9] recognition. Among others, conclusions drawn in these works underline the complementarity of phase-based features with usual parameters extracted from the magnitude spectrum.

The goal of this chapter is two-fold. First of all, a set of new features, mainly based on the glottal source estimation, is extracted through pitch-synchronous analysis. These proposed features are compared according to their relevance for the problem of automatic voice pathology detection. For this, we make use of information theoretic measures. This is advantageous in that the approach is independent of any classifier and allows an intuitive interpretation in terms of discrimination power and redundancy between the features. Secondly, we investigate the potential of using features derived from the phase information for the same purpose, which, to the best of our knowledge, was never explored in the literature. It can be indeed expected that speech representations based on the phase convey relevant information for this task. The detection performance of these latter features is also evaluated, in addition to the information theoretic measures, using an Artificial Neural Network (ANN).

Experiments in this chapter are led on a popular database in the domain of speech pathologies: the MEEI Disordered Voice Database, produced by KayPentax Corp [10]. This database contains sustained vowels and reading text samples, from 53 subjects with normal voice and 657 subjects with a large panel of pathologies. Recordings are linked to information about the subjects (age, gender, smoking or not) and to analysis results. In this work, all the sustained vowels of the MEEI Database resampled at 16 kHz are considered.

This chapter is structured as follows. Section 8.2 provides a necessary background on the measures derived from the Information Theory that are used throughout this chapter. Their interpretation for a classification problem is also highlighted. Section 8.3 then investigates the complementarity of characteristics derived from the glottal and the speech signals. For this, the various features are described in Section 8.3.1, some of them being extracted from an estimation of the glottal source obtained by the Iterative Adaptative Inverse Filtering (IAIF) method (see Section 4.2.1 or [11]). Experiments and results using these features are detailed in Section 8.3.2. It is discussed which features are particularly informative for detecting a voice disorder and which ones are interestingly complementary or synergic. Section 8.4 focuses on the potential of phase-based features for voice disorder analysis. The proposed phase-based features are detailed in Section 8.4.1, and their efficiency is evaluated in Section 8.4.2. Finally Section 8.5 concludes.

8.2 Background on Information Theory-based Measures

The problem of automatic classification consists in finding a set of features X_i such that the uncertainty on the determination of classes C is reduced as much as possible [12]. For this, Information Theory [13] allows to assess the relevance of features for a given classification problem, by making use of the following measures (where $p(.)$ denotes a probability density function, and where c and x_i are a discretized version of variables C and X_i):

- The entropy of classes C is expressed as:

$$H(C) = -\sum_{c} p(c) \log_2 p(c) \tag{8.1}$$

  and can be interpreted as the amount of uncertainty on their determination.

- The mutual information between one feature X_i and classes C:

$$I(X_i; C) = \sum_{x_i} \sum_{c} p(x_i, c) \log_2 \frac{p(x_i, c)}{p(x_i)p(c)} \tag{8.2}$$

  can be viewed as the information the feature X_i conveys about the considered classification problem, i.e the discrimination power of one individual feature.

- The joint mutual information between two features X_i, X_j, and classes C can be expressed as (do not confuse the notation "," for a joint distribution and ";" used to separate random variables in a mutual information expression):

$$I(X_i, X_j; C) = I(X_i; C) + I(X_j; C) - I(X_i; X_j; C) \tag{8.3}$$

  and corresponds to the information that features X_i and X_j, when *used together*, bring to the classification problem. The last term can be written as:

$$I(X_i; X_j; C) = \sum_{x_i} \sum_{x_j} \sum_{c} p(x_i, x_j, c) \cdot \log_2 \frac{p(x_i, x_j)p(x_i, c)p(x_j, c)}{p(x_i, x_j, c)p(x_i)p(x_j)p(c)} \tag{8.4}$$

  An important remark has to be underlined about the sign of this term. It can be noticed from Equation (8.3) that a positive value of $I(X_i; X_j; C)$ implies some **redundancy** between the features, while a negative value means that features present some **synergy** (depending on whether their association brings less or more than the addition of their own individual information).

In order to evaluate the significance of the proposed features, the following measures are computed:

- the relative intrinsic information of one individual feature $\frac{I(X_i;C)}{H(C)}$, i.e the percentage of relevant information conveyed by the feature X_i,
- the relative redundancy between two features $\frac{I(X_i;X_j;C)}{H(C)}$, i.e the percentage of their common relevant information,
- the relative joint information of two features $\frac{I(X_i,X_j;C)}{H(C)}$, i.e the percentage of relevant information they convey together.

For this, Equations (8.1) to (8.4) are calculated. Probability density functions are estimated by a histogram approach. The number of bins was set to 50 for each feature dimension, which results in a trade-off between an adequately high number for an accurate estimation, while keeping sufficient samples per bin. Since features are extracted at the frame level, a total of around 32000 and 107000 pitch-synchronous examples is available in the MEEI database respectively for normal and pathological voices. Mutual information-based measures can then be considered as being accurately estimated. Class labels correspond to the presence or not of a dysphonia.

8.3 On the Complementarity of Glottal and Filter-based Features

This section focuses on the usefullness of glottal-based features for voice pathology detection. First of all, Section 8.3.1 describes the features considered in the following. These features are related to the glottal flow, to the vocal tract or to prosody. Relying on the measures derived from the Information Theory and which were presented in Section 8.2, the complementarity, redundancy and intrisic discrimination power of these features are discussed in Section 8.3.2.

8.3.1 Feature Extraction

The features considered in the present study characterize two signals. Some features are extracted from the speech signal, as in traditional methods of voice pathology detection. Others are extracted from the glottal source estimation in order to take into account the contribution of the glottis in the production of a disordered voice. All proposed features are extracted on pitch-synchronous frames in voiced parts of speech. For this, the pitch and voicing decision are computed using the Snack library [14] while Glottal Closure Instants (GCIs) are located on the speech signals using the DYPSA algorithm [15]. The choice of DYPSA was mainly motivated by the fact that it was the most well-known method of GCI detection at the time of these experiments. Given its low robustness (as studied in Section 3.6), it can be expected that the presence of a voice disorder will affect its performance, and will therefore be reflected in the resulting GCI-synchronous time-domain features (assuming that parameters extracted from the amplitude spectrum are relatively insensitive to an error of synchronization). In this work, glottal source-based frames are two pitch period-long while speech frames have a fixed length of 30 ms. In all cases, a GCI-centered Blackman weighting window is applied.

Speech signal-based features

It is generally considered that the spectrum of speech signal contains information about the presence of a pathology. The spectral content can be exploited using specific correlates between harmonics and formants although it must be noticed that these measures are dependent upon the phonetic context [3]. Another way for summarizing the spectral content is to compute characteristics describing its

spreading in energy. Most of the time these descriptors are designed to highlight the presence of noise. For instance, it is proposed in [16] to divide the [0-11kHz] frequency range into 5 frequency bands and to compute the ratio of energy between some pairs of bands and between each band and the whole frequency range.

In the present study, it has been chosen to follow a similar idea as [16], considering this time the perceptive mel scale, with the hope of being closer to human perception. The power spectral density is weighted by a mel-filterbank consisting of 24 triangular filters equally spaced along the whole mel scale. The Perceptive Energy (PE) associated to each filter and the spectral balances are defined as:

$$Bal1 = \frac{\sum_{i=1}^{4} PE(i)}{\sum_{i=1}^{24} PE(i)} \tag{8.5}$$

$$Bal2 = \frac{\sum_{i=5}^{12} PE(i)}{\sum_{i=1}^{24} PE(i)} \tag{8.6}$$

$$Bal3 = \frac{\sum_{i=13}^{24} PE(i)}{\sum_{i=1}^{24} PE(i)} \tag{8.7}$$

where $PE(i)$ denotes the perceptive energy computed in the i^{th} mel band. The global spreading of spectral energy can also be captured in the spectral centroid, also known as spectral center of gravity (CoG).

In addition, many studies attempt to quantify the presence of noise in the speech signal, as it is supposed to be linked to the perception of voiced disorders. Many parameters have been proposed to quantify the importance of the spectral noise, one of the most popular being the Harmonic-to-Noise Ratio (HNR) [17]. Basically this descriptor aims at quantifying the ratio between the energy of the harmonic and noise components of the spectrum. HNR is here computed using the Praat software [18] for comparison purpose. Besides we compute the so-called *maximum voiced frequency* Fm, as suggested in the Harmonic plus Noise Model (HNM, [19]). According to this model, the maximum voiced frequency dermacates the boundary between two distinct spectral bands, where respectively an harmonic and a stochastic modeling are supposed to hold. The higher Fm, the stronger the harmonicity, and consequently the weaker the presence of noise in speech. An example of Fm determination is displayed in Figure 8.1.

Glottal source-based features

In this part, the glottal source is estimated by the Iterative Adaptive Inverse Filtering (IAIF) method. IAIF is a popular approach proposed by Alku in [11] for improving the quality of the glottal flow estimation. IAIF is here used as it gave the best robustness performance in Section 6.3.1, which is a clear advantage for processing speech containing a voice disorder.

The amplitude spectrum for a voiced source generally presents a low-frequency resonance called *glottal formant*, produced during the glottal open phase (see Section 4.1 or [20]). The glottal formant frequency (Fg) and bandwidth (Bw) are consequently two important characteristics of the glottal signal. As shown in [21], as long as the applied windowing is GCI-centered, relatively sharp and two period-long, considering only the left-part of the window makes a good approximation of the glottal open phase, allowing an accurate estimation of Fg and Bw.

An example of glottal frames for both normal and pathological voices is given in Figure 8.2. It can be noticed that the discontinuity at the GCI is more significant for the normal voice. One possible way to quantify this is to take the minimum value at the GCI ($minGCI$) of frames normalized in energy and length. Apart from Fg, Bw and $minGCI$, spectral balances $Bal1$, $Bal2$, $Bal3$ and center of gravity CoG were also calculated on the glottal source frames.

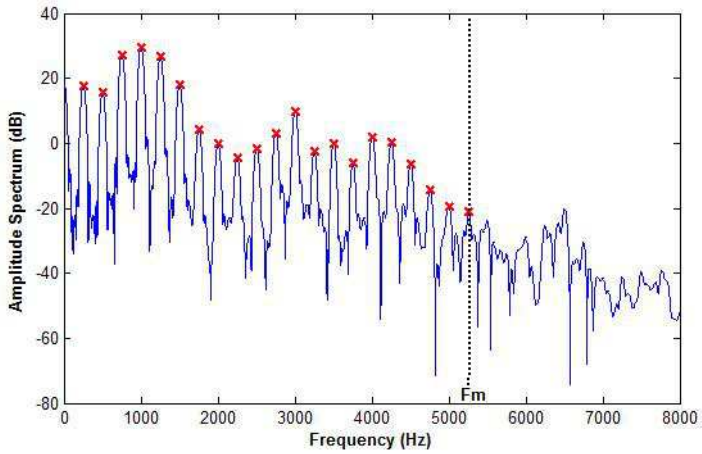

Figure 8.1 - *Example of maximum voiced frequency Fm determination on a frame of normal voice. Harmonics are indicated (x) and Fm corresponds to the last detected harmonic.*

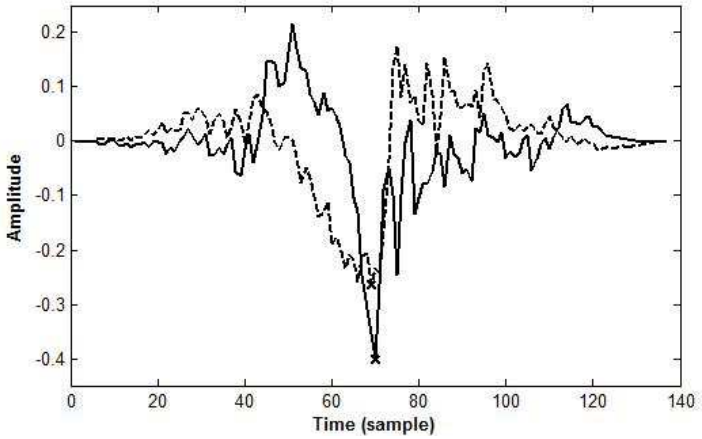

Figure 8.2 - *Comparison between a normal (solid) and pathological (dashed line) glottal frame. The minimum values around the GCI are indicated (x) and are an image of the discontinuity strength occuring at this moment.*

Prosodic features

It is often considered that dysphonic speakers have difficulty in maintaining stable prosodic characteristics during sustained vowels. These perturbations can be quantified by means of jitter and shimmer measures [22]. In the present study, the prosodic features are inspired by these measures. Indeed, for each frame, $DeltaF0$ and $DeltaE$ are respectively defined as the variation of pitch and energy around their respective median value calculated over the whole phonation of sustained vowels.

8.3.2 Results

The values of the measures derived from Information Theory (see Section 8.2) for the features proposed in Section 8.3.1 are presented in Table 8.1. The diagonal of this table indicates the percentage of relevant information conveyed by each feature. It turns out that features describing the speech spectrum contents are particularly informative (57% for $Bal1$), as well as Fm (41.4%) and $minGCI$ (47.7%).

The top-right part of Table 8.1 contains the values of relative joint information of two features, while the bottom-left part shows the relative redundancy between two features. Prosodic features convey few relevant information but are rather synergic with features from two other categories. Although features describing the speech spectrum are intrinsically relevant, they are fairly redundant between them (for example, $Bal1$ and $Bal2$ present 49.4% of redundancy). More interestingly it can be noted that they are relatively complementary to glottal-based features. In particular the association of $minGCI$ and $Bal1$ brings the most important joint mutual information (81.1%). Fig. 8.3 displays the joint distributions of these features for both normophonic and dysphonic voices. A clear separability between both classes can be observed. In addition it is confirmed that dysphonic speakers generally have difficulties in producing an abrupt glottal closure (leading to low values of $minGCI$).

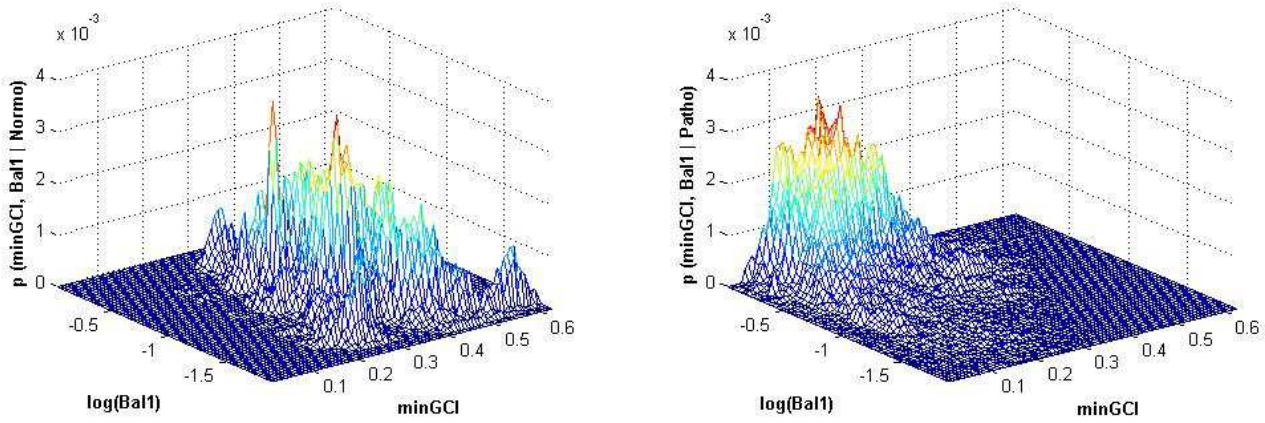

Figure 8.3 - *Example of separability for the two features giving the highest joint information.*

8.4 Using Phase-based Features for Detecting Voice Disorders

This section investigates the usefulness of phase-based features for the automatic detection of a voice pathology. Although not strictly motivated by any physiological process, as it is the case for the glottal flow, these features are related to the vocal chords behavior. It will be shown that they are particularly well suited for highlighting turbulences during the phonation process, and therefore for the discernment of a voice disorder. Section 8.4.1 first presents the phase-based features used in the remaining of this chapter. The potential of these features for voice pathology detection is then evaluated in Section 8.4.2.

8.4.1 Phase-based Features

This section presents the characteristics based on the speech signal phase information that are evaluated in Section 8.4.2. Analysis relying on the group delay function is first detailed. The respect of the mixed-phase model in both normo and dysphonic voices is then discussed.

Group Delay-based Analysis

The group delay function is defined as the derivative of the unwrapped phase spectrum. However, the group delay computed in this way contains spikes due to the presence of some zeros of the signal z-transform close to the unit circle (where the Fourier transform is evaluated). Therefore group

		Glottal source-based features							Speech signal-based features							
		F_g	B_w	$minGCI$	CoG	Bal_1	Bal_2	Bal_3	F_m	HNR	CoG	Bal_1	Bal_2	Bal_3	$Delta_E$	$Delta_{F_0}$
Glottal source-based features	F_g	25.7	46.6	58.8	39.9	40.8	41.3	39.6	57.4	49.5	68.0	73.3	69.3	46.4	36.1	32.2
	B_w	3.3	24.2	57.9	37.6	36.8	37.9	37.2	54.2	48.8	64.8	72.0	67.6	44.9	34.4	31.1
	$minGCI$	14.6	14.0	47.7	53.4	54.6	55.6	52.5	66.1	64.6	74.7	81.1	78.1	59.8	53.9	50.9
	CoG	8.9	9.7	17.3	23.1	30.8	31.4	29.5	51.4	47.5	64.8	70.4	65.0	43.8	33.7	27.8
	Bal_1	2.4	4.9	10.5	9.8	17.5	24.2	27.8	49.7	43.5	66.1	69.6	64.9	41.2	27.8	22.0
	Bal_2	2.0	3.8	9.5	9.2	10.9	17.5	29.1	48.8	42.4	67.0	71.2	66.5	40.4	27.0	21.6
	Bal_3	7.5	8.3	16.5	14.9	11.0	9.7	21.3	49.9	45.2	65.0	71.1	66.3	42.1	31.3	25.0
Speech signal-based features	F_m	9.7	11.3	22.9	13.1	9.2	10.1	12.8	41.4	49.2	71.4	77.0	72.2	56.5	48.2	44.8
	HNR	10.1	9.3	17.0	9.5	8.0	9.0	10.0	26.0	33.9	73.7	73.9	70.2	60.7	40.6	34.6
	CoG	6.2	7.9	21.5	6.8	0.0	-0.9	4.8	18.5	8.8	48.5	71.6	66.8	60.9	55.3	58.8
	Bal_1	9.4	9.2	23.6	9.7	5.0	3.4	7.3	21.4	17.1	34.0	57.0	63.0	70.6	64.9	66.9
	Bal_2	11.8	12.0	25.0	13.5	8.0	6.4	10.4	24.6	19.1	37.2	49.4	55.4	65.8	61.7	61.7
	Bal_3	-3.4	-3.4	5.1	-3.5	-6.4	-5.6	-3.5	2.2	-9.5	5.0	3.7	6.9	17.3	27.0	28.5
	$Delta_E$	-2.4	-2.2	1.7	-2.7	-2.3	-1.5	-2.0	1.1	1.3	1.3	0.1	1.7	-1.7	8.0	14.1
	$Delta_{F_0}$	-2.9	-3.3	0.7	-1.1	-0.9	-0.5	-0.1	0.2	2.9	-6.7	-6.3	-2.7	-7.6	-2.5	3.6

Table 8.1 - *Mutual information-based measures for the proposed features.* On the diagonal*: the relative intrinsic information* $\frac{I(X_i;C)}{H(C)}$. In the bottom-left part*: the relative redundancy between the two considered features* $\frac{I(X_i;X_j;C)}{H(C)}$. In the top-right part*: the relative joint information of the two considered features* $\frac{I(X_i,X_j;C)}{H(C)}$.

delay processing has been avoided for a long time. Nevertheless, some new representations aiming at reducing the effect of these spikes have been recently suggested, leading to some improvements in speech recognition, especially in noisy conditions [9], [23], [24]. This section explores the use of these new modes of representation for determining the presence of a voice disorder. For this, 5 types of spectrograms are considered in this work:

- The **Fourier Magnitude (FM) spectrogram** is the commonly adopted representation in speech processing, here introduced as a baseline.
- The **STRAIGHT spectrogram** is based on a restructuration of the speech representation by using a pitch-adaptive time-frequency smoothing of the FM spectrogram [25].
- The **Modified Group Delay (ModGD) spectrogram** is a function introduced by Hegde *et al.* in [9] using a cepstral smoothing in order to reduce the effect of the spikes on the group delay function.
- The **Product of the Power and Group Delay (PPGD) spectrogram** also aims at reducing the source of spikes, and is defined in [23] as the product of the power spectrum and the group delay function.
- Finally, the **Chirp Group Delay (CGD) spectrogram** is a representation proposed by Bozkurt *et al.* in [24] which relies on a chirp (i.e the Fourier transform is evaluated on a contour in the z-plane different from the unit circle) analysis of the zero-phase version of the speech signal. Note that this approach is completely different from the chirp mixed-phase separation presented in Chapter 7. The chirp technique (with a fixed predefined radius in this case) is here used to provide a high-resolved representation of the formant peaks, as initially suggested by Rabiner in [26].

In order to give an illustration, Figure 8.4 compares the five spectrograms for a segment of sustained vowel */a/* for two typical normophonic and dysphonic voices. It can be noticed from these plots that spectrograms of the normophonic voice present a regular structure in time, while their equivalents for the dysphonic speech contain time-varying irregularities. These are probably due to the difficulties of the patient suffering from a voice disorder in sustaining a regular phonation. Indeed, during the production of a sustained vowel, the vocal tract can be assumed as a stationary system on a short-time period, even for pathological voices. The speech signal then results from the excitation of this stationary system by the glottal source. As a consequence, a regular glottal flow will be characterized by a smooth spectrogram. On the opposite, if the resolution of the spectrogram is sufficiently high, turbulences or cycle-to-cycle variations present in the glottal flow will be reflected by irregularities in the spectrogram structure, as exhibited in the bottom plots of Fig. 8.4. It is also observed in Figure 8.4 that these irregularities are particularly well emphasized in the CGD spectrogram. This approach is indeed known for giving both a smooth and high-resolved representation for showing up resonance peaks in the speech spectrum [24].

Adequacy with the Mixed-Phase Model of Speech

The principle of mixed-phase separation has been deeply studied in Chapter 5. It has been shown, among others, that it allows an efficient estimation of the glottal open phase by isolating the maximum-phase component of speech. In this work, mixed-phase deconvolution is achieved using the Complex Cepstrum-based Decomposition (CCD) proposed in Section 5.3.2 (or [27]). If CCD is applied to a windowed segment of voiced speech exhibiting characteristics of the mixed-phase model, source-tract

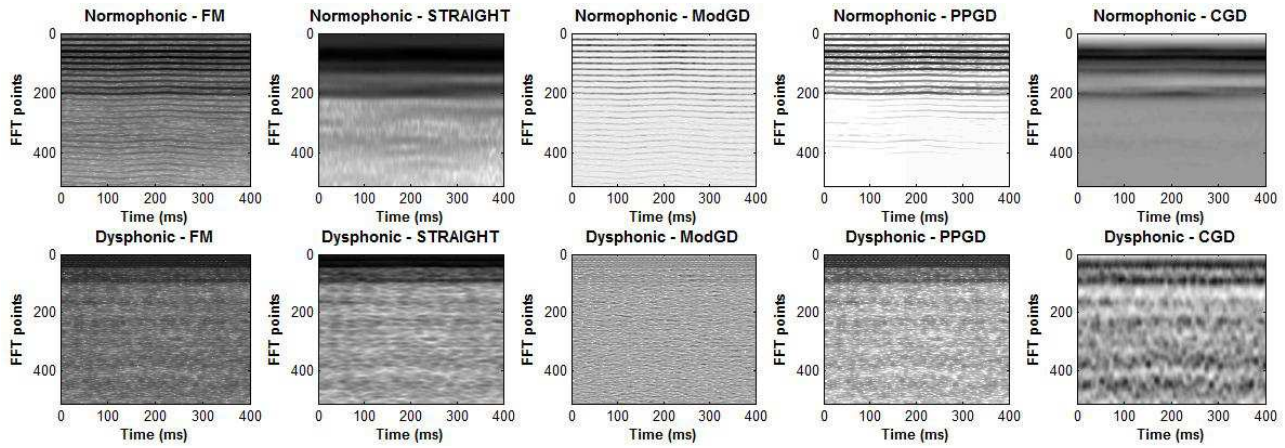

Figure 8.4 - *Illustration of the five types of spectrograms for a segment of sustained vowel /a/ for both normophonic (top plots) and dysphonic (bottom plots) voices. The spectrograms are related to the following representation (from left to right): the Fourier Magnitude (FM), the STRAIGHT, the Modified Group Delay (ModGD), the Product of the Power and Group Delay (PPGD), and the Chirp Group Delay (CGD) spectrograms.*

separation can be correctly carried out. Two cycles of the resulting anticausal component are displayed in Fig. 8.5(a), providing a reliable estimation of the glottal source (i.e corroborating the glottal flow models). If this is not the case, the decomposition fails and the resulting anticausal contribution has an irrelevant shape (as shown in Fig. 8.5(b)), generally characterized by a high-frequency noise.

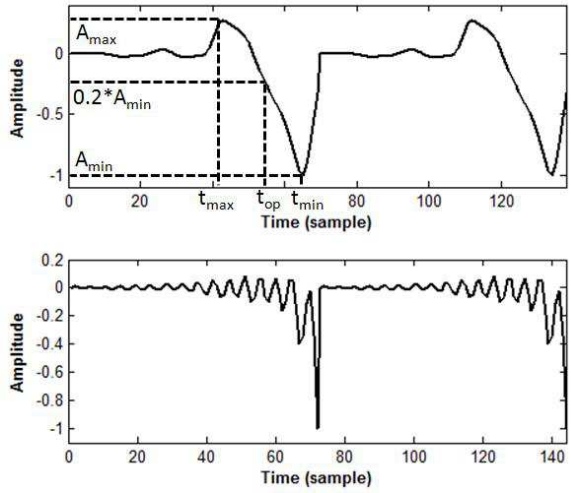

Figure 8.5 - *Two cycles of the anticausal component isolated by the mixed-phase decomposition* (top plot)*: when the speech segment exhibits characteristics of the mixed-phase model,* (bottom plot)*: when this is not the case. The particular instants defining the two proposed time constants are also indicated.*

In order to assess the quality of the mixed-phase separation, two time features are extracted from the maximum-phase signal. These two features characterize the glottal open phase response and are defined with the help of Figure 8.5. If T_0 denotes the pitch period, the first time constant T_1 is defined as $\frac{t_{min}-t_{max}}{T_0}$, while the second one T_2 is computed as $\frac{t_{min}-t_{op}}{T_0}$. These parameters should then contain relevant information about the consistency of the mixed-phase decomposition, i.e whether this

deconvolution leads to a reliable estimation of the glottal flow or not.

8.4.2 Evaluation of the Proposed Phase-based Features

Method Implementation

The five spectrograms defined in Section 8.4.1 are computed using Blackman-windowed frames shifted every $10ms$ and whose length is $30ms$. For this, the Fourier spectrum is estimated by a DFT of 1024 points. All other parameters are fixed to the values recommended in the corresponding references. For each spectrogram, the relative difference between two consecutive frames is calculated, since this feature should convey relevant information about glottal source irregularities. In the rest of the chapter, the prefix d ahead of the name of a spectrogram type will be used for denoting this feature. It is worth emphasizing that since we consider frame-to-frame variations, the resulting features are robust to differences between recording conditions or equipments, and among others to phase distortion.

The mixed-phase decomposition is achieved using the CCD algorithm proposed in [27]. Since this method requires a GCI-synchronous process, Glottal Closure Instants (GCIs) are located on the speech signals using the DYPSA algorithm [15], for the same reasons as in Section 8.3.1. Therefore the possible failure of GCI estimation for pathological voices is implicitly captured in the mixed-phase based features. For each resulting frame, one cycle of the anticausal component of speech is isolated and time constants T_1 and T_2 are extracted from it. In order to be synchronous with the other features, these streams are then interpolated at $100Hz$.

The three spectral balances Bal_1, Bal_2 and Bal_3 proposed in Section 8.3.1 are also extracted from the FM spectrogram. These balances were defined as the power spectral density in three perceptual subbands (see Equations (8.5) to (8.7)), and were shown in Table 8.1 to be highly discriminant for detecting a voice disorder.

Mutual Information-based Evaluation

In this part, the proposed features are compared according to their relevance for the problem of voice pathology detection. For this, we make use of information-theoretic measures described in Section 8.2. This is advantageous since the approach is independent of any classifier and allows an intuitive interpretation in terms of discrimination power. More precisely, the normalized Mutual Information (MI) of the proposed features is here studied. As a reminder, this measure is the relative intrinsic information of one individual feature $\frac{I(X_i;C)}{H(C)}$, i.e the percentage of relevant information conveyed by the feature X_i about the classification problem.

Table 8.2 presents the values of the normalized MI for the 10 features. As expected from the results of Section 8.3.2 , the two first spectral balances are strongly informative. From the various spectrogram representations, the Chirp Group Delay (CGD) provides the highest amount of relevant information (covering 56% of the total uncertainty). Although in a lesser extent, the Modified Group Delay (ModGD) and the two time constants characterizing the mixed-phase decomposition show an interesting potential for voice pathology detection. However, it is worth noting that the normalized MI is a measure of the intrinsic discrimination power of each feature separately, and consequently is not informative about the redundancy between them. The mutual information-based measures of redundancy are not exhaustively presented here as it was done in Table 8.1. Nevertheless, the main conclusions that can be drawn from them are the following. Albeit spectral balances have the highest normalized MI in Table 8.2, they are also highly redundant. In this way, the joint use of Bal_1 and Bal_2 only brougth 63.0% of MI. On the contrary, the best combination of two features is surprisingly

T_2 and Bal_1 (which is not straightforward at the only sight of Table 8.2), with 79.31% of MI. This is possible as these two features present a very low amount of redundancy.

Feature	dFM	dSTRAIGHT	dModGD	dPPGD	dCGD
Normalized MI (%)	22.32	16.32	30.56	15.43	55.97
Feature	T_1	T_2	Bal_1	Bal_2	Bal_3
Normalized MI (%)	32.02	23.09	57.02	55.38	17.29

Table 8.2 - *Values of the normalized mutual information for the 10 features.*

Classifier-based Evaluation

This section aims at evaluating the proposed features using a classifier for the automatic detection of voice disorders. For this, an Artificial Neural Network (ANN, [28]) is used for its discriminant learning capabilities. The ANN employed in this study has only one hidden layer and uses sigmoid as activation functions. The hidden layer is made of 16 neurons, as this gave the best trade-off between complexity and generalization capabilities in our attempts. Evaluation is achieved using a 10-fold cross validation framework. The system is finally assessed at both frame and patient levels (a patient is diagnosed as dysphonic if the majority of his frames are recognized as pathological).

Feature set name	**Features used**	**Error rate (%) (frame level)**	**Error rate (%) (patient level)**
Fourier Magnitude	dFM	17.2	8.73
Chirp Group Delay	dCGD	9.40	4.93
Mixed-phase model-based	T_1,T_2	13.28	5.35
Two best features	Bal_1,T_2	8.65	5.07
Spectral balances	Bal_1,Bal_2,Bal_3	9.97	7.89
Group Delay-based	dModGD,dPPGD,dCGD	7.92	4.08
Spectrogram-based	dFM,dSTRAIGHT, dModGD,dPPGD,dCGD	8.25	4.65
Phase-based	T_1,T_2, dModGD,dPPGD,dCGD	7.97	4.08
Whole feature set	All 10 features	6.16	4.08

Table 8.3 - *Results of voice pathology detection using an ANN classifier for various feature sets.*

Results we obtained are presented in Tab.8.3 for different feature sets. Several conclusions can be drawn from these results. First of all, the efficiency of dCGD is confirmed as this feature alone leads to only 4.93% of patients incorrectly classified. Its advantage over the traditional Fourier Magnitude (dFM) spectrum can be noted. In the experiments using only 2 or 3 parameters, the improvement brought by the proposed features compared to the spectral balances is clearly seen. Indeed the two time constants characterizing the respect of the mixed-phase model lead, at the patient level, to a better classification than when using the 3 spectral balances. In addition, the use of T_2 in combination with Bal_1 is more performant than the 3 spectral balances, confirming the discussion about redundancy in the previous subsection. It is worth noting the high performance achieved when using the 3 GD-based features. Adding the 2 magnitude-based spectrograms to these latter even leads to a slight degradation of accuracy. Similarly, it can be observed that adding the two mixed-phase model-based time constants

makes the performance almost unchanged. Finally, considering all 10 features correctly identifies 93.84% of frames and 95.92% of patients. However, it is interesting to notice that the best performance at the patient level was already achieved using only the 3 GD-based representations. Although not reported in Tab.8.3, it is worth noting that for these 3 latter features, the rates of false positive and negative patients are respectively of 16.98% and 3.04%. The high rate of false positive detections can be explained by the unbalance of the MEEI database, leading to an overestimation of pathologies. Nonetheless, relying on a Receiver Operating Characteristic (ROC) curve, one could modify these latter rates by playing on the posterior threshold for deciding whether a frame is pathological or not (i.e a frame could be pathological with a probability greater or lower than 0.5).

8.5 Conclusion

This chapter focused on the problem of automatic detection of voice pathologies from the speech signal. The use of the glottal source estimation was investigated and a set of new features was proposed. The resulting extracted features were assessed through mutual information-based measures. This allowed their interpretation in terms of discrimination power and redundancy. It turned out that features describing the speech spectrum contents are particularly informative, as well as the maximum voiced frequency and the glottal discontinuity at the GCI. It was also shown that speech and glottal-based features are relatively complementary, while they present some synergy with prosodic characteristics. The potential of using phase-based features for detecting voice pathologies was also explored. It was shown that representations based on group delay functions are particularly suited for capturing irregularities in the speech signal. The adequacy of the mixed-phase model during the voice production was discussed and shown to convey relevant information. Besides it was underlined that phase-based and magnitude spectrum-based features may present interesting complementarity for this task, showing among others a very weak redundancy. The efficiency of these phase-based features may be explained by their higher sensitivity to turbulences during the phonation process.

Bibliography

[1] J. Stemple, L. Glaze, and B. Klaben. Clinical voice pathology: Theory and management. *Singular Editions, 3rd Edition*, 2000.

[2] M. Hirano. Psycho-acoustic evaluation of voice: Grbas scale for evaluating the hoarse voice. *Springer-Verlag*, 1981.

[3] P. Gomez-Vilda, R. Fernandez, V. Rodellar, V. Nieto, A. Alvarez, L.M. Mazaira, R. Martinez, and J. Godino. Glottal source biometrical signature for voice pathology detection. *Speech Communication*, 51(9):759–781, 2009.

[4] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–586, 1999.

[5] P. Alku. Parametrisation methods of the glottal flow estimated by inverse filtering. In *Proc. of VOQUAL' 03*, pages 81–87, 2003.

[6] M. Airas and P. Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. Interspeech*, pages 1410–1413, 2007.

[7] B. Atal K. Paliwal. Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Trans. Speech Audio Processing*, 1:3–14, 1993.

[8] S. Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. In *IEEE SigPro Lett.*, volume 13, pages 52–55, 2006.

[9] R. Hegde, H. Murthy, and V. Gadde. Continuous speech recognition using joint features derived from the modified group delay function and mfcc. In *Proc. ICSLP*, 2004.

[10] Kay Elemetrics Corp. Disordered voice database model (version 1.03). In *Massachussetts Eye and Ear Infirmary Voice and Speech Lab*, 1994.

[11] P. Alku, J. Svec, E. Vilkman, and F. Sram. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[12] L. Huan and H. Motoda. Feature selection for knowledge discovery and data mining. *The Springer International Series in Engineering and Computer Science*, 454, 1998.

[13] T. Cover and J. Thomas. Elements of information theory. *Wiley Series in Telecommunications, New York*, 1991.

[14] Online. The snack sound toolkit. In *http://www.speech.kth.se/snack/*.

[15] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Trans. Speech Audio Process.*, 15(1):34–43, 2007.

[16] J.B. Alonso, J. de Leon, I. Alonso, and A.M. Ferrer. Automatic detection of pathologies in the voice by hos based parameters. *EURASIP Journal on Applied Signal Proceesing*, 4:275–284, 2001.

[17] K. Shama, A. Krishna, and N. Cholayya. Study of hrmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on Applied Signal Proceesing*, 2007.

[18] P. Boersma and D. Weenik. Praat: doing phonetics by computer (version 5.1.03). *[Computer Program]*, 2009.

[19] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9:21–29, 2001.

[20] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

[21] T. Drugman, T. Dubuisson, A. Moinet, N. D'Alessandro, and T. Dutoit. Glottal source estimation robustness. In *IEEE Int. Conf. on Signal Processing and Multimedia Applications*, 2008.

[22] D. Michaelis, M. Fröhlich, and H. Strube. Selection and combination of acoustic features for the description of pathological voices. *J. Acoust. Soc. Am.*, 103:1628–1639, 1998.

[23] D. Zhu and K. Paliwal. Product of power spectrum and group delay function for speech recognition. In *Proc. ICASSP*, 2004.

[24] B. Bozkurt, L. Couvreur, and T. Dutoit. Chirp group delay analysis of speech signals. In *Speech Comm.*, volume 49, pages 159–176, 2007.

[25] H. Kawahara, P. Zolfaghari, and A. de Cheveigne. On f0 trajectory for very high-quality speech manipulation. In *Proc. ICSLP 2002*, 2002.

[26] L. Rabiner, R. Schafer, and C. Rader. The chirp-z transform algorithm and its application. *Bell System Technical Journal*, 48(5):1249–1292, 1969.

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

[28] N. Karayiannis and A. Venetsanopoulos. *Artificial Neural Networks - Learning Algorithms, Performance Evaluation, and Applications*. Kluwer Academic Publishers, 1993.

Chapter 9

Glottal-based Analysis of Expressive Speech

Contents

Abstract

More and more efforts in speech processing are nowadays devoted to the analysis and synthesis of expressive speech (i.e speech revealing affect). This interest is motivated by a need in more and more natural human-machine interactions. The focus of research has then shifted from *read speech* to *conversational* styles of speech. This chapter aims at studying the glottal-based modifications of two particular types of expressive speech: Lombard speech and hypo or hyperarticulated speech. The Lombard effect refers to the speech changes due to the immersion of the speaker in a noisy environment. These are (generally unconsciously) intended so as to maximize the intelligibility of the delivered message. We analyze how the glottal behaviour is altered in Lombard speech, as a function of the type and level of the surrounding noise. As a second specific mode of expressivity, hypo and hyperarticulation are studied. Hyperarticulated speech refers to the voice for which clarity tends to be maximized, while hypoarticulation results from a production with minimal efforts. We investigate how characteristics of both the vocal tract and the glottis are affected.

This chapter is based upon the following publications:

- Thomas Drugman, Thierry Dutoit, *Glottal-based Analysis of the Lombard Effect*, Interspeech Conference, Makuhari, Japan, 2010.
- Benjamin Picart, Thomas Drugman, Thierry Dutoit, *Analysis and Synthesis of Hypo and Hyperarticulated Speech*, 7th ISCA Speech Synthesis Workshop, Kyoto, Japan, 2010.

Many thanks to Benjamin Picart for his collaboration on the experiments of Section 9.3.

9.1 Introduction

Arising from a need in more natural interactions with the machine, a part of the speech processing community has focused on the analysis and synthesis of expressive speech. A gain of interest has shifted from *read speech* to *conversational* styles of speech. Indeed, in real human-machine interactions, there is need for more than just the intelligible portrayal of linguistic information; there is also a need for the expression of affect [1]. One has then to deal with all subtleties conveyed in the extralinguistic and paralinguistic information, besides the usual phonetic content of read speech.

In the previous chapters (see for example Section 6.4), it was already shown on a large corpus of expressive speech (the De7 database) that the glottal source is informative about the produced voice quality (modal, soft or loud). A study of glottal modifications led on a large corpus of expressive speech (e.g sad, dictated or narrative speech) was recently carried out by Sturmel in [2]. This chapter now focuses on the analysis of two other modes of expressivity: Lombard speech (Section 9.2) and hypo or hyperarticulated speech (Section 9.3). Although acoustic changes also concern prosodic or vocal tract-based features, we here focus on the glottal modifications in expressive speech, emphasizing the usefullness of the methods described in Part II. Albeit the glottal behaviour was expected to change in expressive speech, the automatic analysis of such voices on large corpora (i.e not limited to a few sustained vowels) is rather limited in the literature. Results presented in the following were possible thanks to the development of efficient methods of glottal flow estimation and parametrization, as studied in the previous chapters of Part II. Finally Section 9.4 concludes the chapter.

9.2 Glottal-based Analysis of Lombard Speech

9.2.1 The Lombard Effect

The Lombard effect, as originally highlighted by Dr. Lombard in 1909 [3], refers to the speech changes due to the immersion of the speaker in a noisy environment. In such a context, the speaker tends (generally unconsciously) to modify its way of uttering so as to maximize the intelligibility of its message [4]. On a physiological point of view, an hyper-articulation is observed when the subject speaks in noisy conditions, which is reflected by an amplification of the articulatory movements [5]. As a consequence the Lombard effect encompasses a set of acoustic and phonetic modifications in the speech signal. These modifications affect the efficiency of speech processing systems and have to be compensated for an optimal performance. Among these systems, the compensation of the Lombard effect in speech recognition [6], [7] and speaker recognition [8] have already been studied.

The analysis of Lombard speech has been studied in several works ([4], [9], [7], [5], [8]). In these studies, the acoustic and phonetic features that were inspected include the vocal intensity, phoneme durations, the fundamental frequency, the spectral tilt, and the formant frequencies. In this way, the Lombard effect is known to result in an increased vocal intensity and fundamental frequency. Duration of vowels and semi-vowels was shown to increase with the noise level, while the duration of consonants was observed to be shorter. Regarding the spectral contents, the proportion of high frequencies is more important in Lombard speech, when compared to the neutral condition [5]. This is reflected by a weaker average spectral tilt as the noise increases [4], [8]. Finally the formant frequencies were observed to be reorganized in the $F1 - F2$ plane [4], [5]. While $F1$ was shown to increase in noisy conditions, no general rule were noticed for $F2$. In any of these studies, modifications were observed to be dependent on the noise type and level, as well as the considered speaker who may adapt its speaking style more or less strongly.

Among all these works, no one reported studies based on features related to the glottal flow (at the exception of the widely used fundamental frequency). However, it is expected that, during the

production of speech in noise, the vocal folds work in a way different from their normal behaviour in silent conditions. Indeed, significant differences in the glottal source have already been observed between various phonation types [10]. To the best of our knowledge, only one study investigated the information extracted from the excitation for analyzing Lombard speech [11]. In that study, authors inspected the changes present in two signals: 1) a zero-frequency filtered signal, 2) the LP residual signal. Although these two signals are informative about the excitation of the vocal tract, they do not correspond to the actual glottal flow produced by the vocal folds.

This section focuses on the analysis of the Lombard speech based on features extracted from the glottal flow. This signal corresponds to the airflow arising from the trachea and modulated by the vocal folds, and is then motivated by physiological considerations. In Section 9.2.2 the methods for glottal flow estimation and parametrization are described. Section 9.2.3 then presents the results of our experiments led on a large database containing an important number of speakers, noise types and levels.

9.2.2 Glottal Flow Estimation and Characterization

In this section, the glottal flow is estimated using the Closed Phase Inverse Filtering (CPIF) technique described in Section 6.2.1. Our choice for CPIF is motivated by the fact that it gave in Chapter 6 good results on real speech, and that it showed interesting robustness properties which are required here since recordings used for our experiments in Section 9.2.3 are not of very high quality.

As a reminder (see Section 6.2.1), this method is based on a Discrete All Pole (DAP, [12]) inverse filtering process estimated during the closed phase. The closed phase period is determined using the Glottal Opening and Closure Instants (GOIs and GCIs) located by the SEDREAMS algorithm detailed in Chapter 3 (or [13]). For high-pitched voices, two analysis windows were used as suggested in [14], [15] and [16]. As speech signals sampled at 16 kHz are considered in the following, the order for DAP analysis is fixed to 18 ($=F_s/1000+2$, as commonly used in the literature).

Once the glottal flow has been estimated, each glottal cycle is characterized by the following features: the Normalized Amplitude Quotient (NAQ) and the Quasi Open Quotient (QOQ) in the time domain; and the H1-H2 ratio and the Harmonic Richness Factor (HRF) in the spectral domain. These four parameters have been described in Section 4.3. In this study, these features were extracted with the TKK Aparat toolkit freely available in [17]. Besides, since the fundamental frequency has been extensively used in the literature ([4], [7], [8]), pitch is estimated using the Snack Sound Toolkit [18].

Finally, as a last feature related to the glottal behaviour, an averaged spectrum is computed for characterizing the utterances of a speaker in given recording conditions (i.e for a given noise type and level). Note that the use of a long-term averaged spectrum for characterizing speech is not a new idea [19]. In this work, this is achieved in a way inspired from the technique described in [8]. Voiced regions of speech are isolated and nasal segments are removed. For each resulting frame, the amplitude spectrum is computed. Periodograms are then averaged. This averaged magnitude spectrum then contains a mix of the average glottal and vocal tract contributions. If the dataset is sufficiently large and phonetically balanced, formants tend in average to cancel each other. An example of averaged spectrum for a male speaker and for three recording conditions is exhibited in Figure 9.1. Since these spectra were computed for the same speaker, it is reasonable to think that the main difference between them is due to the spectral tilt of the glottal flow regarding the phonation mode. In this way, it can be noticed from Figure 9.1 that the high-frequency contents becomes more important as the noise level increases. This confirms the conclusions about the spectral tilt drawn in [4] and [8]. In order to characterize the content of the averaged spectrum $S(f)$, the following ratios of energy are defined:

$$E_{2-1} = \frac{\int_{1000}^{3000} |S(f)|^2 df}{\int_{0}^{1000} |S(f)|^2 df}, E_{3-1} = \frac{\int_{3000}^{8000} |S(f)|^2 df}{\int_{0}^{1000} |S(f)|^2 df}.$$

The division according to these 3 subbands arises from the observation of Figure 9.1 where the distinction between these 3 spectral regions is well marked.

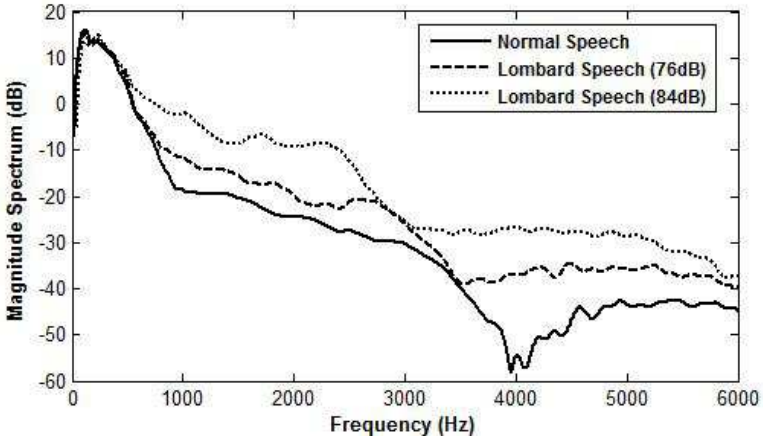

Figure 9.1 - *Averaged spectrum for a male speaker uttering in a silent environment, or with a factory noise of 76 dB and 84 dB.*

9.2.3 Experiments

Database

The database used in this study was first designed by the Multitel research center in order to develop robust speech recognition systems[1]. It consists of speech uttered by 25 speakers (11 females and 14 males). For recordings in clean conditions, the dataset consists of about 350 phonetically balanced sentences, and 57 sequences of words and numbers. For Lombard speech, four types of noise (car, crowd, factory and pop music noises) with two levels (76 and 84 dB-SPL (Sound Pressure Level)) were used. For the noisy conditions, only the 57 sequences of words and numbers were recorded. The speech signals were captured by a close-talk microphone and sampled at 16kHz.

Results

For all recordings, the glottal flow is estimated and characterized as described in Section 9.2.2. Among these parameters, the fundamental frequency F_0 has been extensively used in the literature. An example of F_0 distribution for a given male speaker is displayed in Figure 9.2 for both normal and Lombard speech. A clear increase of pitch is noticed as the noise level becomes stronger. This observation corroborates the conclusions drawn in [4], [7] or [8].

Regarding the features characterizing the glottal waveform both in time and frequency domains, Figure 9.3 exhibits their histograms for the same male speaker. Also maybe less marked than for the F_0 distribution with this speaker, significant differences in the histograms of the glottal features can be nevertheless observed. In this way, Lombard speech is characterized by a clear drop of NAQ, QOQ and $H1-H2$ parameters, while the Harmonic Richness Factor HRF is increased. These modifications

[1]Many thanks to Multitel ASBL for providing the database.

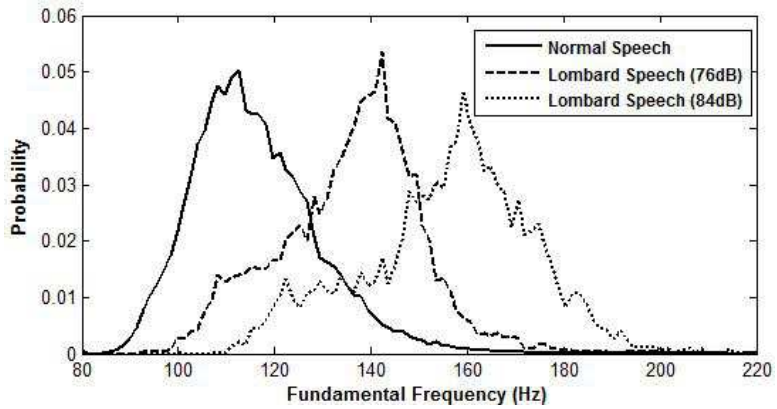

Figure 9.2 - *Pitch distribution for a given male speaker uttering in clean and noisy conditions. For this example, a factory noise at 76 and 84 dB was used.*

are mainly due to the stronger vocal effort in Lombard speech, and are in line with the study of the pressed phonation type ([20], [21]).

According to the evolution of the spectral features ($H1-H2$ and HRF), the content of the glottal spectrum is shown to present more high-frequency energy in Lombard speech. Indeed, in Lombard speech, the amplitude levels between the two first glottal harmonics becomes less important, and the amount of harmonics in the whole glottal spectrum gets richer. On the other hand, the evolution of the time-domain features NAQ and QOQ is difficult to interpret intuitively. To give an idea of their impact, Figure 9.4 displays the glottal open phase according to the LF model [22], for normal and Lombard speech, taking the mean values of NAQ and QOQ from Figure 9.3(a) and (b). Indeed these two parameters are known to control the shape of the glottal open phase. Differences in the glottal waveforms are observed in Figure 9.4, mainly in the rapidity of the open phase time response.

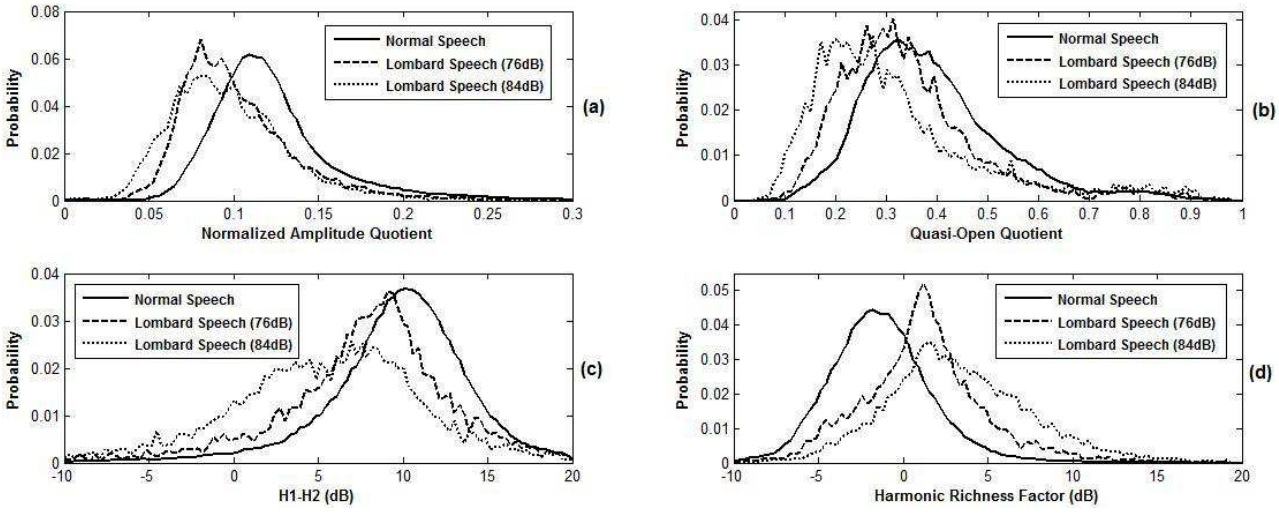

Figure 9.3 - *Distributions, for a given male speaker uttering in a quiet environment or in noisy conditions (with a factory noise at 76 and 84 dB), of the following glottal features:* (a): *the Normalized Amplitude Quotient NAQ,* (b): *the Quasi-Open Quotient QOQ,* (c): *the ratio of the amplitudes at the two first harmonics $H1-H2$,* (d): *the Harmonic Richness Factor HRF.*

Feature	**Normal**	Car76	Car84	Crowd76	Crowd84	Factory76	Factory84	Music76	Music84
NAQ	**0.131**	−15.9%	−23.7%	−14.7%	−22.5%	−20.2%	−26.4%	−5.8%	−15.8%
QOQ	**0.411**	−7.5%	−10.5%	−4.7%	−10.8%	−9.0%	−12.6%	−2.1%	−6.6%
H1H2	**9.45 dB**	−1.8 dB	−2.3 dB	−1.8 dB	−2.5 dB	−1.9 dB	−2.9 dB	−0.6 dB	−1.1 dB
HRF	**-1.72 dB**	+1.7 dB	+3.0 dB	+1.9 dB	+3.3 dB	+2.9 dB	+4.1 dB	+1.5 dB	+2.6 dB
F0	**164.7 Hz**	+9.8%	+25.7%	+20.8%	+25.8%	+13.7%	+29.4%	+25.2%	+31.1%
E_{2-1}	**-22.62 dB**	+8.0 dB	+11.5 dB	+8.1 dB	+11.0 dB	+9.4 dB	+12.8 dB	+10.3 dB	+12.6 dB
E_{3-1}	**-28.34 dB**	+4.6 dB	+6.8 dB	+3.7 dB	+4.3 dB	+7.8 dB	+10.7 dB	+8.3 dB	+9.3 dB

Table 9.1 - *Quantitative summary of the glottal modifications in Lombard speech. The reference values of the glottal features are given for the normal speech. Their relative modifications in Lombard speech are detailed for the 4 noise types at 76 and 84 dB.*

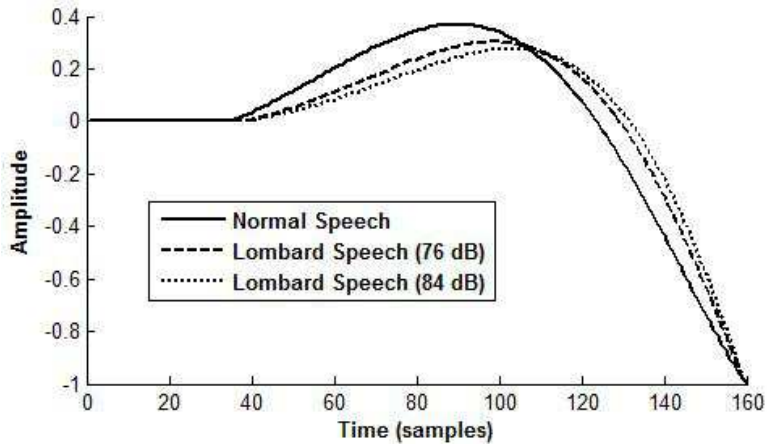

Figure 9.4 - *Illustration of the differences in the glottal open phase according to the LF model for normal and Lombard speech.*

Table 9.1 sums up the modifications of glottal features when speech is produced in silent or noisy environments. These results are averaged for the 25 speakers of the database and detailed in the table according to the noise type and level. Unformally we observed that all speakers tend to modify their glottal source in the same fashion, although these changes were more important for some speakers than for others. Regarding the features extracted from the glottal flow, it turns out that the noise types leading to the strongest modifications are (by order of increasing changes): the music, crowd, car and factory noises. Besides important variations of NAQ from normal to Lombard speech can be noted (up to 26% in the factory noise at $84dB$). As expected ([4], [7], [8]), it is also observed that speakers tend to increase F_0 in Lombard speech. Finally, regarding the spectral balances E_{2-1} and E_{3-1} defined as in Section 9.2.2, it can be concluded that speakers produce a higher amount of high-frequency in Lombard speech, confirming the results from [4], [8] and [5]. Among others, the energy in the $[1kHz - 3kHz]$ is particularly increased. One possible reason for this is that speakers (maybe unconsciously via their own auditory feedback) aim at enhancing their intelligibility by increasing the SNR where the human ear is the most sensitive.

9.3 Analysis of Hypo and Hyperarticulated Speech

9.3.1 Hypo and Hyperarticulated Speech

This section focuses on the study of different speech styles, based on the degree of articulation: neutral speech, hypoarticulated (or casual) and hyperarticulated speech (or clear speech). It is worth noting that these three modes of expressivity are neutral on the emotional point of view, but can vary amongst speakers, as reported in [23]. The influence of emotion on the articulation degree has been studied in [24], [25] and is out of the scope of this section.

The "H and H" theory [26] proposes two degrees of articulation of speech: hyperarticulated speech, for which speech clarity tends to be maximized, and hypoarticulated speech, where the speech signal is produced with minimal efforts. Therefore the degree of articulation provides information on the motivation/personality of the speaker vs the listeners [23]. Speakers can adopt a speaking style that allows them to be understood more easily in difficult communication situations.

The degree of articulation is influenced by the phonetic context, the speech rate and the spectral dynamics (vocal tract rate of change, [23]). The common measure of the degree of articulation consists

in defining formant targets for each phone, taking coarticulation into account, and studying the differences between the real observations and the targets versus the speech rate. Because defining formant targets is not an easy task, Beller proposed in [23] a statistical measure of the degree of articulation by studying the joint evolution of the vocalic triangle area and the speech rate.

The goal of this study is to have a better understanding of the specific acoustic characteristics governing hypo and hyperarticulated speech. Section 9.3.2 presents the database which has been created for this study. Modifications are studied in Section 9.3.3 as a function of the degree of articulation. The acoustic analysis highlights evidence of both vocal tract and glottal characteristics changes. Note that a more complete study including a phonetic analysis and the integration of these changes in speech synthesis based on Hidden Markov Models (HMMs) is given in [27].

9.3.2 Database with various Degrees of Articulation

For the purpose of our research, a new French database was recorded by a professional male speaker, aged 25 and native Belgian French speaking. The database contains three separate sets, each set corresponding to one degree of articulation (neutral, hypo and hyperarticulated). For each set, the speaker was asked to pronounce the same 1359 phonetically balanced sentences, as neutrally as possible from the emotional point of view. A headset was provided to the speaker for both hypo and hyperarticulated recordings, in order to induce him to speak naturally while modifying his articulation degree.

While recording hyperarticulated speech, the speaker was listening to a version of his voice modified by a "Cathedral" effect. This effect produces a lot of reverberations (as in a real cathedral), forcing the speaker to talk slowlier and as clearly as possible (producing more efforts to produce speech). In contrast, while recording hypoarticulated speech, the speaker was listening to an amplified version of his own voice. This effect produced the impression of talking very close to someone in a narrow environment, allowing the speaker to talk faster and less clearly (making less efforts to produce speech). Proceeding that way allowed us to create a "standard recording protocol" to obtain repeatable conditions if required in the future. It also avoided the data from being dependent on some subjective understanding of what *"hyper"* and *"hypo"* articulation actually is.

9.3.3 Acoustic Analysis of Hypo and Hyperarticulated Speech

Acoustic modifications in expressive speech have been extensively studied in the literature [28], [29], [30]. In the frame of this study, one can expect important changes related to the vocal tract function. Indeed, during the production of hypo and hyperarticulated speech, the articulatory strategy adopted by the speaker may dramatically vary. Although it is still not clear whether these modifications consist of a reorganization of the articulatory movements, or of a reduction/amplification of the normal ones, speakers generally tend to consistently change their way of articulating. According to the "H and H" theory [26], speakers minimize their articulatory trajectories in hypoarticulated speech, resulting in a low intelligibility, while an opposite strategy is adopted in hyperarticulated speech. As a consequence, the vocal tract configurations may be strongly affected. The resulting changes are studied in Section 9.3.3.

In addition, the produced voice quality is also altered. Since voice quality variations are mainly considered to be controlled by the glottal source [30], Section 9.3.3 focuses on the modifications of glottal characteristics with regard to the degree of articulation.

Vocal Tract-based Modifications

In order to study the variations of the vocal tract resonances, the evolution of the vocalic triangle [23] with the degree of articulation is analyzed. This triangle consists of the three vowels /*a*/, /*i*/ and /*u*/ represented in the space of the two first formant frequencies $F1$ and $F2$ (here estimated via Wavesurfer [31]). For the three degrees of articulation, the vocalic triangle is displayed in Figure 9.5 for the original sentences. For information, ellipses of dispersion are also indicated on these plots. The first main conclusion is the significant reduction of the vocalic space as speech becomes less articulated. Indeed, as the articulatory trajectories are less marked, the resulting acoustic targets are less separated in the vocalic space. This may partially explain the lowest intelligibility in hypoarticulated speech. On the contrary, the enhanced acoustic contrast is the result of the efforts of the speaker under hyperarticulation. These changes of vocalic space are summarized in Table 9.2, which presents the area defined by the average vocalic triangles.

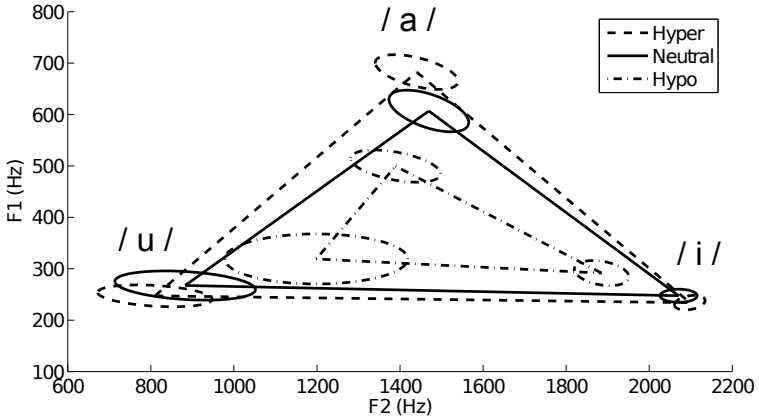

Figure 9.5 - *Vocalic triangle, for the three degrees of articulation. Dispersion ellipses are also indicated.*

Dataset	Hyper	Neutral	Hypo
Original	0.285	0.208	0.065

Table 9.2 - *Vocalic space (in kHz^2) for the three degrees of articulation.*

Inspecting the ellipses, it is observed that dispersion can be high for the vowel /*u*/, while data is relatively well concentrated for /*a*/ and /*i*/.

Glottal-based Modifications

As the most important perceptual glottal feature, pitch histograms are displayed in Figure 9.6. It is clearly noted that the more speech is articulated, the higher the fundamental frequency. Besides these prosodic modifications, we investigate how characteristics of the glottal flow are affected. In a first part, the glottal source is estimated by the Complex Cepstrum-based Decomposition algorithm (CCD) presented in Section 5.3.2 (or [32]). This method was shown in Chapter 5 to be the best performing technique of glottal flow estimation on high-quality recordings. Using this approach, Figure 9.7 shows the averaged magnitude spectrum of the glottal source for the three degrees of articulation. First of all, a strong similarity of these spectra with models of the glottal source (such as the LF model [22])

can be noticed. Secondly it turns out that a high degree of articulation is reflected by a glottal flow containing a greater amount of high frequencies. Finally, it is also observed that the glottal formant frequency increases with the degree of articulation (see the zoom in the top right corner of Figure 9.7). In other words, the time response of the glottis open phase turns to be faster in hyperarticulated speech.

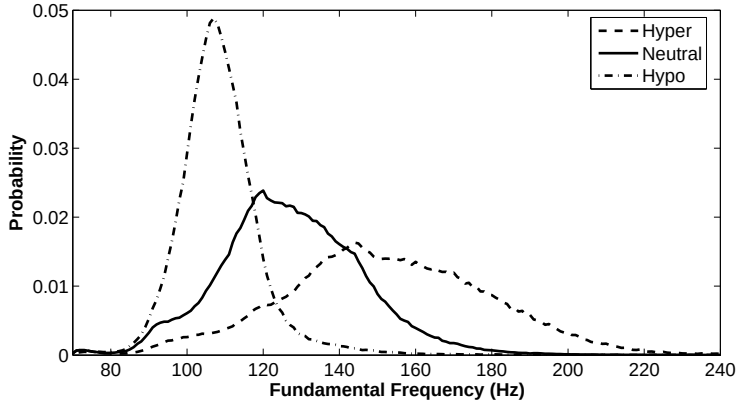

Figure 9.6 - *Pitch histograms for the three degrees of articulation.*

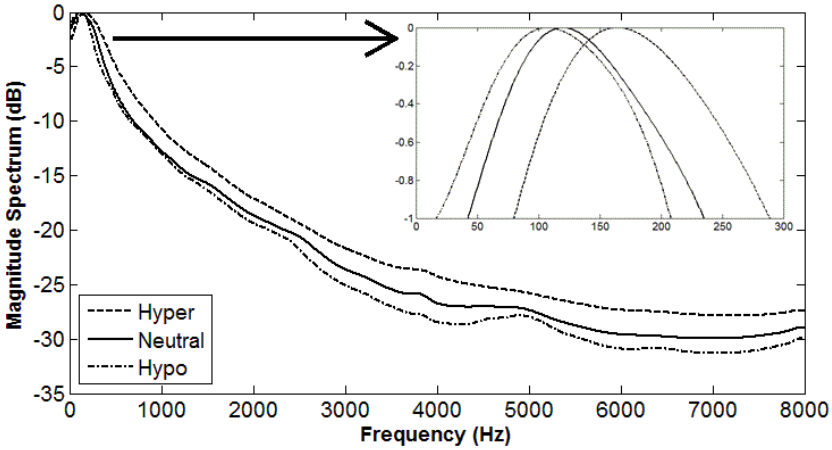

Figure 9.7 - *Averaged magnitude spectrum of the glottal source for the three degrees of articulation.*

In a second part, the maximum voiced frequency is analyzed. In some approaches, such as the Harmonic plus Noise Model (HNM, [33]) or the Deterministic plus Stochastic Model of the residual signal (DSM, see Chapter 11 or [34]), the speech signal is considered to be modeled by a non-periodic component beyond a given frequency. This maximum voiced frequency (F_m) demarcates the boundary between two distinct spectral bands, where respectively an harmonic and a stochastic modeling (related to the turbulences of the glottal airflow) are supposed to hold. In this paper, F_m was estimated using the algorithm described in [33]. The corresponding histograms are illustrated in Figure 9.8 for the three degrees of articulation. It can be noticed from this figure that the more speech is articulated, the higher the F_m, the stronger the harmonicity, and consequently the weaker the presence of noise in speech. Note that the average values of F_m are respectively of 4215 Hz, 3950 Hz (confirming our choice of 4 kHz in [34]) and 3810 Hz for the three degrees of articulation.

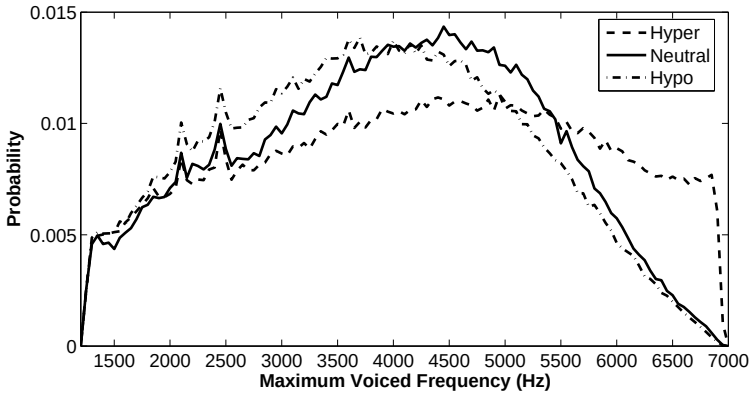

Figure 9.8 - *Histograms of the maximum voiced frequency for the three degrees of articulation.*

9.4 Conclusion

The goal of this chapter was to confirm and quantify, on large corpora, how the glottal source is modified during the production of expressive speech. First, we focused on the glottal analysis of Lombard speech. For this, the glottal flow was estimated by a closed phase inverse filtering process and characterized by a set of time and spectral features. Through an analysis on a database containing 25 speakers talking in quiet and noisy environments (with 4 noise types at 2 levels), it was shown that the glottal source is considerably modified in Lombard speech. These variations have to be taken into account in applications such as speech or speaker recognition systems. Moreover the results presented in this study could be turned into advantage by integrating them in a parametric speech synthesizer based on a source-filter model. It is indeed expected that this approach should enhance the delivered intelligibility by adapting the voice quality. In a second time, speech with various degrees of articulation has been studied. A new French database matching our needs was created, composed of three identical sets, pronounced with three different degrees of articulation (neutral, hypo and hyperarticulated speech). The acoustic analysis investigated changes related to the vocal tract as well as to the glottis. It was shown that hyperarticulated speech is characterized by a larger vocalic space (more efforts to produce speech, with maximum clarity), higher fundamental frequency, a glottal flow containing a greater amount of high frequencies and an increased glottal formant frequency. Conclusions drawn in this chapter are of interest for being applied in applications such as expressive/emotional speech recognition/labeling or synthesis.

Bibliography

[1] N. Campbell. *Expressive / Affective Speech Synthesis.* Springer Handbook on Speech Processing and Speech Communication, 2007.

[2] N. Sturmel. *Analyse de la qualité vocale appliquée à la parole expressive.* PhD thesis, Université Paris Sud 11, Faculté des Sciences d'Orsay, France, 2011.

[3] E. Lombard. Le signe de l'elevation de la voix. In *Annales des Maladies de l'Oreille et du Larynx*, volume 37, pages 101–119, 1911.

[4] W. Van Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes. Effects of noise on speech production: acoustic and perceptual analyses. *J. Acoust. Soc. Am.*, 84:917–928, 1988.

[5] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck. An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *Proc. Interspeech Conf.*, 2006.

[6] H. Boril, P. Fousek, and H. Hoge. Two-stage system for robust neutral/lombard speech recognition. In *Proc. Interspeech Conf.*, pages 1074–1077, 2007.

[7] J. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, 20:151–173, 1996.

[8] J. Hansen and V. Varadarajan. Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 17:366–378, 2009.

[9] J. Junqua. The lombard reflex and its role on human listeners. *J. Acoust. Soc. Am.*, 93:510–524, 1993.

[10] A. Ni Chasaide and C. Gobl. Voice source variation. *The Handbook of Phonetic Sciences*, pages 427–461, 1997.

[11] G. Bapineedu, B. Avinash, S. Gangashetty, and B. Yegnanarayana. Analysis of lombard speech using excitation source information. In *Proc. Interspeech Conf.*, 2009.

[12] A. El Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans. on Signal Processing*, 39 (2):411–423, 1991.

[13] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech*, 2009.

[14] D. Brookes and D. Chan. Speaker characteristics from a glottal airflow model using glottal inverse filtering. *Institue of Acoust.*, 15:501–508, 1994.

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

[16] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7:569–586, 1999.

[17] Online. http://aparat.sourceforge.net/index.php/main_page. *TKK Aparat Main Page*, 2008.

[18] Online. The snack sound toolkit. In *http://www.speech.kth.se/snack/*.

[19] V. Delplancq B. Harmegnies, J. Esling. Quantitative study of the effects of setting changes on the ltas. In *European Conference on Speech Communication and Technology*, pages 139–142, 1989.

[20] P. Alku, T. Backstrom, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112:701–710, 2002.

[21] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[22] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[23] G. Beller. Analyse et modèle génératif de l'expressivité - application à la parole et à l'interprétation musicale. In *PhD Thesis (in French), Universite Paris VI - Pierre et Marie Curie, IRCAM*, 2009.

[24] G. Beller. Influence de l'expressivité sur le degré d'articulation. In *RJCP*, 2007.

[25] G. Beller, N. Obin, and X. Rodet. Articulation degree as a prosodic dimension of expressive speech. In *Fourth International Conference on Speech Prosody*, 2008.

[26] B. Lindblom. Economy of speech gestures. *The Production of Speech, Spinger-Verlag, New-York*, 1983.

[27] B. Picart, T. Drugman, and T. Dutoit. Analysis and synthesis of hypo and hyperarticulated speech. In *7th ISCA Speech Synthesis Workshop*, 2010.

[28] D. Klatt and L. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[29] D. Childers. *Speech Processing and Synthesis Toolboxes*. Wiley and Sons, Inc., 1999.

[30] E. Keller. The analysis of voice quality in speech processing. *Lecture Notes in Computer Science*, pages 54–73, 2005.

[31] K. Sjolander and J. Beskow. Wavesurfer - an open source speech tool. In *ICSLP*, volume 4, pages 464–467, 2000.

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

[33] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9:21–29, 2001.

[34] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech Conf.*, 2009.

Chapter 10

Conclusion on the Glottal Flow Estimation and its Applications

Part II has focused on the problem of the automatic estimation of the glottal flow directly from the speech waveform, emphasizing some of its potential applications. The main difficulty with this issue is the unavailability of any ground truth reference, since neither the vocal tract nor the glottal contribution are observable. This makes the problem a typical case of blind separation, which also implies that no quantitative assessment of the performance of glottal source estimation techniques is possible on natural speech. The main contributions of Part II are the following:

- Chapter 5 provided a complete theoretical framework for mixed-phase (or causal-anticausal) separation. A new algorithm relying on a Complex Cepstrum-based Decomposition (CCD) was proposed. This technique was shown to be functionally equivalent but much faster than an existing method for mixed-phase deconvolution based on the Zeros of the Z-Transform (ZZT), which is advantageous for the design of real-time applications. Effects of windowing on the quality of the mixed-phase separation were studied on synthetic signals and a set of optimal constraints on the window to apply was derived. Relying on these conclusions, it was shown, on both synthetic and real speech signals, that the proposed Complex Cepstrum-based Decomposition can be effectively used for glottal flow estimation.

- Chapter 6 aimed at providing a review and quantitative evaluation of the main state-of-the-art glottal flow estimation techniques. The effectiveness of the CCD method proposed in Chapter 5 was compared to two other well-known techniques of glottal source estimation, representatives of the main approaches to this problem: the Closed Phase Inverse Filtering (CPIF) and the Iterative Adaptive Inverse Filtering (IAIF) methods. The robustness and influence of various factors on the estimation were first studied on synthetic signals. In clean conditions, CCD was the best performing method, while CPIF also led to an efficient parametrization of the glottal flow. Although achieving the worst results in high-quality recordings, IAIF turned out to be the most robust technique, outperforming other approaches for Signal-to-Noise Ratio (SNR) below 40 dB. A slight degradation was observed with an increasing fundamental frequency, and decreasing first formant frequency. In a second time, experiments were performed on a large corpus of expressive speech. The separability of three voice qualities was considered as a measure of the ability of the methods to discriminate different phonation types. It was shown that CCD, and CPIF in a lesser extent, are the best methods. It was also discussed which features are the best suited for revealing differences of voice quality.

- Chapter 7 extended the formalism of mixed-phase decomposition to chirp analysis. This allowed to remove the constraint of being synchronized on a Glottal Closure Instant (GCI), as required in the traditional approach introduced in Chapter 5. An automatic technique for determining the optimal chirp contour for both the CCD and ZZT methods was proposed. The resulting technique was shown to perform an efficient estimation of the glottal flow, while presenting the advantage of operating asynchronously as it is done in usual speech processing systems.

- Chapter 8 investigated the use of features related to the glottis for the automatic detection of voice pathologies. First we focused on the discrimination power and redundancy of vocal tract-based, glottal and prosodic features. Their assessment relied on measures derived from Information Theory. This allowed an objective evaluation, independently of any subsequent classifier. Some of the proposed glottal features were observed to be particularly relevant for detecting voice disorders. It was also shown that glottal characteristics exhibit an interesting complementarity with other types of features. Secondly the potential of using phase-based features for detecting voice pathologies was explored. It was shown that representations based on group delay functions are particularly suited for capturing irregularities in the speech signal. The adequacy of the mixed-phase model during the voice production was discussed and shown to convey relevant information. The efficiency of the proposed phase-based features was explained by their higher sensitivity to turbulences during the phonation process.

- Chapter 9 focused on glottal-based analysis of expressive voices. Two particular modes of expressivity were studied: Lombard speech and hypo or hyperarticulated speech. We analyzed how the glottal behaviour is modified in Lombard speech as a function of the type and level of noise. We then investigated how vocal tract-based and glottal characteristics are affected during the production of speech with various degrees of articulation (i.e with various efforts of pronunciation).

Part III

The Deterministic plus Stochastic Model of the Residual Signal and its Applications

Chapter 11

The Deterministic plus Stochastic Model of the Residual Signal

Contents

Abstract

The modeling of speech production often relies on a source-filter approach. Although methods parameterizing the filter have nowadays reached a certain maturity, there is still a lot to be gained for several speech processing applications in finding an appropriate excitation model. This chapter presents a Deterministic plus Stochastic Model (DSM) of the residual signal. The DSM consists of two contributions acting in two distinct spectral bands delimited by a maximum voiced frequency. Both components are extracted from an analysis led on a speaker-dependent dataset of pitch-synchronous residual frames. The deterministic part models the low-frequency contents and arises from an orthonormal decomposition of these frames. As for the stochastic component, it is a high-frequency noise modulated both in time and frequency. Some interesting phonetic and computational properties of the DSM are also highlighted. The applicability of the DSM in speech synthesis and speaker recognition is respectively studied in Chapters 12 and 13.

This chapter is based upon the following publications:

- Thomas Drugman, Geoffrey Wilfart, Thierry Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Interspeech Conference, Brighton, U.K, 2009 [ISCA Best Student Paper award].
- Thomas Drugman, Thierry Dutoit, *The Deterministic plus Stochastic Model of the Residual Signal and its Applications*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*

Many thanks to Geoffrey Wilfart for his helpful discussions.

11.1 Introduction

In speech processing, the modeling of the speech signal is generally based on a source-filter approach [1]. In such an approach, the source refers to the excitation signal produced by the vocal folds at the glottis, while the filtering operation refers to the action of the vocal tract cavities. In several speech processing applications, separating these two contributions is important as it could lead to their distinct characterization and modeling. The actual excitation signal is the airflow arising from the trachea and passing through the vocal folds, and is called the glottal flow [1]. Its estimation, parametrization and applicability have been studied all troughout Part II. However, it has been emphasized that glottal flow estimation directly from the speech waveform is a typical blind separation problem since neither the glottal nor the vocal tract contributions are observable.

This makes the glottal flow estimation a complex issue [2], and explains why it is generally avoided in usual speech processing systems. For this reason, it is generally preferred to consider, for the filter, the contribution of the spectral envelope of the speech signal, and for the source, the residual signal obtained by inverse filtering. Although not exactly motivated by a physiological interpretation, this approach has the advantage of being more practical while giving a sufficiently good approximation to the actual deconvolution problem.

Methods parameterizing the spectral envelope such as the well-known LPC or MFCC-like features [3], are widely used in almost every field of speech processing. On the contrary, methods modeling the excitation signal are still not well established and there might be a lot to be gained by incorporating such a modeling in several speech processing applications.

The goal of this chapter is to propose a Deterministic plus Stochastic Model (DSM) of the residual signal. The usefulness of this model both speech synthesis and speaker recognition will be shown respectively in Chapters 12 and 13. The proposed DSM of the residual signal results from an analysis led on a speaker-dependent set of residual frames that are synchronous with a Glottal Closure Instant (GCI) and whose length is set to two pitch periods (see Section 11.2). This process is required for matching residual frames so that they are suited for a common modeling. Each residual frame $r(t)$ is modeled as the sum of two components: *i)* a low-frequency deterministic component $r_d(t)$, based on a Principal Component Ananlysis (PCA) decomposition and detailed in Section 11.4, and *ii)* a high-frequency modulated noise component $r_s(t)$ described in Section 11.5. These two components are separated in the spectral domain by a particular frequency called *maximum voiced frequency*, as explained in Section 11.3. Finally, two important properties of the DSM, namely speed of convergence and phonetic independence, are respectively discussed in Sections 11.6 and 11.7. Finally, Section 11.8 concludes this chapter.

11.2 A Dataset of Pitch-Synchronous Residual Frames

The workflow for obtaining pitch-synchronous residual frames is presented in Figure 11.1. For this, a speaker-dependent speech database is analyzed. First the locations of the Glottal Closure Instants (GCIs) are estimated from the speech waveform using the SEDREAMS algorithm proposed in Section 3.3 (or [4]). GCIs refer to the instants of significant excitation of the vocal tract. These particular time events correspond to the moments of high energy in the glottal signal during voiced speech. In our process, GCI positions are used as anchor points for synchronizing residual frames. In parallel, a Mel-Generalized Cepstral (MGC) analysis is performed on the speech signals, as these features have shown their efficiency to capture the spectral envelope [3]. As recommended in [5], we used the parameter values $\alpha = 0.42$ ($Fs = 16kHz$) and $\gamma = -1/3$ for MGC extraction. In this paper, we opted for the MGCs as they are widely used in speech synthesis [5], albeit other filter coefficients could be used

as an alternative. Residual signals are then obtained by inverse filtering. Pitch-synchronous residual frames are finally isolated by applying a GCI-centered, two-pitch-period long Blackman windowing. The resulting dataset serves as a basis for extracting the components of the proposed DSM of the residual signal, as explained in the following sections.

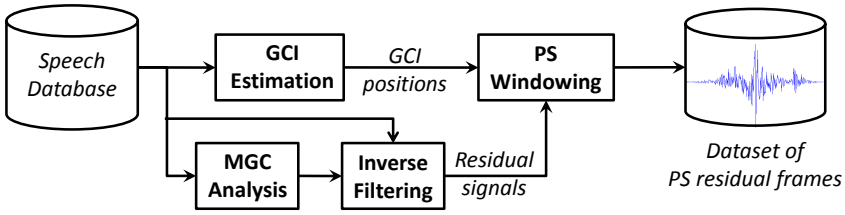

Figure 11.1 - *Workflow for obtaining the pitch-synchronous residual frames.*

11.3 The Maximum Voiced Frequency

As previously mentioned, the DSM consists of the superposition of a deterministic $r_d(t)$ and a stochastic $r_s(t)$ component of the residual signal $r(t)$. In this model, similarly to what is done in the Harmonic plus Noise Model (HNM, [6]), these two contributions are supposed to hold in two distinct spectral bands. The boundary frequency between these two spectral regions is called the *maximum voiced frequency*, and will be denoted F_m in the following. Some methods have already been proposed for estimating F_m from the speech waveform [6], [7]. Figure 11.2 illustrates the distribution of F_m estimated by the technique described in [6] for three voice qualities (loud, modal and soft) produced by the same German female speaker. For this example, we used the De7 database originally designed for creating diphone databases for expressive speech synthesis [8]. A first conclusion drawn from this figure is that significant differences between the distributions are observed. More precisely, it turns out that, in general, the soft voice has a low F_m (as a result of its breathy quality) and that the stronger the vocal effort, the more harmonicity in the speech signal, and consequently the higher F_m. However, it is worth noting that, although statistical differences are observed, obtaining a reliable trajectory of F_m for a given utterance is a difficult problem [9]. For this reason, as it is done in [9] or [10], we prefer in this work to consider a fixed value of F_m for a given speaker with a given voice quality. Therefore, we use in the rest of this paper the mean value of F_m extracted on a given dataset. Regarding the example of Figure 11.2, this leads to $F_m = 4600$ Hz for the loud, 3990 Hz for the modal, and 2460 Hz for the soft voice.

11.4 Modeling of the Deterministic Component

In order to model the low-frequency contents of the pitch-synchronous residual frames (extracted as explained in Section 11.2), it is proposed to decompose them on an orthonormal basis obtained by Principal Component Analysis (PCA, [11]). Preliminarily to this, the residual frames are normalized in prosody as exposed in Figure 11.3, i.e they are normalized both in pitch period and energy. This step ensures the coherence of the dataset before applying PCA. Note that, assuming the residual signal as an approximation of the glottal source, resampling the residual frames by interpolation or decimation should preserve their shape and consequently their most important glottal features (such as the open quotient or the asymmetry coefficient [12]).

It is worth noticing that, for speech synthesis purpose, particular care has to be taken when choosing the number of points for length normalization. Indeed, in order to avoid the appearance of

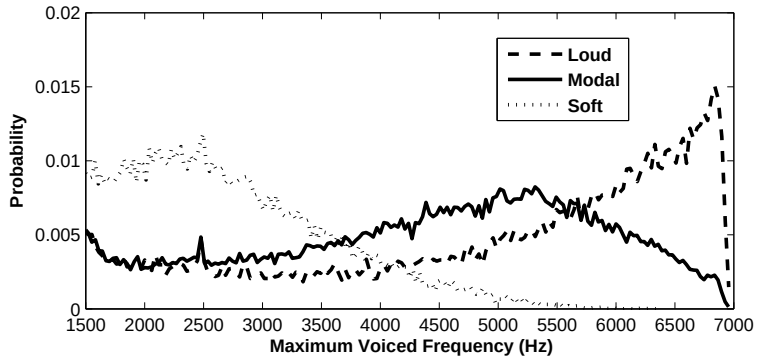

Figure 11.2 - *Histogram of the maximum voiced frequency F_m for the same female speaker with three different voice qualities.*

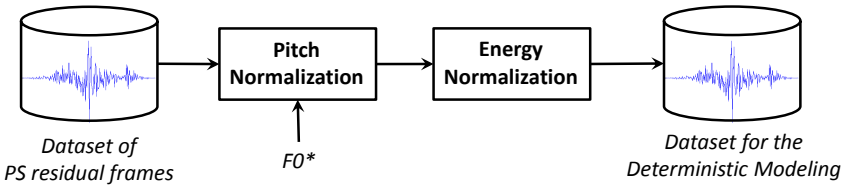

Figure 11.3 - *Workflow for obtaining the dataset for the deterministic modeling.*

energy holes at synthesis time (occuring if the useful band of the deterministic part does not reach F_m after pitch denormalization, see Section 12.2), the pitch value F_0^* for the normalization has to respect the condition:

$$F_0^* \leq \frac{F_N}{F_m} \cdot F_{0,min} \tag{11.1}$$

where F_N and $F_{0,min}$ respectively denote the Nyquist frequency and the minimum pitch value for the considered speaker.

PCA can now be calculated on the resulting dataset, allowing dimensionality reduction and feature decorrelation. PCA is an orthogonal linear transformation which applies a rotation of the axis system so as to obtain the best representation of the input data, in the Least Squared (LS) sense [11]. It can be shown that the LS criterion is equivalent to maximizing the data dispersion along the new axes. PCA can then be achieved by calculating the eigenvalues and eigenvectors of the data covariance matrix [11].

Let us assume that the dataset consists of N residual frames of m samples. PCA computation will lead to m eigenvalues λ_i with their corresponding eigenvectors μ_i (here called *eigenresiduals*). λ_i is known to represent the data dispersion along axis μ_i [11]. Using the k first eigenresiduals (with $k \leq m$), the Cumulative Relative Dispersion (CRD) is defined as:

$$CRD(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}, \tag{11.2}$$

and is a relative measure of the dispersion covered over the dataset using these k eigenresiduals. Figure 11.4 displays a typical evolution of this variable for a given male speaker (F_s=16kHz, m=280 and thus F_0^*=114Hz for this example). It is observed that PCA allows a high dimensionality reduction since very few eigenresiduals are sufficient to cover the greatest amount of dispersion.

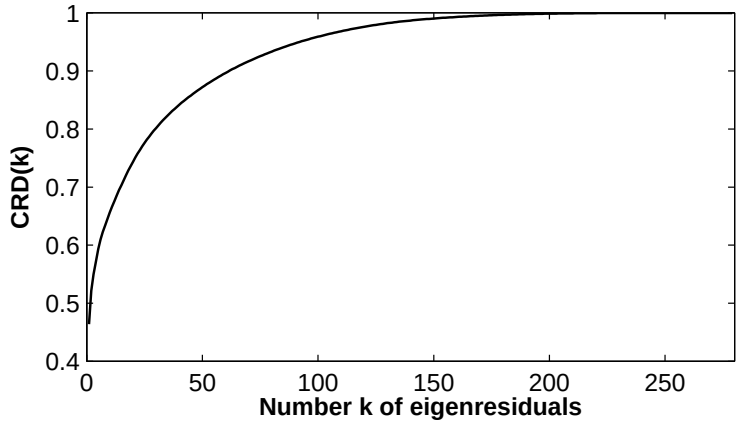

Figure 11.4 - *Evolution of the Cumulative Relative Dispersion (CRD) as a function of the number of eigenresiduals for a given male speaker.*

It is worth noting here that the eigenresiduals of highest orders contribute mainly to the reconstruction of the high-frequency contents of the residual frames. In practice, we observed that, with the usual value of F_m/F_N, the use of only the first eigenresidual (whose relative dispersion is of 46% in the example of Figure 11.4) is sufficient for a good modeling below F_m, and that the effect of higher order eigenresiduals is almost negligible in that spectral band. Since its importance on the spectral contents below F_m is predominant, and as it will be confirmed in the applicative parts (Chapters 12 and 13), the first eigenresidual $\mu_1(n)$ (just called eigenresidual for the sake of conciseness in the following) can be considered to model the deterministic component of the DSM. To illustrate what this waveform looks like, Figure 11.5 shows the first eigenresidual for the same speaker as in Figure 11.4. It is interesting to note the strong similarity with the glottal flow derivative waveform used in many glottal flow models (such as the LF model [13]), mainly during the glottal open phase, and the clear discontinuity at the GCI position. However it is worth noticing that the first eigenresidual is a modeling of the residual signal, and not of the glottal source. Nonetheless, the residual signal conveys information about the glottal behaviour, which turns out to be reflected in the first eigenresidual shape.

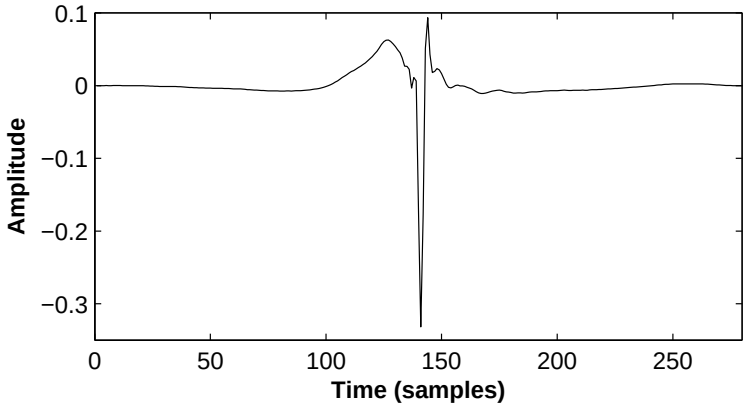

Figure 11.5 - *Illustration of the first eigenresidual $\mu_1(n)$ for a given male speaker.*

11.5 Modeling of the Stochastic Component

In the proposed DSM of the residual signal $r(t)$, the stochastic modeling $r_s(t)$ is similar to the noise part in the HNM [6]. It corresponds to a white Gaussian noise $n(t)$ convolved with an auto-regressive model $h(t)$, and whose time structure is controled by an energy envelope $e(t)$:

$$r_s(t) = e(t) \cdot [h(t) \star n(t)]. \tag{11.3}$$

The use of $h(t)$ and $e(t)$ is required to account respectively for the spectral and temporal modulations of the high-frequency contents of the residual. In order to estimate these two contributions, the dataset of pitch-synchronous residual frames (as extracted in Section 11.2) is considered, and the modifications exhibited in Figure 11.6 are brought to it. More precisely, frames are normalized in energy and only their contents beyond F_m is kept. On the resulting dataset, $h(t)$ is estimated as the Linear Predictive modeling of their averaged amplitude spectrum. Indeed, since F_m has been fixed and since the residual spectral envelope is almost flat over the whole frequency range, it is reasonable to consider that $h(t)$ has fairly the same effect on all frames: it acts as a high-pass filter beyond F_m. As for the energy envelope $e(t)$, it is determined as the average Hilbert envelope of the resulting high-filtered residual frames resampled to the normalized pitch value F_0^*. Note that several envelopes were studied in [10] for modeling the temporal characteristics of noise in the context of HNM and for analysis-synthesis purpose. The Hilbert envelope was shown to be one of the most appropriate for this purpose.

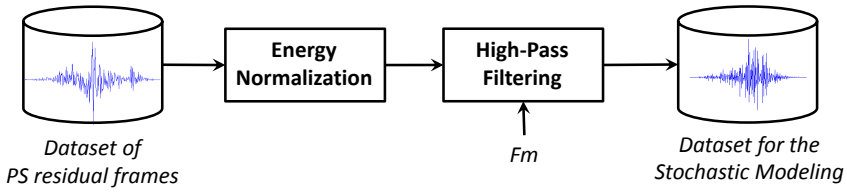

Figure 11.6 - *Workflow for obtaining the dataset for the stochastic modeling.*

11.6 Speed of Convergence

The proposed DSM of the residual signal makes use of two important waveforms: the eigenresidual $\mu_1(n)$ for the deterministic part, and the energy envelope $e(n)$ of the stochastic component. In order to estimate how much data is required for having a reliable estimation of these two signals, the male speaker AWB from the CMU ARCTIC database [14] was analyzed. This database contains about 50 minutes of speech recorded for Text-to-Speech purpose. The two reference waveforms were first computed on a large dataset containing about 150.000 pitch-synchronous residual frames. An additional estimation of these waveforms was then obtained by repeating the same operation on a held out dataset for the same speaker. The Relative Time Squared Error (RTSE) is used for both waveforms as a distance between the estimation $x_{est}(n)$ and the reference $x_{ref}(n)$ signals (where m is the number of points used for pitch normalization):

$$RTSE = \frac{\sum_{n=1}^{m} (x_{est}(n) - x_{ref}(n))^2}{\sum_{n=1}^{m} x_{ref}(n)^2} \tag{11.4}$$

Figure 11.7 displays the evolution of this measure (in logarithmic scale) with the size of the held out dataset. It may be observed that both estimations quickly converge towards the reference. From

this graph, it can be considered that a dataset containing around 1000 residual frames is sufficient for obtaining a correct estimation of both the deterministic and stochastic components of the DSM. To give an idea, this corresponds to about 7s of voiced speech for a male speaker, and about 4 s for a female voice.

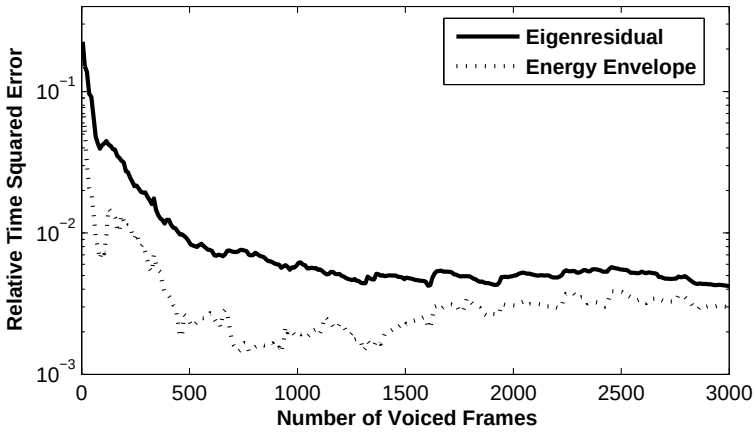

Figure 11.7 - *Speed of convergence for the eigenresidual and the energy envelope.*

11.7 Phonetic Independence

In the proposed DSM of the residual signal, the same modeling is used for any voiced segment. In other words, the same waveforms (eigenresidual or energy envelope) are used for the excitation of all voiced phonetic classes. In order to assess the validity of this assumption, the speaker AWB from the CMU ARCTIC database [14] was also analyzed. On the whole set A (first half) of this database, a reference eigenresidual was extracted. On dataset B (second half), sentences were segmented into phonetic classes, and for each class containing more than 1000 voiced frames (as suggested from Section 11.6 for obtaining a reliable estimation), the corresponding class-dependent eigenresidual was calculated.

Phonetic Class	aa	ae	ah	ao	aw	ax	ay	eh	er	ey	ih
RTSE	0.73	0.72	0.52	0.74	0.46	0.66	0.56	0.59	0.90	0.68	0.29
Phonetic Class	iy	l	m	n	ng	ow	oy	r	uw	w	y
RTSE	0.93	0.91	5.31	3.68	1.53	0.38	0.85	2.89	0.91	2.48	2.00

Table 11.1 - *Relative Time Squared Error (%) between the reference and the class-dependent eigenresiduals.*

Table 11.1 presents the values (in %) of the RTSE between the reference and the class-dependent eigenresiduals. It can be noticed that for most phonetic classes (16 out of the 22 cases), the RTSE is lower than 1% (which is, as seen in Figure 11.7, of the order of the estimation error). The difference with the reference eigenresidual is the highest for the nasalized consonants (/m/ and /n/). This may be explained by the difficulty in modeling the anti-formants of the vocal tract for such sounds. To illustrate the resulting differences, Figure 11.8 shows the reference eigenresidual and the one extracted for the phonetic class /m/ (for which the RTSE is the highest). It can be noticed that the main dissimilarities occur at the right of the GCI, while the left parts are almost identical. Indeed, according to the mixed-

phase model of speech [15], during the production of the speech signal, the response at the left of the GCI is dominated by the open phase of the glottal excitation, while the response at its right is mainly dominated by the vocal tract impulse response. After inverse filtering, the dissimilarities at the right of the GCI might then be explained by an imperfect modeling of the vocal tract transmittance for the phoneme /m/.

Nonetheless, since the results of Table 11.1 suggest that a fairly identical eigenresidual is extracted for the great majority of the voiced phonetic classes (and that for other classes, the difference is still relatively small), and given in addition that our informal attempts of incorporating a class-dependent modeling in analysis-synthesis led to no audible differences, the assumption of using a common modeling independent of the phonetic context can be supposed to hold in the rest of this paper.

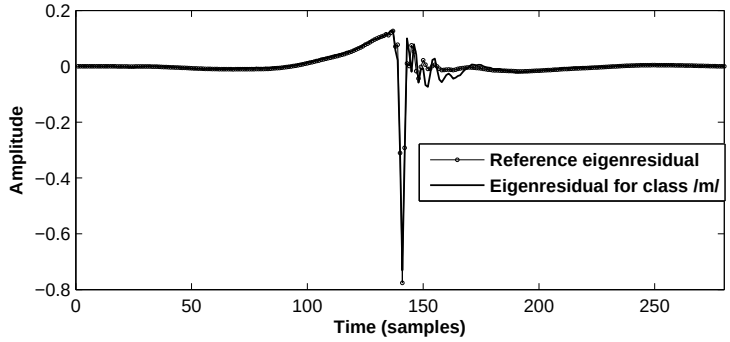

Figure 11.8 - *Waveforms of the reference eigenresidual and the one extracted on the phonetic class /m/.*

11.8 Conclusion

This chapter presented a new excitation model: the Deterministic plus Stochastic Model (DSM) of the residual signal. The DSM estimation is performed by automatic analysis of a speaker-dependent dataset of pitch-synchronous residual frames. These frames are modeled by two components acting in two distinct spectral bands: the deterministic and the stochastic components. The frequency demarcating the boundary between these two spectral regions is called the maximum voiced frequency. The low-frequency deterministic part consists of a decomposition on an orthonormal basis obtained by Principal Component Analysis. As for the high-frequency stochastic component, it is a noise modulated both in time and frequency. After a detailed description of the underlying theoretical framework, some computational and phonetic considerations were examined. It was proved that a speaker-dependent dataset of around 1000 voiced frames is sufficient for having a reliable estimation of the DSM components. It was also shown that the assumption of considering a common excitation modeling for all phonetic classes is valid. The applicability of the proposed DSM will be studied for two major fields of speech processing: speech synthesis (Chapter 12) and speaker recognition (Chapter 13).

Bibliography

[1] T. Quatieri. Discrete-time speech signal processing: Principles and practice. In *Prentice-Hall*, 2002.

[2] T. Drugman, B. Bozkurt, and T. Dutoit. Comparative study of glottal source estimation techniques. In *Computer Speech and Language - to appear*, 2011.

[3] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel generalized cepstral analysis - a unified approach to speech spectral estimation. In *ICSLP*, 1994.

[4] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech Conf.*, 2009.

[5] T. Toda H. Zen and K. Tokuda. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard challenge 2006. In *IEICE Trans. on Information and Systems*, 2006.

[6] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9:21–29, 2001.

[7] S. Han, S. Jeong, and M. Hahn. Optimum MVF estimation-based two-band excitation for HMM-based speech synthesis. *ETRI Journal*, 31(4):457–459, 2009.

[8] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In *Proc. 15th International Conference of Phonetic Sciences*, pages 2589–2592, 2003.

[9] Y. Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *PhD thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[10] Y. Pantazis and Y. Stylianou. Improving the modeling of the noise part in the harmonic plus noise model of speech. In *IEEE ICASSP*, 2008.

[11] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2002.

[12] G. Fant and J. Liljencrants Q. Lin. A four parameter model of glottal flow. In *STL-QPSR4*, pages 1–13, 1985.

[13] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4): 1–13, 1985.

[14] Online. CMU ARCTIC speech synthesis databases. In http://festvox.org/cmu_arctic/, 2009.

[15] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA ITRW VOQUAL03*, pages 21–24, 2003.

Chapter 12

Application of DSM to Speech Synthesis

Contents

Abstract

The Deterministic plus Stochastic Model (DSM) of the residual signal has been presented in Chapter 11 as a new excitation model. This chapter focuses on the use of the DSM vocoder so as to improve the quality delivered by parametric speech synthesizers. The resulting method is assessed within the frame of two applications: pitch modification in an analysis-synthesis context, and Hidden Markov Model (HMM)-based speech synthesis. The problem of pitch modification is first addressed as an important module for an efficient voice transformation system. In that experiment, DSM is compared to three well-known methods for this purpose: TDPSOLA, HNM and STRAIGHT. The four methods are compared through an important subjective test. The influence of the speaker gender and of the pitch modification ratio is analyzed. Despite its higher compression level, the DSM technique is shown to give similar or better results than other methods, especially for male speakers and important ratios of modification. The DSM turns out to be only outperformed by STRAIGHT for female voices. In a second time, we incorporate the DSM vocoder within a HMM-based speech synthesizer. In two subjective evaluations involving a large number of listeners, DSM is compared to the traditional pulse excitation, to the GPF and STRAIGHT methods. Results show that DSM significantly outperforms the pulse and the GPF excitation for both male and female voices and that it provides a quality equivalent to STRAIGHT. In addition, the proposed DSM technique requires few computational load and memory, which is essential for its integration in commercial applications.

This chapter is based upon the following publications:

- Thomas Drugman, Geoffrey Wilfart, Thierry Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Interspeech Conference, Brighton, U.K, 2009 [ISCA Best Student Paper award].
- Thomas Drugman, Thierry Dutoit, *The Deterministic plus Stochastic Model of the Residual Signal and its Applications*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*
- Thomas Drugman, Thierry Dutoit, *A Comparative Evaluation of Pitch Modification Techniques*, 18th European Signal Processing Conference, Aalborg, Denmark, 2010.

Many thanks to Geoffrey Wilfart for his helpful discussions.

12.1 Introduction

Two text-to-speech technologies have clearly emerged these last years. On one hand, the Unit Selection method [1] concatenates speech units picked up from a very large corpus, avoiding signal processing manipulations as much as possible, in order to minimize segmental quality degradations. Its biggest drawbacks lie in the difficulty of producing voice quality variations, required to produce expressive speech, and in the limited voice modification/conversion that are allowed.

On the other hand, Statistical Parametric Speech Synthesis [2] models the speech signal in various contextual situations. Synthesizers based on Hidden Markov Model (HMM) have thus recently gained considerable attention for their flexibility, smoothness and small footprint [3]. Nevertheless their main disadvantage is the quality of the produced speech, which exhibits the typical *buzziness* found in the old Linear Predictive Coding (LPC)-based speech coders. While techniques for modeling the filter are rather well-established, it is not the case for the source representation.

In order to overcome this hindrance, some works have proposed a more subtle excitation model, enhancing in this way the final quality and naturalness. In the Codebook Excited Linear Predictive (CELP) approach [4], the residual signal is constructed from a codebook containing several typical excitation frames. In [5] we proposed the use of a codebook of pitch-synchronous residual frames to construct the voiced excitation. The Multi Band Excitation (MBE) modeling [6] suggests to divide the frequency axis in several bands, and a voiced/unvoiced decision is taken for each band at any time. A process based on a Multi-Band Resynthesis pitch-synchronous OverLap-Add (MBROLA) of the speech signal has been proposed in [7]. According to the Mixed Excitation (ME) approach [8], the residual signal is the superposition of both a periodic and a non-periodic component. Various models derived from the ME approach have been used in HMM-based speech synthesis [9], [10]. In [9], Yoshimura *et al.* integrated a ME coding method. In this framework, the excitation is obtained using a multi-band mixing model, containing both periodic and aperiodic contributions, and controled by bandpass voicing strengths. In a similar way, Maia *et al.* [10] made use of high-order filters to obtain these components, which were derived through a closed-loop procedure. A popular technique used in parametric synthesis is the STRAIGHT vocoder [11]. STRAIGHT excitation relies on a ME model weighting the periodic and noise components by making use of aperiodicity measurements of the speech signal [11]. Some other techniques, such as [12] or [13], have incorporated excitation signals based on the Liljencrants-Fant (LF) glottal flow model [14] into HMM-based synthesis. All these techniques tend to relatively reduce the produced buzziness, and therefore improve the overall quality.

The goal of this chapter is to integrate the Deterministic plus Stochastic Model (DSM) of the residual signal, as introduced in Chapter 11, into a parametric speech synthesizer and to evaluate its performance. It can be expected that, if the excitation is appropriately modeled via the DSM, the naturalness of the delivered voice will be enhanced. For this, the chapter is structured as follows. Section 12.2 describes the vocoder relying on the DSM. The resulting coding technique is assessed within the frame of two applications: pitch modification in an analysis-synthesis context (Section 12.3), and HMM-based speech synthesis (Section 12.4). For each of these two applications, the performance of DSM is assessed via a subjective test involving a large number of listeners and compared to other well-known state-of-the-art of vocoding. Finally, Section 12.5 concludes the chapter.

12.2 The DSM Vocoder

The DSM of the residual signal has been presented in Chapter 11 as a new excitation model. We here propose a method of speech synthesis relying on this approach. A workflow summarizing the resulting DSM vocoder can be found in Figure 12.1. The vocoder takes only two feature streams as input:

pitch values (F_0) for the source, and MGC coefficients for the filter (with $\alpha = 0.42$ and $\gamma = -1/3$, as indicated in Section 11.2). All other data is precomputed on a training dataset as explained in Chapter 11. As our informal attempts showed that adding eigenresiduals of higher orders (see Section 11.4) has almost no audible effect on the delivered speech synthesis, only the first eigenresidual is considered for speech synthesis purpose. The deterministic component $r_d(t)$ of the residual signal then consists of the (first) eigenresidual resampled such that its length is twice the target pitch period. Following Equation (11.3), the stochastic part $r_s(t)$ is a white noise modulated by the autogressive model and multiplied in time by the energy envelope centered on the current Glottal Closure Instant (GCI). During synthesis, artificial GCIs are created using the F_0 information. Note that the energy envelope is also resampled to the target pitch. Both components are then overlap-added so as to obtain the residual signal $r(t)$. In the case of unvoiced regions, the excitation merely consists of white Gaussian noise. The synthesized excitation is finally the input of the Mel-Log Spectrum Approximation (MLSA, [15]) filter to generate the final speech signal.

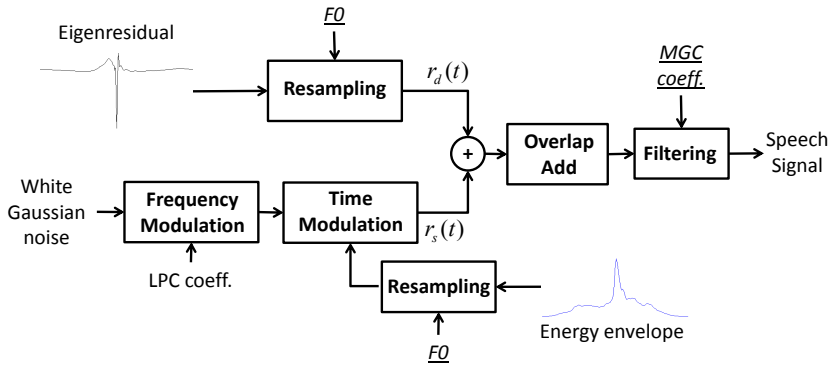

Figure 12.1 - *Workflow of the DSM vocoder. Input features (indicated in italic and underlined) are the target pitch F_0 and the MGC filter coefficients. All other data is precomputed on a training dataset.*

12.3 Evaluation for Pitch Modification in Analysis-Synthesis

Voice transformation refers to the various modifications one may apply to the sound produced by a person such that it is perceived as uttered by another speaker [16]. These modifications encompass various properties of the speech signal such as prosodic, vocal tract-based as well as glottal characteristics. Although all these features should be taken into account in an efficient voice transformation system, this study only focuses on pitch modifications, as pitch is an essential aspect in the way speech is perceived. More precisely, the main goal of this section is to compare the capabilities of the DSM vocoder presented in Section 12.2 to the main state-of-the-art techniques for pitch modification.

This section is structured as follows. Section 12.3.1 gives a brief overview on the methods considered in this study, namely: the DSM of the residual signal, the Time-Domain Pitch-Synchronous Overlap-Add technique (TDPSOLA, [17]), the Harmonic plus Noise Model of speech (HNM, [18]) and STRAIGHT [11]. In Section 12.3.2 these methods are compared through a subjective evaluation regarding their pitch modification performance. Finally Section 12.3.3 discusses in depth the main observations drawn from the results.

12.3.1 Methods for Pitch Modification

Various approaches for pitch modification have already been proposed in the literature. Some of them are based on a parametric modeling (HNM [18], STRAIGHT [11], ARX-LF [19]), or on a phase vocoder [20], [21], while others rely on a non-parametric representation (TDPSOLA [17]). This section briefly presents the methods that will be compared in Section 12.3.2: the DSM, TDPSOLA, HNM and STRAIGHT algorithms. For information, the footprint of each method is presented in number of parameters/second, giving an idea of their compression level.

Deterministic plus Stochastic Model of the Residual Signal (DSM)

We here make use of the DSM vocoder introduced in Section 12.2. As input of the workflow exhibited in Figure 12.1, 25 MGC parameters are used for the vocal tract, and only F_0 for the excitation, all other data being pre-computed on the speaker-dependent database. These features are extracted every 5 ms which leads to a 5200 parameters/s vocoder.

Time-Domain Pitch-Synchronous Overlap-Add (TDPSOLA)

The TDPSOLA technique [17] is probably the most famous non-parametric approach for pitch modification. According to this method, pitch-synchronous speech frames whose length is a multiple of the pitch period are duplicated or eliminated. It is in this way assumed that the pitch can be modified while keeping the vocal tract characteristics unchanged. In our implementation we considered two pitch period-long speech frames centered on the GCIs. GCI positions were located by the SEDREAMS method described in Section 3.3 (or [22]), providing a high-quality phase synchronization. As this technique is based on the speech waveform itself (sampled at 16 kHz in our experiments), 16000 values/s are necessary.

Harmonic plus Noise Model (HNM)

The Harmonic plus Noise Model (HNM, [18]) assumes the speech signal to be composed of a harmonic part and a noise part. The harmonic part accounts for the quasi-periodic component of the speech signal while the noise part accounts for its non-periodic components (e.g., fricative or aspiration noise, etc.). The two components are separated in the frequency domain by a time-varying parameter, referred to as maximum voiced frequency F_m. The lower band of the spectrum (below F_m) is assumed to be represented solely by harmonics while the upper band (above F_m) is represented by a modulated noise component. In this study, we used the HNM algorithm with its default options. Since the number of harmonics (and consequently of parameters) is different depending on F_0 and F_m, the bitrate may vary across speakers and sentences. In average, we found that around 10000 parameters were necessary for coding 1s of speech.

STRAIGHT

STRAIGHT is a well-known vocoding system [11] which showed its ability to produce high-quality voice manipulation and was successfully incorporated into HMM-based speech synthesis. STRAIGHT is basically based on both a source information extractor as well as a smoothed time-frequency representation [11]. In this work, we employed the version publicly available in [23] with its default options. In this implementation, the algorithm extracts every 1 ms: the pitch, aperiodic components of the excitation (513 coeff.) and a representation of the smoothed spectrogram (513 coeff.). This leads to a high-quality vocoder using a bit more than 1 million parameters/s.

12.3.2 Experiments

In this part, methods presented in Section 12.3.1 are evaluated on 3 male (AWB, BDL and JMK) and 2 female (CLB and SLT) speakers from the CMU ARCTIC database [24]. For each speaker, the first three sentences of the database were synthesized using the four techniques, and this for 5 pitch modification ratios: 0.5, 0.7, 1, 1.4 and 2. This leads to a total set containing 300 sentences. The DSM technique was compared to the three other approaches (TDPSOLA, HNM and STRAIGHT) through a Comparative Mean Opinion Score (CMOS, [25]) test composed of 30 pairwise sentences chosen randomly among the total set. 27 people (mainly naive listeners) participated to the test. For each sentence they were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference. The CMOS values range on a gradual scale presented in Table 12.1. The CMOS scores vary from -3 (meaning that DSM is much worse than the other technique) to +3 (meaning the opposite). A score of 0 is given if both versions are found to be equivalent. It is worth noting that, due to the unavailability of a ground truth reference of how a sentence whose pitch has been modified by a given factor should sound, participants were asked to score according to their overall appreciation of the different versions. These scores then reflect both the quality of pitch modification, as well as the possible artifacts that the different signal representations may generate.

Much better	+3
Better	+2
Slightly better	+1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

Table 12.1 - *Grades in the CMOS scale.*

Figure 12.2 displays the CMOS results with their 95% confidence intervals for the three comparisons and according to the gender of the speaker. For male voices, it can be noticed that DSM gives scores similar to TDPSOLA, while its advantage over HNM, and STRAIGHT in a lesser extent, is appreciable. For female speakers, the tendency is inversed. DSM is comparable to HNM while it is superior to TDPSOLA. Albeit for such comparative subjective tests transitional properties can not be assumed to hold, it however seems that STRAIGHT outperforms all other techniques for female voices. It is also worth noticing that the degradation of DSM with regard to STRAIGHT for female speakers is done at the expense of a high gain of compression and complexity. Depending on the considered application, the choice of one of the compared method should then result from a trade-off between these latter criteria (i.e speech quality vs compression rate).

In Figure 12.3 the preference scores for both male and female speakers can be found. Although somehow redundant with the previous results, this figure conveys information about the percentage of preference for a given method and about the ratio of indifferent opinions. Interestingly it can be noted that DSM was in general preferred to other methods, except for female speakers where STRAIGHT showed a clear advantage. In [19], Vincent *et al.* compared an improved ARX-LF framework to TDPSOLA and HNM through a small preference test. Even though these results are obviously not extrapolable, the preference scores they obtain are stronlgy similar to ours (while the DSM technique is much simpler), except for the comparison with TDPSOLA on female voices where ARX-LF was shown to be inferior.

Finally, the evolution of the performance with the pitch modification ratio is analyzed in Figure

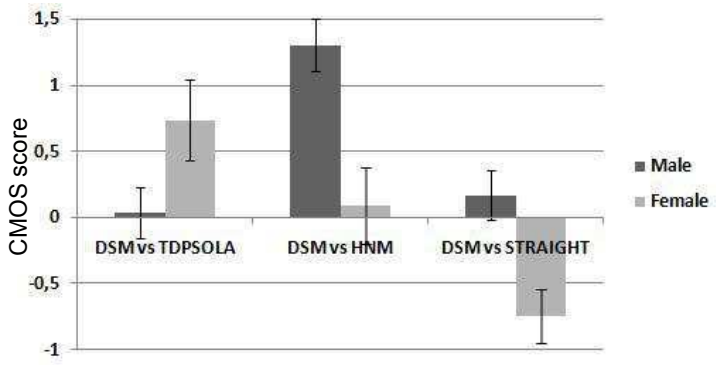

Figure 12.2 - *CMOS results together with their 95% confidence intervals for the three comparisons and for both male and female speakers.*

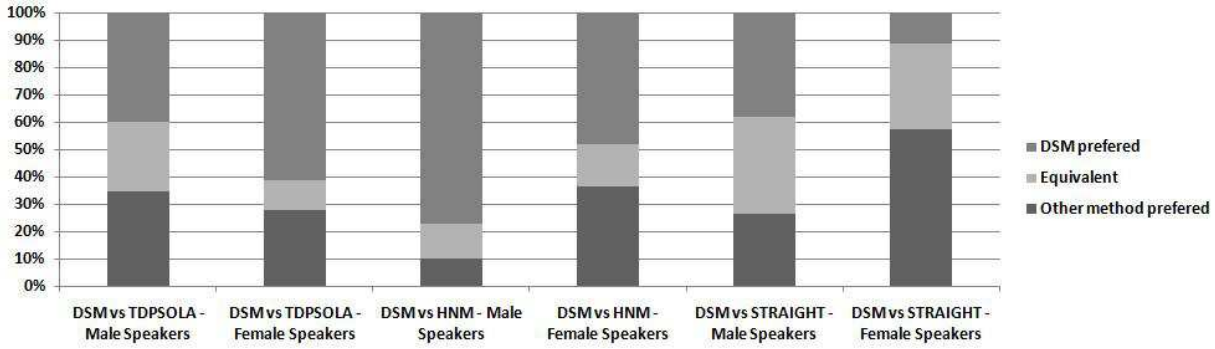

Figure 12.3 - *Preference scores for the three comparisons and for both male and female speakers.*

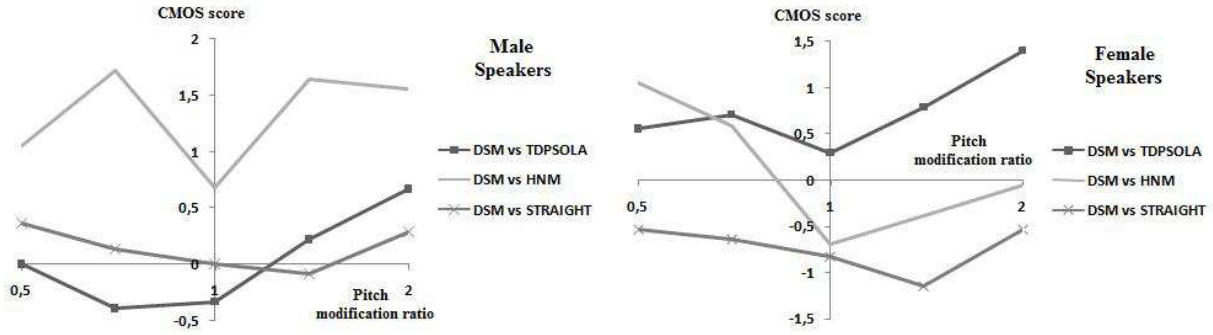

Figure 12.4 - *Evolution of the CMOS results with the pitch modification ratio for both male and female speakers.*

12.4. As a reminder, the higher the CMOS score, the more DSM was preferred regarding the method to which it was compared. A positive (negative) value means that, in average, the DSM (the other method) was preferred. Interestingly, it can be observed that in general plots tend to exhibit a minimum in 1 (where no pitch modification was applied) and go up around this point. This implies that the relative performance of DSM over other techniques increases as the pitch modification ratio is important. This was expected regarding the comparison with TDSPOLA, but the same observation seems to hold for STRAIGHT, and for HNM (though in a lesser extent for male voices). Note that in our implementation

of TDPSOLA, a GCI-synchronous overlap-add was performed even when no pitch modification was required. This may explain why, for female voices (for which GCIs are known to be difficult to be precisely located), listeners slighlty preferred DSM over TDPSOLA, even without pitch modification.

12.3.3 Discussion about the results

Several important conclusions can be drawn from the study led in Section 12.3.2:

- Interestingly, the DSM approach, despite its small footprint, gives similar or better results than other state-of-the-art techniques. Its efficiency probably relies on its ability to implicitly capture and process the essential of the phase information via the eigenresidual. The pitch-dependent resampling operations involved in its process indeed preserve the most important glottal properties (such as the open quotient and asymmetry coefficient). Nevertheless a degradation for female speakers is noticed. This can be mainly explained by the fact that the spectral envelope we used may contain pitch information. Although this effect can be alleviated by the use of Mel-Generalized Cepstral (MGC, [26]) coefficients instead of the traditional LPC modeling (since MGCs make use of a warped frequency axis), it may still occur for high-pitched voices where the risk of confusion between F_0 and the first formant frequency F_1 is more important. After pitch modification, this effect leads to detrimental source-filter interactions, giving birth to some audible artefacts. Note that this effect is almost completely avoided with STRAIGHT, as this method makes use of a time-frequency smooth representation of the spectral envelope [11]. Reducing this drawback within the DSM framework is the object of ongoing work, possibly by applying a GCI-synchronous spectral analysis instead of achieving it in an asynchronous way.

- Results we obtained for HNM corroborate the conclusions from [19] and [27]. In [27], the observation that sinusoidal coders produce higher quality speech for female speakers than for male voices is justified by the concept of critical phase frequency, below which phase information is perceptually irrelevant. Note also that we used the HNM algorithm with its default options. In this version, we observed that the quality of the HNM output was strongly affected for some voices by a too low estimation of the maximum voiced frequency. This led to an unpleasant predominance of noise in the speech signal. Fixing the maximum voiced frequency to a constant value (as in the DSM technique) could lead to a relative improvement for these problematic voices.

- The degradation of TDPSOLA for female speakers is probably due to the difficulty in obtaining accurate pitch marks for such voices. This results in inter-frame phase incoherences, degrading the final quality. Besides note that TDPSOLA requires the original speech waveform as input and is then not suited for parametric speech synthesis.

- It turns out from this study and the one exposed in [19] that approaches based on a source-filter representation of speech lead to the best results. This is possible since these techniques process the vocal tract and the glottal contributions independently. Among these methods, STRAIGHT gives in average the best results but requires heavy computation load. The DSM and the improved ARX-LF technique proposed in [19] seem to lead to a similar quality. Note that STRAIGHT was successfully integrated into a HMM-based speech synthesizer in [28]. As it will be shown in Section 12.4 (or see [29]), this is also the case for the DSM. In [30], it was proposed to incorporate the traditional ARX-LF model in a statistical parametric synthesizer. Although an improvement regarding the basic baseline was reported, it seems that this latter is less significant than it was achieved by STRAIGHT and DSM. It is then clear that the good quality obtained

in [19] and [31] with the improved ARX-LF method is reached thanks to the modeling of the LF-residual (i.e the signal obtained after removing the LF contribution in the excitation). This is possible in an analysis-synthesis task (where the target LF-residual is available), but was not yet carried out in speech synthesis. Finally note that the ARX-LF approach has the flexibility to potentially produce easy modifications of voice quality or emotion ([31], [30]) since it relies directly on a parametric model of the glottal flow (which is not the case for the DSM and STRAIGHT techniques).

Given the good quality of the DSM vocoder in an analysis-synthesis context, and its ability to perform pitch modification, it would be of interest to incorporate it into a parametric speech synthesizer. This is the object of Section 12.4 which focuses on the integration and evaluation of the DSM technique into HMM-based speech synthesis.

12.4 Evaluation for HMM-based Speech Synthesis

Before the last few years, synthetic speech was typically produced by a method based on Unit Selection [1], also called Non Uniform Units (NUU). For this, frames of natural speech selected from a huge database were concatenated, possibly applying signal processing to them so as to smooth discontinuities at jointures [1]. The main advantage of this approach is the high quality at the waveform level, as it relies on segments of *real* speech, therefore embedding all its subtleties. Nonetheless its performance may rapidly degrade due to some possible discontinuities (despite the smoothing process), or to some miss, i.e when an unit to synthesize is not present in the corpus [3]. These hindrances are generally overcome by increasing the size of the database so as to enhance its covering. However this implies to stock a corpus of several hours of speech, and the ensuing synthesizer requires a high amount of ROM memory.

Recently, a new approach of speech synthesis has emerged: the Statistical Parametric Speech (SPS) synthesis [2]. The most famous technique representative of SPS is the HMM-based speech synthesis [3], whose principle is detailed in Section 12.4.1. Basically SPS relies on a statistical modeling of speech parameters. At synthesis time, given the input text, the most likely trajectory of speech parameters is generated from the statistical model. These parameters are then provided to a vocoder, which synthesizes the final speech waveform. The main advantages of SPS are [2], [3]:

- its flexibility: Thanks to the statistical model which it relies on, SPS has the ability to easily modify voice characteristics, to be applied to various languages with little modification, or to produce various speaking styles or emotional speech using a small amount of speech data.
- its small footprint: Albeit it is trained on a large speech corpus, the implementation of the SPS system only requires the storage of the statistical modeling, which results in a huge compression rate. Typically, a HMM-based speech synthesizer holds within 1MB, which makes it suited for small devices and embedded systems.
- its smoothness and stability: Contrarily to the NUU technique, SPS does not suffer from neither discontinuities (i.e the speech parameter trajectories are smooth), nor misses (in the case where the sound to produce is not covered in the training database, SPS extrapolates information relying on its statistics, which tends to give better results than NUU).

However, SPS presents some drawbacks which are inherent to its statistical and parametric nature. First, the statistical process tends to oversmooth the generated trajectories, which results in what is

called a *muffled speech* [32]. Secondly, the parametric representation of speech degrades the quality compared to NUU, which uses segments of *natural* speech. Voice produced by SPS typically exhibits a *buzziness*, as found in old LPC-based speech coders. The goal of this section is precisely to alleviate this latter disadvantage by integrating the DSM vocoder in order to make synthetic speech sound less buzzy, and consequently improve the final quality.

The remainder of this section is organized as follows. Section 12.4.1 describes the principle of HMM-based speech synthesis, as well as how DSM is integrated into it, and what are the specificities of our synthesizer. Sections 12.4.2 and 12.4.3 then evaluate the resulting synthesizer through two subjective tests involving a large number of listeners. Section 12.4.2 compares DSM to the traditional pulse excitation on 5 English and French voices. A comparison with both the pulse excitation and STRAIGHT on two English speakers is provided in Section 12.4.3.

12.4.1 HMM speech synthesis based on DSM

HMM-based speech synthesis aims at generating natural sequences of speech parameters directly from a statistical model, which is previously trained on a given speech database [33]. The general framework of a HMM-based speech synthesizer is displayed in Figure 12.5. Two main steps can be distinguished in this process: training and synthesis.

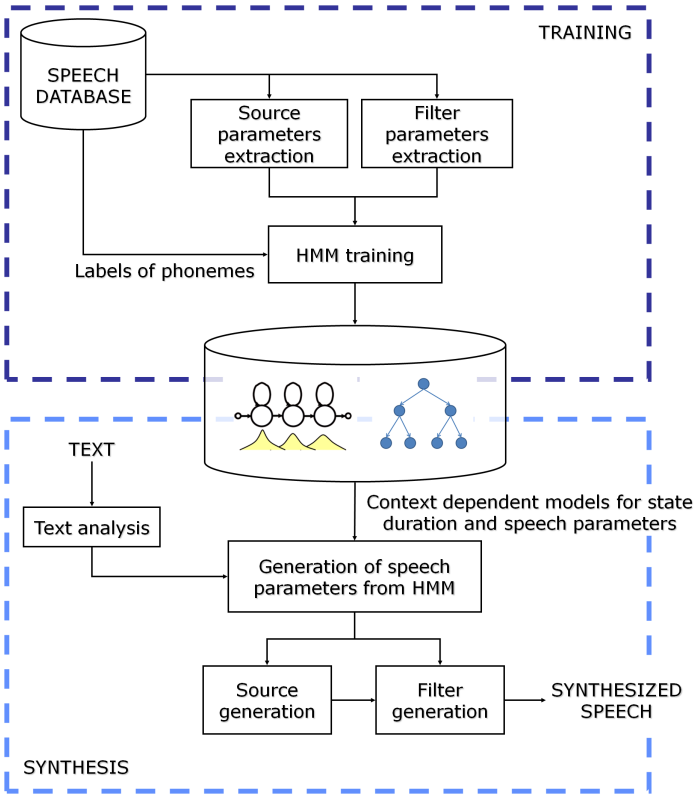

Figure 12.5 - *Framework of a HMM-based speech synthesizer (adapted from [33]).*

The **training** step assumes that a large segmented speech database is available. Labels consist of a phonetic environment description. First, both excitation (source) and spectral (filter) parameters are extracted from the speech signals. Since source modeling may be composed of either continuous values or a discrete symbol (respectively during voiced and unvoiced regions), Multi-Space probability Density

(MSD) HMMs have been proposed [34], as this approach is able to model sequences of observations having a variable dimensionality. Given the speech parameters and the labels, HMMs are trained using the Viterbi and Baum-Welch re-estimation algorithms [33]. Decision tree-based context clustering is used to statistically model data appearing in similar contextual situations. Indeed contextual factors such as stress-related, locational, syntaxical or phonetic factors affect prosodic (duration and source excitation characteristics) as well as spectral features. More precisely an exhaustive list of possible contextual questions is first drawn up. Decision trees are then built for source, spectrum and duration independently using a maximum likelihood criterion. Probability densities for each tree leaf are finally approximated by a Gaussian mixture model.

At **synthesis** time, the input text is converted into a sequence of contextual labels using a Natural Language Processor. From them, a path through the context-dependent HMMs is computed using the duration decision tree. Excitation and spectral parameters are then generated by maximizing the output probability. The incorporation of dynamic features (Δ and Δ^2) makes the coefficients evolution more realistic and smooth [35]. The generated parameters are then the input of the vocoder, which produces the synthetic waveform.

The implementation of our HMM-based speech synthesizer relies on the HTS toolkit publicly available in [36]. As mentioned in Section 12.2, the only excitation feature used for the training is F_0. A five-state left-to-right multistream HMM is used. More precisely, four separate streams are employed: *i)* one single Gaussian distribution with diagonal covariance for the spectral coefficients and their derivatives, *ii)* one MSD distribution for pitch, *iii)* one MSD distribution for pitch first derivative, and *iv)* one MSD distribution for pitch second derivative. In each MSD distribution, for voiced parts, parameters are modeled by single Gaussian distributions with diagonal covariance, while the voiced/unvoiced decision is modeled by an MSD weight. As HMMs are known for oversmoothing the generated trajectories [32], the Global Variance technique [32] is used to alleviate this effect. The generated parameters are then fed into the vocoder described in Section 12.2.

12.4.2 First Evaluation

In this first experiment, the DSM vocoder is compared to the traditional *Pulse* excitation, whose workflow is summarized in Figure 12.6. This technique basically uses either a pulse train during voiced speech, or white noise during unvoiced parts. As for the DSM vocoder, the resulting excitation signal is then the input of the MLSA filter fed by the MGC coefficients.

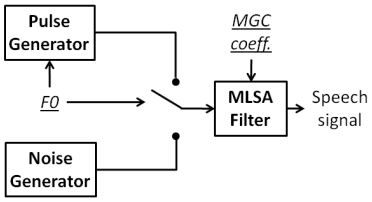

Figure 12.6 - *Workflow of the vocoder using the traditional* Pulse *excitation. Input features (indicated in italic and underlined) are the target pitch F_0 (which also control the switch between voiced and unvoiced excitation) and the MGC filter coefficients.*

Experimental Protocol

The synthetic voices of five speakers are assessed: AWB (Scottish male), Bruno (French male), Julie (French female), Lucy (US female) and SLT (US female). AWB and SLT come from the publicly

available CMU ARCTIC database [24] and about 45 minutes of speech for each were used for the training. Other voices were kindly provided by Acapela Group and were trained on a corpus of around 2 hours. The test consists of a subjective comparison between the proposed and the traditional pulse excitation. For this, 40 people participated to a CMOS test composed of 20 randomly chosen sentences of about 7 seconds. For each sentence they were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference (see the CMOS scale in Table 12.1).

Results

Preference scores can be viewed in Figure 12.7. A clear improvement over the traditional pulse excitation can be observed for all voices. Indeed, DSM was preferred between 78% and 94% of cases, depending on the considered voice, while the proportion of preference for Pulse did not exceed 8%. Compared to the method using a pitch-synchronous residual codebook we proposed in [5], results are almost similar on male speakers, with a minor loss of less than 5% for both AWB and Bruno. On the contrary, the contribution on female voices is much more evident. While only 30% of participants preferred the technique using the codebook for speaker SLT [5], score now reaches more than 90% for the DSM. This trends holds for other female voices. Figure 12.8 exhibits the average CMOS scores with their 95% confidence intervals. Average values vary between 1 and 1.75, confirming a clear significant advantage for the proposed technique.

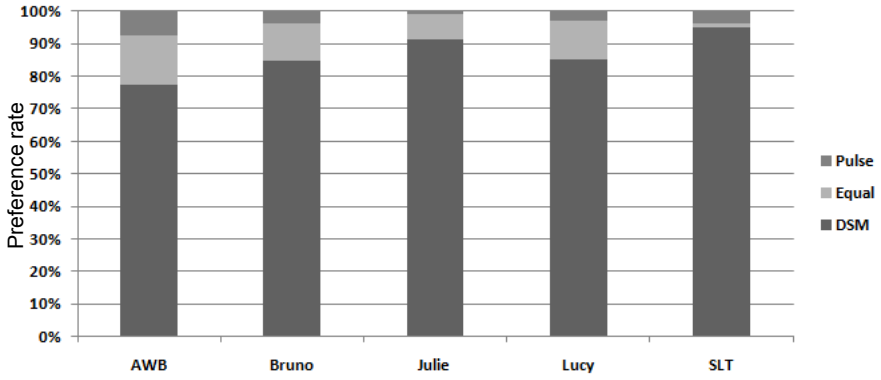

Figure 12.7 - *Preference score for the five speakers.*

12.4.3 Second Evaluation

In this second experiment, the proposed DSM is compared to three other state-of-the-art excitation models for HMM-based speech synthesis purpose. The first method is the traditional *Pulse* excitation, used by default in the HTS toolkit [36], and whose workflow is displayed in Figure 12.6.

The second method, called Glottal Post-Filtering (GPF), was proposed in [37] and aims at combining the LF model [14] with the spectral envelope of STRAIGHT. The workflow of the GPF vocoder is illustrated in Figure 12.9. Basically, it consists of transforming the LF model signal into a spectrally flat signal. The resulting signal can be used to synthesise speech instead of the impulse train. Although the excitation obtained using GPF does not represent the glottal source signal, this excitation is expected to produce more natural speech than the impulse train [37]. This improvement is explained by the fact that the spectrum of this excitation has a harmonic structure less periodic than that of the impulse train spectrum, which reduces the buzziness of the synthetic speech.

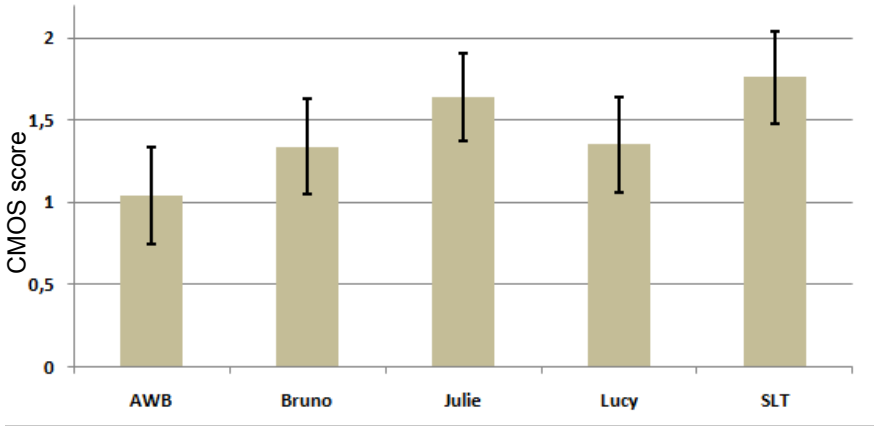

Figure 12.8 - *Average CMOS score in advantage of the DSM for the five speakers, together with their 95% confidence interval.*

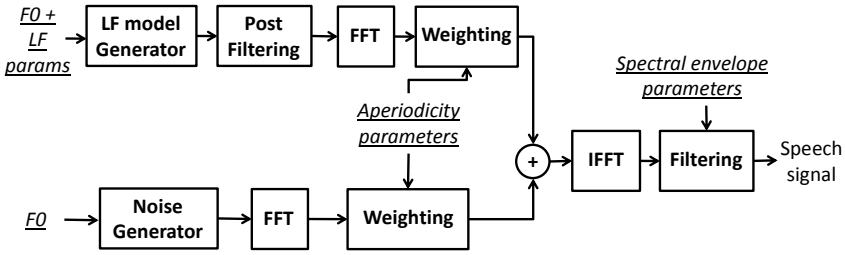

Figure 12.9 - *Workflow of the GPF vocoder. Input features (indicated in italic and underlined) are, for the excitation, the target pitch F_0, the LF model parameters and the aperiodicity coefficients, and the spectral envelope parameters for the filter.*

The third method is the *STRAIGHT* vocoder, known for its high-quality representation of the speech signal. STRAIGHT makes use of a specific spectral envelope obtained via a pitch-adaptive time-frequency smoothing of the FFT speech spectrum. As for the excitation modeling, STRAIGHT relies on aperiodic measurements in five spectral subbands: [0-1], [1-2], [2-4], [4-6] and [6-8] kHz. As a consequence, the excitation features used by the HMM synthesizer now include the 5 aperiodic measurements besides F_0. This results in an additional HMM stream composed of these aperiodicity parameters, together with their first and second derivatives. Once generated, the speech features are the input of the STRAIGHT vocoder presented in Figure 12.10. The source signal is a Mixed Excitation whose periodic and aperiodic components are weighted by the aperiodicity measures. As suggested in [11], the phase of the periodic contribution is manipulated so as to reduce buzziness. Both components are then added, and passed through a minimum-phase filter obtained from the parameters describing the smooth STRAIGHT spectral envelope.

Experimental Protocol

The synthetic voices of two UK English speakers were assessed. The first is a male speaker who recorded about ten hours, while the second is a female speaker with about four hours of speech. The HMM-based speech synthesizers were trained, for both voices, on the whole corpus. This was carried out, as explained in Section 12.4.1, by the Centre for Speech Technology Research of Edinburgh, which kindly

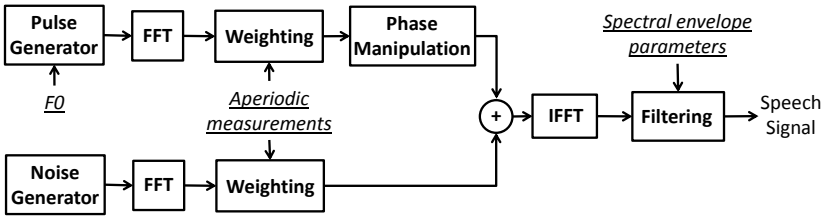

Figure 12.10 - *Workflow of the STRAIGHT vocoder. Input features (indicated in italic and underlined) are, for the excitation, the target pitch F_0 and the 5 aperiodic measurements, and the spectral envelope parameters for the filter.*

provided us the generated parameters[1]. All details about the training and the parameter generation can be found in [37], as well as other experiments with other excitation models.

The test consists of a subjective comparison between the proposed DSM and both the traditional pulse excitation and STRAIGHT. More precisely, we performed a CMOS test composed of 23 sentences, among which the first 3 were provided for calibration. These utterances were randomly chosen out of a set of 120 sentences, half for each speaker. For each sentence, participants were asked to listen to both versions (DSM versus Pulse or STRAIGHT, randomly shuffled) and to attribute a CMOS score according to their overall preference. Participants were divided into two categories: 26 speech experts (i.e people familiar with speech processing), and 34 naive listeners. The test was conducted through the Web.

Results

Results of the CMOS test are exhibited in Table 12.2, and are separated for the two categories of participants. First, it is observed that speech experts significantly perferred DSM over the pulse excitation, with a CMOS score of a bit more than 1.2 for both the male and the female speaker. A similar conclusion can be drawn for the naive listeners, although their averaged CMOS scores are around 0.75 instead of 1.2. As a matter of fact, we observed that naive listeners used the whole CMOS scale in a lesser extent. Indeed, since the only change between the two versions only concerns the excitation modeling (as spectral envelope and prosody were kept unchanged), auditive differences were relatively subtle. It can then be understood that speech experts noticed them more easily.

DSM is also observed to outperform GPF for both categories of listeners, and this to a greater extent for male speakers. Averaged CMOS scores for GPF reached around 75% of those obtained with the pulse excitation for male voices, and about 50% for female speakers. Regarding the comparison with STRAIGHT, it turns out that both methods were found, in average, to deliver a comparable quality. Although speech experts very slightly preferred DSM, the opposite is noted for naive listeners. But taking the 95% confidence intervals into account, no significant advantage for DSM over STRAIGHT, or vice versa, can be highlighted.

In complement to the CMOS scores, Tables 12.3 and 12.4 present the preference results for all test conditions, respectively for the male and female speakers. While speech experts preferred DSM to Pulse in about 75% of cases, this proportion is reduced to around 60% for naive listeners. Nevertheless, the advantage of DSM over Pulse is clear again, as Pulse was only preferred in very few cases. DSM is also noticed to give a better performance than GPF. Indeed only about 15% of speech experts preferred GPF over DSM. On naive listeners, this rate reached 19.5% for male speakers, and 26.8% for female

[1]We are very grateful to Dr. Joao Cabral and Prof. Steve Renals for their precious help.

Speech Experts	Male Speaker	Female Speaker
DSM vs Pulse	**1.205** ± 0.198	**1.241** ± 0.209
DSM vs GPF	**0.840** ± 0.202	**0.655** ± 0.256
DSM vs STRAIGHT	**0.167** ± 0.217	**0.037** ± 0.197
Naive listeners	Male Speaker	Female Speaker
DSM vs Pulse	**0.75** ± 0.196	**0.722** ± 0.188
DSM vs GPF	**0.59** ± 0.176	**0.363** ± 0.181
DSM vs STRAIGHT	**-0.010** ± 0.164	**-0.072** ± 0.201

Table 12.2 - *Average CMOS scores together with their 95 % confidence intervals, for both speech experts and naive listeners.*

voices, confirming a clear advantage in favor of DSM. Regarding the comparison with STRAIGHT, preference results confirm that both methods are almost equivalent. Indeed it is seen in Table Tables 12.3 and 12.4 that the repartition between the three categories is almost one third, reflecting the fact that both methods lead to a similar quality. However, one advantage of DSM over STRAIGHT is that it does not require the addition of a specific stream in the HMM-based synthesizer, making not only the training step lighter, but also more importantly alleviating the computational footprint at running time.

Speech Experts	DSM preferred	Equivalent	Other method preferred
DSM vs Pulse	76.07 %	18.80 %	5.13 %
DSM vs GPF	63.50 %	21.90 %	14.60 %
DSM vs STRAIGHT	33.33 %	40.35 %	26.32 %
Naive listeners	DSM preferred	Equivalent	Other method preferred
DSM vs Pulse	59.38 %	24.38 %	16.24 %
DSM vs GPF	54.66 %	29.19 %	16.15 %
DSM vs STRAIGHT	31.22 %	33.86 %	34.92 %

Table 12.3 - *Preference scores for the **male speaker**, for both speech experts and naive listeners.*

Speech Experts	DSM preferred	Equivalent	Other method preferred
DSM vs Pulse	73.68 %	18.80 %	7.52 %
DSM vs GPF	55.75 %	24.78 %	19.47 %
DSM vs STRAIGHT	35.77 %	32.85 %	31.39 %
Naive listeners	DSM preferred	Equivalent	Other method preferred
DSM vs Pulse	62.78 %	16.11 %	21.11 %
DSM vs GPF	46.93 %	26.26 %	26.82 %
DSM vs STRAIGHT	30.47 %	33.11 %	36.42 %

Table 12.4 - *Preference scores for the **female speaker**, for both speech experts and naive listeners.*

In Section 12.4.2, the first evaluation compared DSM and Pulse for 5 English and French voices, and the CMOS test was submitted to 40 people, among them both speech experts and naive listeners. Since the data, the synthesizer itself and the test conditions are not the same, results are obviously not directly comparable. However, the conclusions drawn from these two experiments both report the

overwhelming advantage of DSM over Pulse in speech synthesis. This superiority was even stronger in Section 12.4.2 where the averaged CMOS scores varied between 1 and 1.8 across the 5 voices, and the preference rates for DSM between 78% and 94%.

12.5 Conclusion

This chapter aimed at applying the DSM of the residual signal introduced in Chapter 11 to speech synthesis. For this, the DSM vocoder was first presented. The resulting method was then evaluated within two applicative contexts: pitch modification and HMM-based speech synthesis. Regarding the pitch modifications capabilities, DSM was compared to 3 other well-known state-of-the-art techniques for this purpose: TDPSOLA, HNM and STRAIGHT. A subjective test showed that DSM clearly outperforms TDPSOLA for female speakers and HNM on male voices. DSM was also observed to be slightly better than STRAIGHT on male speakers, while STRAIGHT obtained higher results on female voices.

In a second application, the DSM vocoder was integrated into a HMM-based speech synthesizer. The quality delivered by the resulting synthesizer was compared to the traditional pulse excitation, the GPF and the STRAIGHT methods. Two subjective comparative evaluations involving a large number of participants, among which speech experts and naive listeners, were performed. In all cases, results showed a significant preference for DSM over the pulse excitation, this advantage being clearer for speech experts. DSM was also shown to provide a better quality than the GPF excitation modeling. As for the comparison with STRAIGHT, both techniques turned out to lead to similar quality for both male and female voices.

Bibliography

[1] A.J Hunt and A.W Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 373–376, 1996.

[2] A. Black, H. Zen, and K. Tokuda. Statistical parametric speech synthesis. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1229–1232, 2007.

[3] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. *IEEE Workshop on Speech Synthesis*, pages 227–230, 2002.

[4] M. Schroeder and B. Atal. Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 10, pages 937–940, 1985.

[5] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit. Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[6] H. Yang, S. Koh, and P. Sivaprakasapillai. Enhancement of multiband excitation (MBE) by pitch-cycle waveform coding. In *Electronics Letters*, volume 30, pages 1645–1646, 1994.

[7] T. Dutoit and H. Leich. Mbr-psola : Text-to-speech synthesis based on an mbe re-synthesis of the segments database. *Speech Communication*, 13:435–440, 1993.

[8] W. Lin, S. Koh, and X. Lin. Mixed excitation linear prediction coding of wideband speech at 8kbps. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1137–1140, 2000.

[9] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura. Mixed-excitation for hmm-based speech synthesis. In *Eurospeech*, pages 2259–2262, 2001.

[10] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda. An excitation model for HMM-based speech synthesis based on residual modeling. In *ISCA SSW6*, 2007.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207, 2001.

[12] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *6th ISCA Workshop on Speech Synthesis*, 2007.

[13] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. In *IEEE Trans. on Audio, Speech, and Language Processing*, volume 19, pages 153–165, 2011.

[14] G. Fant and J. Liljencrants Q. Lin. A four parameter model of glottal flow. In *STL-QPSR4*, pages 1–13, 1985.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. An adaptive algorithm for Mel-cepstral analysis of speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 137–140, 1992.

[16] Y. Stylianou. Voice transformation: a survey. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[17] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–205, 1995.

[18] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9:21–29, 2001.

[19] D. Vincent, O. Rosec, and T. Chovanel. A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[20] J. Laroche and M. Dolson. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 91–94, 1999.

[21] P. Depalle and G. Poirrot. SVP: A modular system for analysis, processing and synthesis of sound signals. In *Proc. International Computer Music Conference*, 1991.

[22] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Proc. Interspeech Conf.*, 2009.

[23] Online. STRAIGHT: a speech analysis, modification and synthesis system. In http://www.wakayama-u.ac.jp/ kawahara/STRAIGHTadv/index_e.html, .

[24] Online. Cmu arctic speech synthesis databases. In http://festvox.org/cmu_arctic/, .

[25] V. Grancharov and W. Kleijn. Speech quality assessment. In *Springer Handbook of Speech Processing*, 2007.

[26] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel generalized cepstral analysis Ů a unified approach to speech spectral estimation. In *ICSLP*, 1994.

[27] D. Kim. On the perceptually irrelevant phase information in sinusoidal representation of speech. *IEEE Trans. Speech Audio Processing*, 9:900–905, 2001.

[28] H. Zen, T. Toda, M. Nakamura, and T. Tokuda. Details of the nitech hmm-based speech synthesis system for the blizzard challenge 2005. *IEICE Trans. Inform. Systems*, E90-D (1):325–333, 2007.

[29] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech Conf.*, 2009.

[30] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi. Glottal spectral separation for parametric speech synthesis. In *Proc. Interspeech Conf.*, pages 1829–1832, 2008.

[31] Y. Agiomyrgiannakis and O. Rosec. Arx-lf-based source-filter methods for voice modification and transformation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3589–3592, 2009.

[32] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, 90(5):816–824, 2007.

[33] Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1229–1232, 2007.

[34] K. Tokuda, T. Masuko, N. Myiazaki, and T. Kobayashi. Multi-space probability distribution HMM. In *IEICE Trans. on Information and Systems*, volume E85-D, pages 455–464, 2002.

[35] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai. An algorithm for speech parameter generation from continuous mixture hmms with dynamic features. In *Eurospeech*, 1995.

[36] Online. HMM-based speech synthesis system (HTS). In http://hts.sp.nitech.ac.jp/, .

[37] Joao Cabral. HMM-based speech synthesis using an acoustic glottal source model, phd thesis. In *School of Informatics, University of Edinburgh*, 2010.

Chapter 13

Application of DSM to Speaker Recognition

Contents

Abstract

Most of current speaker recognition systems are based on features extracted from the magnitude spectrum of speech. However the excitation signal produced by the glottis is expected to convey complementary relevant information about the speaker identity. This chapter explores the use of glottal signatures, derived from the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in Chapter 11, for speaker identification. Experiments using these signatures are performed on both TIMIT and YOHO databases. Promising results are shown to outperform other state-of-the-art approaches based on glottal features.

This chapter is based upon the following publications:

- Thomas Drugman, Thierry Dutoit, *On the Potential of Glottal Signatures for Speaker Recognition*, Interspeech Conference, Makuhari, Japan, 2010.
- Thomas Drugman, Thierry Dutoit, *The Deterministic plus Stochastic Model of the Residual Signal and its Applications*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*

13.1 Introduction

Automatic speaker recognition refers to the use of a machine in order to recognize a person from a spoken phrase [1]. This task is then closely linked to the understanding of what defines the speaker individuality. Although high-level information (such as the word usage) could be of interest, low-level acoustic features are generally employed [1]. Such features are most of the time extracted from the amplitude spectrum of the speech signal. They aim at parameterizing the contribution of the vocal tract, which is an important characteristic of the speaker identity. On the other hand, very few works address the possibility of using features derived from the glottal source in speaker recognition. However significant differences in the glottal waveforms have been observed between different speaker types [2].

In [3], Thevenaz exploits the orthogonality of the LPC residue for text-independent speaker verification. In order to avoid synchronization with pitch epochs and simultaneously to get rid of the residual phase contribution, it was suggested to retain the residual amplitude spectrum. It is concluded that although the residue-based features are less informative than the vocal tract-based ones, they are nonetheless useful for speaker verification, and above all combine favourably with methods based on the LPC filter. In [4], Murty *et al.* demonstrate the complementarity of features based on the residual phase with the traditional MFCCs, commonly used in speaker recognition. Authors led speaker recognition experiments on the NIST-2003 database. By integrating the residual phase information in addition to the common MFCCs, they reported a reduction of equal error rate from 14% to 10.5%. In [5], Plumpe *et al.* focused on the use of the glottal flow estimated by closed phase inverse filtering. On the resulting glottal source, two types of features were extracted. The first ones are time-domain features, parameterizing both the coarse structure (obtained by fitting a LF model [6]) and the fine structure of the glottal flow derivative. The second ones are a Mel-cepstral representation of the glottal source. A clear advantage in favor of the cepstral coefficients was shown. In a similar way, Gudnason *et al.* focus in [7] on the use of Voice Source Cepstrum Coefficients (VSCCs) for speaker recognition. A process based on closed-phase inverse filtering, and which is shown to be robust to LPC analysis errors and low-frequency phase distortion, is proposed. When combined to traditional MFCCs, the resulting features are reported to lead to an appreciable improvement for speaker identification.

The goal of this chapter is to investigate the potential of using *glottal signatures* in speaker recognition. The research of an invariant *voiceprint* in the speech signal, univoquely characterizing a person (as achieved with the fingerprint), has always attracted the speech community [8]. As this seems utopian due to the inherent nature of the phonation mechanism, we here prefer the use of the term "*signature*" for denoting a signal conveying a relevant amount of information about the speaker identity.

It is here focused on the usefulness of glottal signatures derived from the Deterministic plus Stochastic Model (DSM) of the residual excitation, introduced in Chapter 11, for speaker recognition purpose. For this, we suggest to use speaker-dependent waveforms of the DSM: the eigenresiduals for the deterministic part (see Section 11.4), and the energy envelope for the stochastic contribution (see Section 11.5). It was also shown in Section 11.6 that about 1000 voiced frames are sufficient for a reliable estimation of these signatures. This means that about 7s of voiced speech for a male speaker, or 4s for a female voice, are sufficient for a reliable identification using these waveforms. Besides the estimation was shown in Section 11.7 to converge towards the same waveform independently of the considered phonetic class. The signatures could then be used in text-independent speaker recognition.

The chapter is structured as follows. Section 13.2 explains how the DSM-based waveforms are used for speaker identification purpose. In Section 13.3, the protocol used for our experiments is described. Section 13.4 presents our results on the large TIMIT database. First of all, the potential of the proposed waveforms is investigated, as well as the impact of the higher orders eigenresiduals. Then, speaker identification performance using the glottal signatures is assessed. Our experiments on

the YOHO database are reported in Section 13.5. This gives an idea of the inter-session sensitivity of the proposed technique. On both databases, some comparisons with other glottal-based speaker recognition approaches [5], [7] are provided.

13.2 Integrating Glottal Signatures in Speaker Identification

In order to be integrated into a speaker identification system, the proposed DSM-based signatures are estimated on both training and testing sets. A *distance matrix* $D(i,j)$ between speaker i (whose glottal signatures are estimated on the training dataset) and speaker j (estimated on the testing dataset) is then computed. In this work, the Relative Time Squared Error (RTSE) (see Equation 11.4) is chosen as a distance measure between two waveforms. Finally, the identification of a speaker i is carried out by looking for the lowest value in the i^{th} row of the distance matrix $D(i,j)$. The speaker is then correctly identified if the position of the minimum is i. In other words, when a new recording is presented to the system, the identified speaker is the one whose glottal signatures are the closest (in the Euclidian sense) to the signatures extracted on this recording.

In the following, it will be seen that no more than two glottal signatures are used for speaker identification. Many strategies are possible for combining their information and draw a final decision [9]. In this study, two strategies are considered: a weighted multiplication or a weighted sum. More precisely, denoting $D_x(i,j)$ and $D_y(i,j)$ the distance matrices using respectively the glottal signatures $x(n)$ and $y(n)$, the two sources of information are merged in our framework by calculating the final distance matrix $D(i,j)$ respectively as:

$$D(i,j) = D_x(i,j)^{\alpha} \cdot D_y(i,j)^{1-\alpha} \tag{13.1}$$

$$D(i,j) = \beta \cdot D_x(i,j) + (1-\beta) \cdot D_y(i,j) \tag{13.2}$$

where α and β are weights ranging from 0 to 1. They are used to possibly emphasize the importance of a given glottal signature with regard to the other. When the weight is 0, only $y(n)$ is considered, while a weight equal to 1 means that only $x(n)$ is used for identification.

13.3 Experimental Protocol

In this Section, the maximum voiced frequency F_m is fixed to 4 kHz (usual value for a modal voice quality, as shown in Section 11.3) and the normalized pitch value F_0^* is set to 100 Hz for all speakers. Experiments are carried out on both TIMIT and YOHO databases, for comparison purpose with [5] and [7]. In [5], Plumpe *et al.* reported speaker identification results on TIMIT using either Time-Domain features (TDGF) or a Mel-Cepstral (MCGF) representation of the estimated Glottal Flow. As for [7], Gudnason *et al.* performed tests on both TIMIT and YOHO using their proposed Voice Source Cepstrum Coefficients (VSCC). For both methods, classification was performed using an approach based on a Gaussian Mixture Model(GMM).

The TIMIT database [10] comprises 10 recordings from 630 speakers (438 males, 192 females) and sampled at 16 kHz. As for the YOHO database [11], it contains speech from 138 speakers (108 males, 30 females) sampled at 8 kHz. Since $F_s = 8kHz$ for YOHO, only the deterministic part of the DSM holds, and the unvoiced energy envelope cannot therefore be used for the recognition. Recordings of YOHO were collected in a real-world office environment through 4 sessions over a 3 month period. For each session, 24 phrases were uttered by each speaker.

In the following experiments, the data is split for each speaker (and each session for YOHO) into 2 equal parts for training and testing. This is done in order to guarantee that, for both steps, enough residual frames are available for reliably estimating the signatures (see Section 11.6). However, it is worth noting that although there was always a sufficient number of frames for YOHO, it happened for some low-pitched voices of the TIMIT database that only around 500 voiced frames were used for the training or the test. This consequently led to an imperfect estimation of the glottal signatures in such cases.

13.4 Results on the TIMIT database

13.4.1 Usefulness of the glottal signatures

To give a first idea on the potential of using the glottal signatures in speaker recognition, Figure 13.1 displays the distributions of $D_{\mu_1}(i,j)$ (i.e the distance matrix using only the first eigenresidual $\mu_1(n)$) respectively when $i = j$ and when $i \neq j$. In other words, this plot shows the histograms of the RTSE (in logarithmic scale) between the first eigenresiduals estimated respectively for the same speaker, and for different speakers. It is clearly observed that the error measure is much higher (about 15 times higher in average) when the tested signature does not belong to the considered speaker. It is also noticed that, for the same speaker, the RTSE on the eigenresidual is about 1%, which is of the same order of magnitude as for the inherent estimation process, confirming our results of Sections 11.6 and 11.7. However a weak overlap between both distributions is noted, which may lead to some errors in terms of speaker identification.

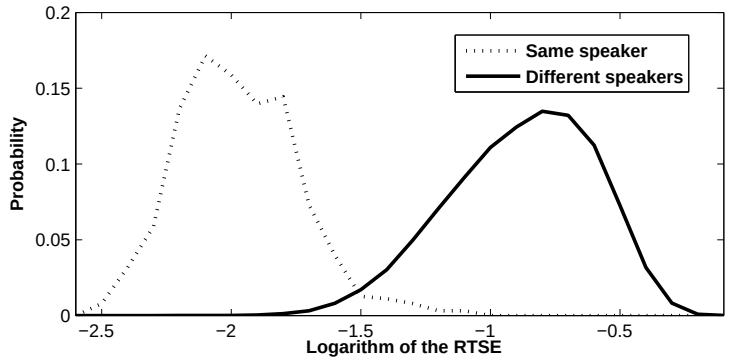

Figure 13.1 - *Distributions of the Relative Time Squared Error (RTSE) between the first eigenresiduals $\mu_1(n)$ estimated respectively for the same speaker and for different speakers.*

13.4.2 Effect of the higher order eigenresiduals

It was mentioned in Section 11.4, that only considering the first eigenresidual is sufficient for a good modeling of the residual signal below F_m, and that the effect of higher order eigenresiduals is almost negligible in that spectral band. One could argue however that higher order waveforms can be useful for speaker recognition. Figure 13.2 shows the identification rate on the whole TIMIT database (630 speakers), for each eigenresidual $\mu_i(n)$. It is clearly observed that higher order eigenresiduals are less discriminative about the speaker identity. More particularly, the identification rate dramatically drops from 88.6% to 39.8% when going from the first to the second eigenresidual used individually.

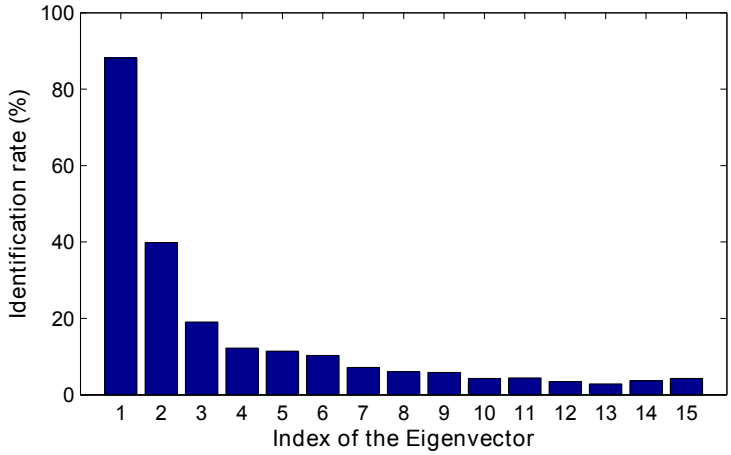

Figure 13.2 - *Speaker identification capability on the whole TIMIT database using* individually *the eigenresiduals of higher orders.*

In order to assess the contribution of higher order eigenresiduals, Figure 13.3 shows the evolution of the identification rate as a function of α and β, when the first and second eigenresiduals $\mu_1(n)$ and $\mu_2(n)$ are combined according to Equations (13.1) and (13.2). In both strategies, it turns out that considering $\mu_2(n)$ in addition to $\mu_1(n)$ does not bring anything, since optimal performance is reached for α=1 and β=1. Therefore, the effect of higher order eigenresiduals for speaker identification can be neglicted and only $\mu_1(n)$ is considered in the following experiments.

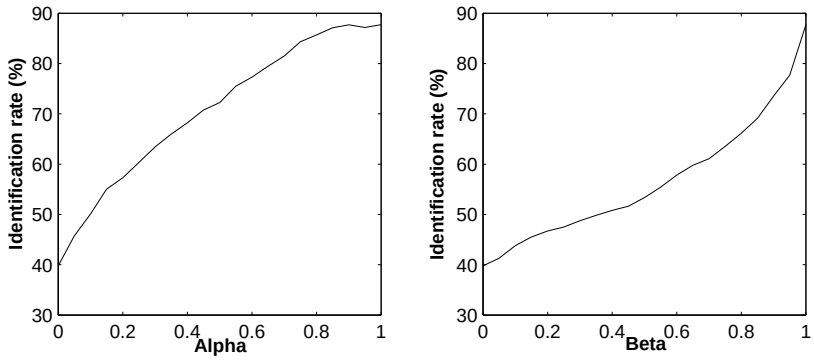

Figure 13.3 - *Evolution of the identification rate as a function of α and β, when* ***the first and second eigenresiduals ($\mu_1(n)$ and $\mu_2(n)$) are combined*** *according to Equations (13.1) and (13.2).*

13.4.3 Combining the eigenresidual and the energy envelope

Contrarily to higher order eigenresiduals, the energy envelope $e(n)$ of the stochastic part (see Section 11.5) showed a high discrimination power with an identification rate of 82.86% on the whole TIMIT database. It can then be expected that using the first eigenresidual $\mu_1(n)$ in complement to $e(n)$ could improve the performance. For this, they are combined as in Equations (13.1) and (13.2), and the influence of α and β is displayed in Figure 13.4. First, the advantage of using both signatures together is clearly confirmed. Secondly, the optimal performance using Eq. (13.1) or Eq. (13.2) is identical. In the rest of our experiments, we used Equation 13.1 with α=0.5 which, although slightly suboptimal in

this example, makes the combination as a simple element-by-element multiplication of $D_{\mu_1}(i,j)$ and $D_e(i,j)$.

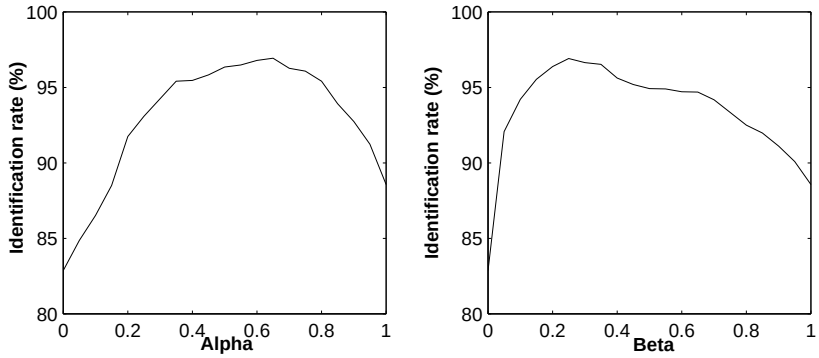

Figure 13.4 - *Evolution of the identification rate as a function of α and β, when* ***the first eigenresidual*** $\mu_1(n)$ ***and the energy envelope*** $e(n)$ ***are combined*** *according to Equations (13.1) and (13.2).*

13.4.4 Speaker identification results

Figure 13.5 exhibits the evolution of the identification rate with the number of speakers considered in the database. Identification was achieved using only one of the two glottal signatures, or using their combination as suggested in Section 13.2. As expected the performance drops as the number of speakers increases, since the risk of confusion becomes more important. However this degradation is relatively slow in all cases. One other important observation is the clear advantage of combining the information of the two signatures. Indeed this leads to an improvement of 7.78% compared to using only the first eigenresidual on the whole database.

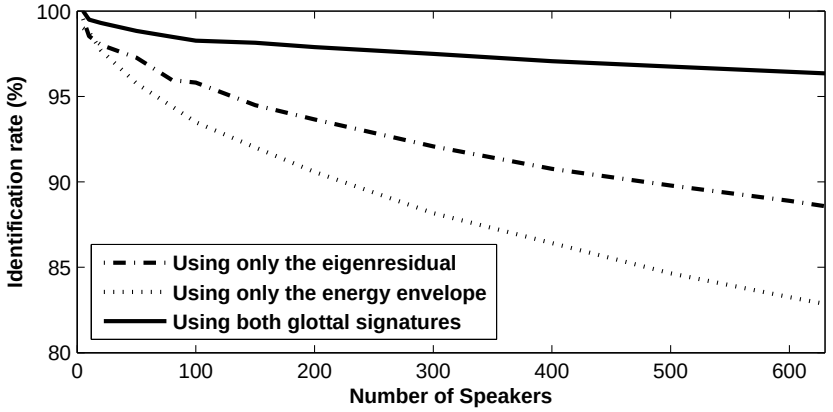

Figure 13.5 - *Evolution of the identification rate with the number of speakers for the TIMIT database.*

Table 13.1 summarizes the results obtained on the TIMIT database. Identification rates for 168 speakers are also given for comparison purpose. Using the time-domain parametrization of the glottal flow (TDGF), Plumpe *et al.* [5] reported an average misclassification rate of 28.65%. This result was importantly reduced to 4.70% by making use of the Mel-cepstral representation of the glottal flow (MCGF). On the same subset, Gudnason *et al.* reported in [7] a misclassification rate of 5.06%

using their proposed VSCC. These results can be compared to the 1.98% we achieved using the two glottal signatures. Finally also note that, relying on the VSCC, Gudnason *et al.* [7] obtained a misidentification rate of 12.95% on the whole TIMIT database (630 speakers). With the proposed signatures, a misclassification rate of 3.65% is reached. It is worth noting that no specific disparity bewteen male and female speakers was observed. More precisely, 6 out of the 192 female speakers (3.13%), and 17 out of the 438 male speakers (3.88%) were misclassified using the two glottal signatures.

	168 speakers	630 speakers
TDGF [5]	28.65	/
MCGF [5]	4.70	/
VSCC [7]	5.06	12.95
Using only the eigenresidual	5.88	11.43
Using only the energy envelope	8.76	17.14
Using both glottal signatures	**1.98**	**3.65**

Table 13.1 - *Misidentification rate (%) on the TIMIT database obtained using state-of-the-art glottal approaches or the proposed DSM-based signatures.*

13.5 Results on the YOHO database

As mentioned above, recordings in the YOHO database are sampled at 8 kHz, and therefore only the first eigenresidual is used for speaker identification. Besides, as the 4 sessions were spaced over 3 months, we evaluate here the inter-session variability of the proposed glottal signature. Table 13.2 reports our speaker identification results as a function of the period separating the training and testing sessions. In addition the proportions of cases for which the correct speaker is recognized in second or third position (instead of first position) are also given. When recordings are from the same session, an almost perfect performance is carried out, with 99.73% of correct identification. This is above the approximative 95% rate reached on TIMIT with the eigenresidual for the same number of speakers (see Figure 13.5). This might be explained by the greater amount of data available in YOHO for the estimation of the glottal signature.

On the contrary, when the test is performed one session later, the identification dramatically drops by 30%. This first degradation accounts for two phenomena: the mismatch between training and testing recording conditions, and the intra-speaker variability. It then turns out that the identification rate decreases of about 5% for any later session. This is mainly attributable to speaker variability, which increases with the period separating the two sessions. As future work, we plan to design a filter whose variable phase response has the ability to compensate the mismatch between different recording environments.

It is worth noting that when recording sessions differ, between 13% and 16% of speakers are identified in second or third position. By integrating a complementary source of information, such as the traditional features describing the vocal tract function, it can be expected that most of the ambiguity on these signatures will be removed. Finally note that Gudnason *et al.* reported in [7] an identification rate of 63.7% using the VSCC, but with test recordings coming from the 4 sessions. By averaging our results over all sessions, the use of only the eigenresidual leads to an identification rate of 71.1%, confirming the good performance of the DSM-based signatures for speaker recognition.

	First Position	Second Position	Third Position
Same session	99.73%	0.27%	0%
One session later	69.29%	7.88%	5.19%
Two sessions later	64.31%	8.83%	4.57%
Three sessions later	58.70%	11.78%	4.35%

Table 13.2 - *Proportion of speakers classified in first (correct identification), second and third position, when recordings are spaced over several sessions.*

13.6 Conclusion

This chapter investigated the potential of using glottal signatures for speaker recognition. These signatures were derived from the DSM of the residual signal proposed in Chapter 11, which is a new speaker-dependent excitation modeling. Their usefulness was studied on the large TIMIT database, and the identification carried out relying on DSM-based signatures was observed to outperform by large the use of other glottal-based features proposed in the literature. An identification error of only 3.65% was obtained with the 630 speakers of the TIMIT corpus using the two proposed glottal signatures. In a second test on the YOHO database, we evaluated the inter-session sensitivity of these signatures, highlighting the degradation due to a mismatch between recording conditions, and the intra-speaker variability.

Several improvements could be brought to the current approach. Indeed results were obtained using *only* the proposed glottal signatures. Regarding the evidence of a complementarity between excitation-based and vocal tract-based features ([4], [5], [7]), it is reasonable to expect that combining the proposed signatures with a conventional speaker recognition system (e.g. with a typical GMM-MFCC approach) would lead to an appreciable improvement. Secondly, applying some channel compensation could alleviate the mismatch between training and testing sessions. Indeed different recording conditions impose different characteristics to the speech signal. Among these, differences in phase response may dramatically affect the estimation of the signatures (since the information of the residual is essentially contained in its phase).

Bibliography

[1] D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 4072–4075, 2002.

[2] I. Karlsson. Glottal waveform parameters for different speaker types. In *STL-QPSR*, volume 29, pages 61–67, 1988.

[3] P. Thevenaz and H. Hugli. Usefulness of the LPC-residue in text-independent speaker verification. In *Speech Communication*, volume 17, pages 145–157, 1995.

[4] S. Murty and B. Yegnanarayana. Combining evidence from residual phase and MFCC features for speaker recognition. In *IEEE Signal Processing Letters*, volume 13, pages 52–55, 2006.

[5] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech Audio Process.*, 7(5): 569–576, September 1999.

[6] G. Fant and J. Liljencrants Q. Lin. A four parameter model of glottal flow. In *STL-QPSR4*, pages 1–13, 1985.

[7] J. Gudnason and M. Brookes. Voice source cepstrum coefficients for speaker identification. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4821–4824, 2008.

[8] L.G. Kersta. Voiceprint identification. *Nature*, 196:1253–1257, 1962.

[9] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1993.

[10] W. Fisher, G. Doddington, and K. Goudie-Marshall. The darpa speech recognition research database: Specifications and status. In *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[11] J. Campbell. Testing with the yoho cd-rom voice verification corpus. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 341–344, 1995.

Chapter 14

General Conclusion

Contents

14.1 Contributions of this thesis

This thesis presented several advances in the field of glottal analysis. Solutions proposed in this work were shown to provide appreciable improvements in various speech processing applications. The original contributions of this thesis are the following:

- **Pitch tracking**

  The *Summation of Residual Harmonics* (SRH) method has been proposed in Chapter 2 for robust pitch tracking. A comparison with six state-of-the-art pitch trackers was performed in both clean and noisy conditions. A clear advantage of the proposed approach was its robustness to additive noise. In 9 out of the 10 noisy experiments (5 noise types at 0dB of Signal-to-Noise Ratio, for male and female speakers), SRH was shown to lead to a significant improvement, while its performance was comparable to other techniques in clean conditions.

- **Glottal closure instant detection**

  The *Speech Event Detection using the Residual Excitation And a Mean-based Signal* (SEDREAMS) algorithm has been proposed in Chapter 3 for accurate, reliable and robust Glottal Closure Instant (GCI) detection. SEDREAMS was compared to four of the most effective methods on a total amount of data of approximately four hours. In the experiments on clean speech, SEDREAMS provided the best results, reaching on all databases an identification rate greater than 98% while more than 80% of GCIs were located with an accuracy of 0.25 ms. In a second experiment, robustness to additive noise, as well as to reverberation, was investigated. SEDREAMS was shown to have the highest robustness, with an almost unchanged reliability. Another advantage of SEDREAMS is that it allows a very fast implementation.

- **Source-tract separation**

  The *Complex Cepstrum-based Decomposition* (CCD) has been proposed in Chapter 5 as a non-parametric approach for source-filter deconvolution. CCD allows the glottal flow estimation

directly from the speech waveform, as an alternative to the Zeros of the Z-Transform (ZZT) algorithm. Both techniques were shown to be functionally equivalent to each other, while the complex cepstrum is advantageous for its much higher speed, making it suitable for real-time applications. Windowing effects were studied in a systematic way on synthetic signals. It was emphasized that windowing plays a crucial role. More particularly we derived a set of constraints the window should respect so that the windowed signal matches the mixed-phase model. The potential of CCD was then confirmed by analyzing a large corpus of real speech containing various voice qualities. Interestingly some significant differences between the voice qualities were observed in the excitation.

- **Comparative evaluation of glottal flow estimation methods**

  Chapter 6 objectively compared the effectiveness of the main state-of-the-art techniques for glottal flow estimation: CCD, the Closed Phase Inverse Filtering (CPIF), and the Iterative Adaptive Inverse Filtering (IAIF) methods. Thorough tests were performed on both synthetic and real speech. Our first conclusion was that the usefulness of the NAQ, H1-H2 and HRF features is confirmed for parameterizing the glottal flow. We also confirmed other works in the literature showing that these parameters can be effectively used as measures for discriminating different voice qualities. Our results showed that the effectiveness of CPIF and CCD appears to be similar and rather high, with a slight preference towards CCD. However, it should be emphasized here that in our real speech tests, clean signals were used; for applications requiring the analysis of noisy signals (such as telephone applications) further testing is needed. The impact of factors such as the fundamental or formant frequencies on the estimation performance was also studied.

- **Asynchronous glottal flow estimation**

  The *Chirp Mixed-Phase Decomposition* has been proposed in Chapter 7 for the *asynchronous* estimation of the glottal flow. This was made possible by an extension of the framework used by both the ZZT and CCD methods. We also suggested an automatic way to carry out this decomposition on real speech. The resulting method was shown to be much more robust to GCI location errors than its traditional (non chirp) equivalent. Interestingly a reliable estimation of the glottal flow was obtained in an asynchronous way on real connected speech. Thanks to its low computational load, the chirp CCD method is then suited for being incorporated within a real-time asynchronous speech processing application.

- **Voice pathology detection**

  A set of new glottal and phase-based features has been proposed in Chapter 8 for the automatic detection of voice pathologies. The resulting extracted features were assessed through mutual information-based measures. This allowed their interpretation in terms of discrimination power and redundancy. It was shown that speech and glottal-based features are relatively complementary, while they present some synergy with prosodical characteristics. It was also shown that representations based on group delay functions are particularly suited for capturing irregularities in the speech signal. The adequacy of the mixed-phase model during voice production was discussed and shown to convey relevant information. Integrated within a classifier, the proposed features also led to an interesting improvement in terms of detection rate.

- **Glottal-based analysis of expressive voices**

  Chapter 9 confirmed and quantified, on large corpora, how the glottal source is modified during the production of expressive speech. First, we focused on the glottal analysis of Lombard

speech. Through an analysis on a database containing 25 speakers uttering in quiet and noisy environments, it was shown that the glottal source is considerably modified in Lombard speech. These variations, studied for several noise levels and types, have to be taken into account in applications such as speech or speaker recognition systems. In a second experiment, speech with various degrees of articulation has been considered. The acoustic analysis investigated changes related to the vocal tract as well as the glottis. It was shown that hyperarticulated speech is characterized by a larger vocalic space (more efforts to produce speech, with maximum clarity), higher fundamental frequency, a glottal flow containing a greater amount of high frequencies and a higher glottal formant frequency. These conclusions are of interest for being applied in applications such as expressive/emotional speech recognition/labeling or synthesis.

- **Parametric speech synthesis**

  The Deterministic plus Stochastic Model (DSM) of the residual signal has been proposed in Chapter 11 for modeling the source excitation. DSM has been integrated into a vocoder in Chapter 12 for improving the quality delivered by parametric speech synthesizers. The resulting method was then evaluated within two applicative contexts: pitch modification and HMM-based speech synthesis. Regarding the pitch modifications capabilities, DSM was compared to 3 other well-known state-of-the-art techniques for this purpose: TDPSOLA, HNM and STRAIGHT. A subjective test showed that DSM clearly outperforms TDPSOLA for female speakers and HNM on male voices. DSM was also observed to be slightly better than STRAIGHT on male speakers, while STRAIGHT obtained higher results on female voices. In a second application, the DSM vocoder was integrated into a HMM-based speech synthesizer. The quality delivered by the resulting synthesizer was compared to the traditional pulse excitation and the STRAIGHT method. Two subjective comparative evaluations involving a large number of participants, among which speech experts and naive listeners, were performed. In all cases, results showed a significant preference for DSM over the pulse excitation, this advantage being clearer for speech experts. As for the comparison with STRAIGHT, both techniques turned out to lead to similar quality for both male and female voices. However, one advantage of DSM over STRAIGHT is that it makes the training step lighter and more importantly alleviates the computational footprint at running time.

- **Speaker recognition**

  A new approach for speaker recognition has been proposed in Chpater 13. This technique is based on glottal signatures derived from the proposed DSM of the residual signal. The usefulness of these signatures was studied on the large TIMIT database, and the recognition results they carried out were observed to outperform by large the use of other glottal-based features proposed in the literature. An identification error of only 3.65% was obtained with the 630 speakers of the TIMIT corpus using the two proposed glottal signatures. In a second test on the YOHO database, we evaluated the inter-session sensitivity of these signatures, highlighting the degradation due to a mismatch between recording conditions, and the intra-speaker variability.

14.2 Perspectives

We have presented several new speech analysis tools. Each of the proposed approaches has been evaluated and compared on a large amount of data with various state-of-the-art techniques for the same purpose. We can therefore consider that the efficiency of these methods has been confirmed

through a comprehensive assessment, that no further validation is necessary, and that only minor improvements could possibly be brought to them.

These tools have been incorporated within several speech processing applications, as it was illustrated in Figure 1.5. However, the scope of such methods is not strictly limited to speech processing, and could be of interest in various other fields of signal processing. The possible further investigations ensuing from the work presented in this thesis can be summarized as follows:

- **Voice Disorder Detection**

  New features have been proposed in Chapter 8 for automatic voice pathology detection, where their usefulness was emphasized through a preliminary study. Further work might include their assessment on connected speech (without limiting the data to sustained vowels), and their use for the recognition, quantification and qualification of various voice disorders.

- **Expressive Voice Analysis**

  Chapter 9 focused on the glottal analysis of expressive speech. Lombard and hyper/hypo articulated speech have been studied. A comparable analysis could obviously be extended to other types of expressive voice, e.g speech presenting affects such as anger, boredom, disgust, anxiety, happiness or sadness. Such a framework could be also integrated into a system for emotion recognition, useful for example for enhanced human-computer interactions, or for automatically detecting events in public places in a surveillance application.

- **Speech Synthesis**

  A new excitation model called DSM has been shown in Chapter 11 to significantly enhance the quality delivered by parametric speech synthesizers. Its efficiency was confirmed through several subjective experiments. As perspectives, let us mention the possible improvements of DSM in voice conversion or modification. Indeed, this requires to process not only characteristics related to the vocal tract response, but also features derived from the source signal. This is also true for generating expressive speech, for which one could rely on the conclusions drawn in Chapter 9. An important advantage of DSM is that it only requires a very small amount of data (a few seconds of speech), which is particularly convenient for such applications. At last, it would be worth developing a speech synthesizer exploiting the Complex Cesptrum Decomposition proposed in Chapter 5, since this could allow the distinct characterization of the glottal flow and the vocal tract configuration, as physiologically motivated.

- **Speaker Recognition**

  A preliminary evaluation of a speaker identification system using glottal signatures derived from DSM has been given in Chapter 13. However, several improvements could be brought to the current approach. Indeed results were obtained using *only* the proposed glottal waveforms. Regarding the evidence of a complementarity between excitation-based and vocal tract-based features, it is reasonable to expect that combining the proposed signatures with a conventional speaker recognition system would lead to an appreciable improvement. Secondly, applying some channel compensation could alleviate the mismatch between training and testing conditions, consequently enhancing the system inter-session robustness.

- **Speech Recognition**

  As illustrated in Figure 1.5, the applicability of glottal analysis methods to speech recognition has not been addressed in this thesis. According to the author's point of view, this is probably the

field of speech processing for which such methods have the less promising contribution. Indeed, it was shown in Chapter 11 that the vocal tract can be assumed to be sollicited by a comparable excitation for all voiced phonetic classes. Besides the recognition of the pronounced message is independent of the produced voice quality. Glottal features would consequently be not suited for discriminating between different voiced phonemes. The only useful glottal information for speech recognition would therefore be the presence of voicing in the signal.

- **Speech Enhancement**

  Speech enhancement refers to the cleaning process which aims at reducing the presence of noise in a corrupted signal, or the task of enhancing its intelligibility. As seen in Figure 1.5, this issue is not explicitly tackled within the frame of this thesis. Nevertheless, methods of pitch tracking and GCI detection have been shown in Chapters 2 and 3 to maintain high performance even under adverse conditions. Thanks to their robustness, these approaches could therefore be of interest for speech enhancement. Also, if it can be assumed that one can reliably estimate the DSM components on degraded speech, using the DSM at synthesis time could strongly reduce the phase perturbations induced by the noise.

- **Music, Audio and Biomedical Signal Processing**

  Techniques described in this thesis have been designed for the speech signal, most of the time relying on some phonation properties. Nonetheless, these tools are extrapolable to other unidimensional signals, such as music, audio or biomedical signals. In this way, a slightly modified version of the proposed methods for pitch tracking and GCI detection could be advantageous for applications where an accurate and reliable synchronization of periodic signals is necessary. In addition, the SEDREAMS algorithm detailed in Chapter 3 could be beneficial for locating particular events in the considered signal. As an illustration, this could allow the detection of abnormalities in audio recordings (typically for surveillance purpose) or in biomedical signals (typically for patient monitoring). Besides, since the proposed methods are known to be robust, they are suited for biomedical applications where the use of some sensors inexorably leads to the inherent capture of parasitical signals.

  Finally, for some instruments or for singing voice, mechanisms of production are comparable to those involved during speech generation. Such mechanisms can be modeled by a source-filter approach. It would be therefore interesting to investigate the potential of using algorithms suggested in this thesis for the analysis, and even the synthesis of such music signals.

Appendix A

Calculation of the radius modifying a Blackman window for a chirp analysis

In Section 7.2.1, it is argued that for a Blackman window, the radius R necessary to modify its shape so that its new maximum lies in position t^* ($< L$) is expressed as:

$$R = exp[\frac{2\pi}{L} \cdot \frac{41\tan^2(\frac{\pi t^*}{L}) + 9}{25\tan^3(\frac{\pi t^*}{L}) + 9\tan(\frac{\pi t^*}{L}))}]. \tag{A.1}$$

This latter expression can be demonstrated as follows. Let us consider a Blackman window $w(t)$ of length L starting in $t = 0$:

$$w(t) = 0.42 - 0.5 \cdot \cos(\frac{2\pi t}{L}) + 0.08 \cdot \cos(\frac{4\pi t}{L}). \tag{A.2}$$

It is also known from Equation 7.3 that the evaluation of the chirp z-transform of a signal $x(t)$ on a circle of radius R is equivalent to evaluating the z-transform of $x(t) \cdot exp(log(1/R) \cdot t)$ on the unit circle. The new window $w_2(t)$ modified for a chirp analysis of radius R can then be written as:

$$w_2(t) = w(t) \cdot exp(log(1/R) \cdot t). \tag{A.3}$$

For a radius R, its maximum is in position t^* such that $w_2'(t^*) = 0$, where $w_2'(t)$ can be expressed as:

$$w_2'(t) = w'(t) \cdot exp(log(1/R) \cdot t) + w(t) \cdot log(1/R) \cdot exp(log(1/R) \cdot t) \tag{A.4}$$

$$= exp(log(1/R) \cdot t) \cdot [w'(t) + w(t) \cdot log(1/R)] \tag{A.5}$$

In t^*, we have $w_2'(t^*) = 0$ and consequently:

$$w'(t^*) + w(t^*) \cdot log(1/R) = 0, \tag{A.6}$$

$$R = exp(\frac{w'(t^*)}{w(t^*)}). \tag{A.7}$$

Denoting $\alpha = \frac{\pi t}{L}$, the first derivative of the Backman window function can be written as:

$$w'(t) = 0.5 \cdot \frac{2\pi}{L}\sin(2\alpha) - 0.08 \cdot \frac{4\pi}{L}\sin(4\alpha) \tag{A.8}$$
$$= \frac{\pi}{L}[\sin(2\alpha) - 0.32 \cdot \sin(4\alpha)] \tag{A.9}$$
$$= \frac{\pi}{L}[\sin(2\alpha) - 0.64 \cdot \sin(2\alpha) \cdot \cos(2\alpha)] \tag{A.10}$$
$$= \frac{\pi}{L} \cdot \sin(2\alpha) \cdot [1 - 0.64 \cdot \cos(2\alpha)] \tag{A.11}$$

which, denoting $\theta = \tan(\alpha)$, becomes:

$$w'(t) = \frac{\pi}{L} \cdot \frac{2\theta}{1+\theta^2} \cdot [1 - 0.64 \cdot \frac{1-\theta^2}{1+\theta^2}] \tag{A.12}$$
$$= \frac{\pi}{L} \cdot \frac{2\theta}{1+\theta^2} \cdot \frac{1+\theta^2 - 0.64 \cdot (1-\theta^2)}{1+\theta^2} \tag{A.13}$$
$$= \frac{\pi}{L} \cdot \frac{2\theta}{1+\theta^2} \cdot \frac{0.36 + 1.64 \cdot \theta^2}{1+\theta^2} \tag{A.14}$$
$$= \frac{\pi}{L} \cdot \frac{0.72 \cdot \theta + 3.28 \cdot \theta^3}{(1+\theta^2)^2} \tag{A.15}$$

With the same notation, the Blackman window function $w(t)$ can now be expressed as:

$$w(t) = 0.42 - 0.5 \cdot \cos 2\alpha + 0.08 \cdot \cos(4\alpha) \tag{A.16}$$
$$= 0.42 - 0.5 \cdot \cos 2\alpha + 0.08 \cdot [1 - 2 \cdot \sin^2(2\alpha)] \tag{A.17}$$
$$= 0.5 - 0.5 \cdot \cos 2\alpha - 0.16 \cdot \sin^2(2\alpha) \tag{A.18}$$
$$= 0.5 - 0.5 \cdot \frac{1-\theta^2}{1+\theta^2} - 0.16 \cdot \frac{4 \cdot \theta^2}{(1+\theta^2)^2} \tag{A.19}$$
$$= \frac{0.5 \cdot (1+\theta^2)^2 - 0.5 \cdot (1-\theta^2) \cdot (1+\theta^2) - 0.64 \cdot \theta^2}{(1+\theta^2)^2} \tag{A.20}$$
$$= \frac{0.5 \cdot (1 + 2 \cdot \theta^2 + \theta^4) - 0.5 \cdot (1-\theta^4) - 0.64 \cdot \theta^2}{(1+\theta^2)^2} \tag{A.21}$$
$$= \frac{0.36 \cdot \theta^2 + \theta^4}{(1+\theta^2)^2} \tag{A.22}$$
$$\tag{A.23}$$

and the ratio $\frac{w'(t)}{w(t)}$ can be written as:

$$\frac{w'(t)}{w(t)} = \frac{\pi}{L} \cdot \frac{0.72 \cdot \theta + 3.28 \cdot \theta^3}{0.36 \cdot \theta^2 + \theta^4} \tag{A.24}$$

$$= \frac{\pi}{L} \cdot \frac{\frac{18}{25} + \frac{82}{25} \cdot \theta^2}{\frac{9}{25} \cdot \theta + \frac{25}{25} \cdot \theta^3} \tag{A.25}$$

$$= \frac{\pi}{L} \cdot \frac{82 \cdot \theta^2 + 18}{25 \cdot \theta^3 + 9 \cdot \theta} \tag{A.26}$$

$$= \frac{2\pi}{L} \cdot \frac{41 \cdot \theta^2 + 9}{25 \cdot \theta^3 + 9 \cdot \theta} \tag{A.27}$$

$$\tag{A.28}$$

Injecting this expression in Equation A.7, and taking into account the notation for θ and α, the radius R necessary to modify the shape of a Blackman window so that its new maximum lies in position t^* is:

$$R = exp[\frac{2\pi}{L} \cdot \frac{41 \tan^2(\frac{\pi t^*}{L}) + 9}{25 \tan^3(\frac{\pi t^*}{L}) + 9 \tan(\frac{\pi t^*}{L}))}] \tag{A.29}$$

Appendix B

Publications

B.1 Patents

- **T. Drugman**, G. Wilfart, T. Dutoit, *Speech Synthesis and Coding Methods*, European patent EP 2242045 A1, PCT patent WO 2010/118953 A1, 2010.

B.2 Regular papers in Journals

- **T. Drugman**, T. Dutoit, *The Deterministic plus Stochastic Model of the Residual Signal and its Applications*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*

- **T. Drugman**, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit, *Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review*, IEEE Transactions on Audio, Speech and Language Processing, *Accepted for publication.*

- **T. Drugman**, B. Bozkurt, T. Dutoit, *Causal-Anticausal Decomposition of Speech using Complex Cepstrum for Glottal Source Estimation*, Speech Communication Journal, Volume 53, Issue 6, July 2011, Pages 855-866, 2011.

- **T. Drugman**, B. Bozkurt, T. Dutoit, *A Comparative Study of Glottal Source Estimation Techniques*, Computer, Speech and Language Journal, Elsevier, *Accepted for publication.*

- **T. Drugman**, B. Bozkurt, T. Dutoit, *Glottal Source Estimation Using an Automatic Chirp Decomposition*, Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing, volume 5933, pp. 35-42, 2010.

B.3 Papers in Conference Proceedings

- **T. Drugman**, A. Alwan, *Robust Pitch Tracking Based on Residual Harmonics*, Interspeech Conference, Firenze, Italy, 2011.

- B. Picart, **T. Drugman**, T. Dutoit, *Continuous Control of the Degree of Articulationin HMM-based Speech Synthesis*, Interspeech Conference, Firenze, Italy, 2011, *(submitted).*

- **T. Drugman**, J. Urbain, T. Dutoit *Assessment of Audio Features for Automatic Cough Detection*, 19th European Signal Processing Conference (EUSIPCO11), Barcelona, Spain, 2011, *(submitted)*.
- **T. Drugman**, T. Dubuisson, T. Dutoit, *Phase-based Information for Voice Pathology Detection*, IEEE International Conference on Acoustics, Speech and Signal Processing 2011 (ICASSP11), Prague, Czech Republic, 2011.
- **T. Drugman**, T. Dutoit, *Chirp Complex Cepstrum-based Decomposition for Asynchronous Glottal Analysis*, Interspeech10, Makuhari, Japan, 2010.
- **T. Drugman**, T. Dutoit, *On the Potential of Glottal Signatures for Speaker Recognition*, Interspeech10, Makuhari, Japan, 2010.
- **T. Drugman**, T. Dutoit, *Glottal-based Analysis of the Lombard Effect*, Interspeech10, Makuhari, Japan, 2010.
- B. Picart, **T. Drugman**, T. Dutoit, *Analysis and Synthesis of Hypo and Hyperarticulated Speech*, 7th ISCA Speech Synthesis Workshop, Kyoto, Japan, 2010.
- **T. Drugman**, T. Dutoit, *A Comparative Evaluation of Pitch Modification Techniques*, 18th European Signal Processing Conference (EUSIPCO10), Aalborg, Denmark, 2010.
- **T. Drugman**, T. Dutoit, *Reconnaissance du Locuteur basee sur des Signatures Glottiques*, XXVIIIe Journees d'Etude sur la Parole, Mons, Belgium, 2010.
- **T. Drugman**, B. Bozkurt, T. Dutoit, *Analyse et Modification de la Qualite Vocale basee sur l'Excitation*, XXVIIIe Journees d'Etude sur la Parole, Mons, Belgium, 2010.
- **T. Drugman**, T. Dutoit, *On the Glottal Flow Estimation and its Usefulness in Speech Processing*, EuroDocInfo Conference, Valenciennes, France, 2010.
- T. Dubuisson, **T. Drugman**, T. Dutoit, *On the Mutual Information of Glottal Source Estimation Techniques for the Automatic Detection of Speech Pathologies*, 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA09), Florence, Italy, 2009.
- **T. Drugman**, G. Wilfart, T. Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Interspeech09, Brighton, U.K, 2009, *[ISCA Best Student Paper award]*.
- **T. Drugman**, B. Bozkurt, T. Dutoit, *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*, Interspeech09, Brighton, U.K, 2009.
- **T. Drugman**, T. Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Interspeech09, Brighton, U.K, 2009.
- **T. Drugman**, T. Dubuisson, T. Dutoit, *On the Mutual Information between Source and Filter Contributions for Voice Pathology Detection*, Interspeech09, Brighton, U.K, 2009.
- **T. Drugman**, G. Wilfart, T. Dutoit, *Eigenresiduals for Improved Parametric Speech Synthesis*, 17th European Signal Processing Conference (EUSIPCO09), Glasgow, Scotland, 2009.

- **T. Drugman**, B. Bozkurt, T. Dutoit, *Chirp Decomposition of Speech Signals for Glottal Source Estimation*, ISCA Workshop on Non-Linear Speech Processing 2009 (NOLISP09), Vic, Spain, 2009.
- **T. Drugman**, G. Wilfart, A. Moinet, T. Dutoit, *Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis*, IEEE International Conference on Acoustics, Speech and Signal Processing 2009 (ICASSP09), Taipei, Taiwan, 2009.
- **T. Drugman**, T. Dutoit, *Hidden Markov Models-based speech synthesis*, EuroDocInfo 2009, Mons, Belgium, 2009.
- **T. Drugman**, T. Dubuisson, A. Moinet, N. D'Alessandro, T. Dutoit, *Glottal Source Estimation Robustness*, IEEE International Conference on Signal Processing and Multimedia Applications (SIGMAP08), Porto, Portugal, 2008.
- **T. Drugman**, T. Dubuisson, N. D'Alessandro, A. Moinet, T. Dutoit, *Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase*, 16th European Signal Processing Conference (EUSIPCO08), Lausanne, Switzerland, 2008.
- **T. Drugman**, A. Moinet, T. Dutoit, *On the use of Machine Learning in Statistical Parametric Speech Synthesis*, 17th Annual Belgian-Dutch Conference on Machine Learning (Benelearn08), Spa, Belgium, 2008.
- M. Gurban, **T. Drugman**, J-P. Thiran, T. Dutoit, *Dynamic modality weighting for multi-stream HMMs in Audio-Visual Speech Recognition*, 10th IEEE International Conference on Multimodal Interfaces (ICMI08), Chania, Greece, 2008.
- **T. Drugman**, M. Gurban, J-P. Thiran, *Relevant Feature Selection for Audio-Visual Speech Recognition*, IEEE International Workshop on Multimedia Signal Processing (MMSP07), Chania, Creta, 2007.
- J-P. Thiran, A. Valles, **T. Drugman**, M. Gurban, *Definition et selection d'attributs visuels pour la reconnaissance audio-visuelle de la parole*, Traitement et Analyse de l'Information : Methodes et Applications (TAIMA07), Hammamet, Tunisia, 2007.

B.4 Scientific Reports

- L. Couvreur, F. Bettens, **T. Drugman**, T. Dubuisson, S. Dupont, C. Frisson, M. Jottrand, M. Mancas, *Audio Thumbnailing*, Numediart Project, Mons, Belgium, March 2008.
- L. Couvreur, F. Bettens, **T. Drugman**, C. Frisson, M. Jottrand, M. Mancas, A. Moinet, *Audio Skimming*, Numediart Project, Mons, Belgium, March 2008.
- N. D'Alessandro, **T. Drugman**, T. Dubuisson, *Transvoice Table*, Numediart Project, Mons, Belgium, March 2008.