Novel multiscale methods for nonlinear speech analysis

Vahid Khanagha

N° d'ordere: 4737

THÈSE

pour obtenir le titre de Docteur en Sciences
de l'Université de Bordeaux 1
Mention : Informatique

NOUVELLES MÉTHODES MULTI-ÉCHELLES POUR L'ANALYSE NON-LINÉAIRE DE LA PAROLE

Présentée et soutenue par

Vahid Khanagha

Équipe GEOSTAT
INRIA Bordeaux Sud-Ouest
École Doctoral Mathématiques et Informatique
Université Bordeaux 1, Sciences et Technologies

Soutenue le 16 Janvier 2013
Devant le jury composé de:

Directeur de thèse	Hussein Yahia	Chargé de recherche (HDR)	GEOSTAT-INRIA Bordeaux Sud-Ouest
Codirecteur de thèse	Khalid Daoudi	Chargé de recherche	GEOSTAT-INRIA Bordeaux Sud-Ouest
Rapporteurs:	Régine André-Obrecht	Professeur	Institut de Recherche en Informatique de Toulouse
	Max A. Little	Chercheur	Massachusetts Institute of Technology
Examinateurs:	Christophe D'Alessandro	Directeur de recherche	LIMSI-CNRS (Paris)
	Gael Richard	Professeur	TELECOM ParisTech
	Antonio María Turiel	Directeur de recherche	Institute for Marine Sciences of Barcelona
Membre invité	Yannis Stylianou	Professeur	Institute of Computer Science-Forth (Greece)

PhD thesis

to obtain the degree of PhD in computer science
of the University of Bordeaux 1

NOVEL MULTISCALE METHODS FOR NONLINEAR SPEECH ANALYSIS

Using the Microcanonical Multiscale Formalism

Presented and defended by

Vahid Khanagha

GEOSTAT team
INRIA Bordeaux Sud-Ouest
Doctoral school of mathematics and computer science
University of Bordeaux 1

Defended on 16 January 2013
In front of the jury composed of:

Thesis director	Hussein Yahia	Chargé de recherche (HDR)	GEOSTAT-INRIA Bordeaux Sud-Ouest
Thesis co-director	Khalid Daoudi	Chargé de recherche	GEOSTAT-INRIA Bordeaux Sud-Ouest
Reviewers:	Régine André-Obrecht	Professeur	Institut de Recherche en Informatique de Toulouse
	Max A. Little	Chercheur	Massachusetts Institute of Technology
Examiners:	Christophe D'Alessandro	Directeur de recherche	LIMSI-CNRS (Paris)
	Gael Richard	Professeur	TELECOM ParisTech
	Antonio María Turiel	Directeur de recherche	Institute for Marine Sciences of Barcelona
Invited member	Yannis Stylianou	Professeur	Institute of Computer Science-Forth (Greece)

Résumé

Cette thèse présente une recherche exploratoire sur l'application du Formalisme Microcanonique Multiéchelles (FMM) à l'analyse de la parole. Dérivé de principes issus en physique statistique, le FMM peut permettre une analyse géométrique précise de la dynamique nonlinéaire des signaux complexes. Il est fondé sur l'estimation des paramètres géométriques locaux, appelés les exposants de singularité, qui quantifient le degré de prédictibilité à chaque point du domaine du signal. Si correctement définis est estimés, ils fournissent des informations précieuses sur la dynamique locale de signaux complexes et ont été utilisés dans plusieurs applications allant de la représentation des signaux à l'inférence ou la prédiction. Nous démontrons la pertinence du FMM en analyse de la parole et développons plusieurs applications qui montrent le potentiel et l'efficacité du FMM dans ce domaine. Ainsi, dans cette thèse, nous introduisons: un nouvel algorithme performant pour la segmentation phonétique indépendante du texte, un nouveau codeur du signal de parole, un algorithme robuste pour la détection précise des instants de fermeture glottale, un algorithme rapide pour l'analyse par prédiction linéaire parcimonieuse et une solution efficace pour l'approximation multipulse du signal source d'excitation.

Les mots-clés: analyse non-linéaire de la parole, analyse multi-échelles de la parole, formalisme microcanonique multi-échelles, exposants de singularité, segmentation phonétique, codage de la parole, détection des instants de fermeture glottale, analyse par prédiction linéaire parcimonieuse.

Abstract

This thesis presents an exploratory research on the application of a nonlinear multiscale formalism, called the Microcanonical Multiscale Formalism (the MMF), to the analysis of speech signals. Derived from principles in Statistical Physics, the MMF allows accurate analysis of the nonlinear dynamics of complex signals. It relies on the estimation of local geometrical parameters, the singularity exponents (SE), which quantify the degree of predictability at each point of the signal domain. When correctly defined and estimated, these exponents can provide valuable information about the local dynamics of complex signals and has been successfully used in many applications ranging from signal representation to inference and prediction.

We show the relevance of the MMF to speech analysis and develop several applications to show the strength and potential of the formalism. Using the MMF, in this thesis we introduce: a novel and accurate text-independent phonetic segmentation algorithm, a novel waveform coder, a robust accurate algorithm for detection of the Glottal Closure Instants, a closed-form solution for the problem of sparse linear prediction analysis and finally, an efficient algorithm for estimation of the excitation source signal.

Key words: non-linear speech analysis, multi-scale speech analysis, microcanonical multiscale formalism, singularity exponents, phonetic segmentation, speech coding, glottal closure instant detection, sparse linear prediction analysis.

There was a Door to which I found no Key,
There was a Veil through which I could not see,
Some little Talk awhile of Me and Thee
"Hey there seemed" and then, no trace of Thee and Me.

— O. Khayyám

Acknowledgments

This work would not have been possible without the guidance and help of several individuals who in one way or another contributed in preparation and completion of this study; to only some of whom it is possible to give particular mention here.

I would like to express my sincerest gratitude to my advisors Dr. Hussein Yahia and Dr. Khalid Daoudi, for their support and guidance during my Ph.D study and research. I have had the privilege of being directed by the expertise of Khalid in speech processing and also Hussein as one of the founders of the emerging signal processing methodology upon which this thesis is built. I am grateful to Hussein for welcoming me in GEOSTAT team and giving me the excellent opportunity of conducting my PhD study on this subject and for his patience and kindness throughout this project. I am grateful to Khalid for encouraging me to progress, for constantly directing me towards the right track and for rigorously supervising the scientific quality of my work.

I wish also to thank the help and support of my colleagues in GEOSTAT. I am thankful to Josy Baron for her kind presence, to Oriol Pont for helping me catching up upon my arrival, to Suman Kumar Maji for being with me and to Hicham Badri for bringing lots of motivation to the whole team. Thank you to Harish Kumar, Joshua, Nicolas, Andrea, Luca and Denis for your company.

Special thanks to many friends I met in Bordeaux who made my stay an unforgettable experience. Your presence helped me overcome setbacks and stay focused. It would not have been possible without you all.

The greatest sadness in these years has been being away from my dear family and my beloved hometown. Although far away, their love and support was the greatest motivation and encouragement for going on. I missed my brothers Ali and Sina every day and also Saeed and Saeedeh who have encouraged me a lot. Finally, I whole-heartedly dedicate this thesis to my parents, for they brought me forth, raised me and taught me to be everything that I am.

Contents

Acronyms

ACC	ACCumulative function
BIC	Bayesian Information Criterion
dEGG	differentiated Electro-Glotto-Graph
DPCM	Differential Pulse Code Modulation
EGG	Electro-Glotto-Graph
FA	False Alarm
FDT	Fully Developed Turbulence
GCI	Glottal Closure Instance
HMM	Hidden Markov model
HR	Hit Rate
LE	Lyapanov Exponents
LLRT	Log Likelihood Ratio Test
LPA	Linear Prediction Analysis
LPC	Linear Predictive Coding
SE	Singularity Exponent
MFCC	Mel Frequency Cepstral Coefficients
MMF	Microcanonical Multiscale Formalism
MOS	Mean Opinion Score
MPE	Multi Pulse Excitation
MSM	Most Singular Manifold
OS	Over Segmentation
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
RCT	Relative Computation Time
SVF	Spectral Variation Function
TD	Text Dependent
TI	Text Independent

Notations

$ACC(t)$	The Accumulative function for phonetic segmentation.
$h(t)$	Singularity Exponent at time t.
$h_{r_i}(t)$	Partial singularity exponents at scale r_i.
F_h	The level-sets of the singularity exponents: $F_h = \{t : h(t) = h\}$.
$D(h)$	The singularity spectrum: the Hausdorff dimension of F_h.
$D_{r_i}(h)$	The singularity spectrum at scale r_i.
$\delta D_{r_1,r_2}$	The Directed Weighted Average Difference function.
$\Gamma_r(\cdot)$	The multi-scale functional at time t and scale r.
$\mathcal{D}(t)$	An appropriate differential operator for computation of $\Gamma_r(\cdot)$.
h_∞	The smallest value of h(t).
F_∞	The level set representing the MSM: $F_\infty = \{t : h(t) = h_\infty\}$.
$\rho_r(h)$	The empirical histogram of singularity exponents.
$\Psi_r(t)$	The scaled version of a given mother wavelet ($\Psi(t)$).
	$\Psi_r(t) := r^{-d}\,\Psi(r/t)$).
$\mathbb{T}_\Psi[s](r,t)$	The continuous wavelet transform of the signal $s(t)$.
	$\mathbb{T}_\Psi[s](r,t) := (\Psi_r * s)(t)$.
$\mathcal{L}_c(t)$	The level-change functional for GCI detection.
$s(t)$	The continious signal under test.
$s[n]$	The descritisized version of $s(t)$: $s[n] = s(t = nT_s)$.
	T_s is the Sampling period.
$KS_{D_{n1,n2}}$	Kolmogorov-Smirnov distance between data samples $n1$ and $n2$.
F_{k,n_k}	The cumulative distribution function of the kth data sample.
	n_k denotes the length of the data sample.

1

Introduction

This thesis presents an exploratory research on the application of a novel multi-scale non-linear approach to the analysis of the speech signal. Speech is indeed the simplest and the most efficient mean of human communication but yet, the high complexity of the machinery involved in its production puts it among the most difficult phenomena to be perceived by a machine. The raw speech signal that a machine has access to, is no more than a continuously varying acoustic pressure wave, transduced into an electrical signal by the microphone. These changes in pressure however, are generated by elaborate intervening of several organs ranging from the ones in the respiratory system to the articulators in oral cavity.

Speech analysis refers to the task of converting these raw recordings of the speech signal (the pressure wave) into more characteristic representations, so as to facilitate the access to descriptive and invariant attributes of this signal. In simpler terms, speech analysis involves the extraction of “interesting" information from the raw digitized speech signal. This information, which may include parametric representations of the functionality of organs involved in speech production or pure spectral descriptions, will later be employed in more sophisticated processing systems that might be specialized in many different end-user applications in speech technology: retrieval of speech linguistic content, its compression for efficient transmission and storage, recognition of the identity of the speaker, artificial synthesis/re-production and a variety of medical applications. As such, speech analysis forms the lowest layer, or the backbone, of all these high-level applications.

Despite the high degree of complexity of this signal, as we often tend to use the simplest solutions, the field of speech analysis is arguably dominated by *linear* signal processing models and methods. The reason might be the widespread knowledge about the well-established linear approaches and their analytical simplicity. However, there exist a handful of evidences regarding the existence of non-linear dynamics in production mechanism of this signal and the insufficiency of existing linear models for taking all these non-linear aspects into account. These evidences may even invalidate some of the underlying assumptions of linear techniques. Indeed, these linear approaches have had some success in advancement of the speech technology during past decades. Also, they usually do acknowledge the existence of

non-linear dynamics by considering exceptions and/or imposing special constraints to their linear models. But still, the emergence of powerful tools for non-linear processing and analysis of digital signals, has caused a trend towards non-linear speech analysis aiming both at improving the existing linear methods and also at enhancement of our knowledge regarding this complex signal (we will make a brief review on the evidences regarding the non-linear character of speech and the insufficiency of linear methods in capturing them, in chapter 2).

1.1 Objectives of the study

In this thesis, our strategy is to take the analogies coming from the study of highly disordered systems in statistical physics and adapt them to the specific case of the speech signal: we consider speech as a complex signal (as a realization of a complex physical system), and we attempt to find and measure key parameters characterizing some aspects of the non-linear dynamics of the speech signal. As we will see in chapter 2, previous studies show that such parameters exist, but they might be very difficult to compute for real data. Our goal is to develop an alternative approach, drawn from first principles and out of any limiting assumptions.

To do so, we will try to evaluate the parameters associated to the local concept of predictability in such complex signals. This is made possible by the emergence and establishment of a novel formalism, called the Micro-canonical Multi-scale Formalism (the MMF), which provides efficient tools and methods for characterization of the non-linear aspects of any given complex signal, in a local geometric manner as well as a global statistical way. The MMF has recently proven to be promising in many signal processing applications ranging from signal compression to inference and prediction in a quite diverse set of scientific disciplines such as satellite imaging [52, 53, 133, 134], adaptive optics [79, 80], computer graphics [15] and natural image processing [127, 128]. The heart of the formalism is the precise estimation of local parameters at each point in the signal domain, out of any limiting constraint such as stationarity assumption. This makes it a suitable technique for analysis of a signal with rapidly time-varying dynamics such as the speech signal. These local parameters unlock the access to a hierarchy of multi-scale organizations responsible for the complex dynamics of the signal. Particularly, they give access to a subset of points called the Most Singular Manifold (the MSM), which is shown to be the most informative subset of points in case of natural images, in the sense that the whole image can be reconstructed using exclusively the information contained in this subset.

We therefore concentrate in this thesis on the study of the applicability of the MMF to speech as a potential alternative non-linear approach for speech analysis. We start by an extensive set of experiments to investigate theoretical validity of

the application of MMF to this particular signal. We will then study the relevance of the tools and concepts already developed in the formalism to several speech analysis problems. Therefore, through finding attractive applications of these tools and methods, we try to understand the implications of this possible correspondence. In the meantime, to cope with particularities of the speech signal, we will try to adapt these tools to [possibly] multifarious dynamics of speech.

We emphasize that our goal in thesis is not to achieve best of performances in the speech analysis applications we encounter. But rather, our goal is to do an exploratory research on understanding the possible correspondence of the components of the MMF with particularities of this complex signal. Of course, we always put our algorithms in direct comparison with the state-of-the-art methods and whenever possible, we will try to overcome the difficulties of conventional methods in terms of efficiency and performance. In the development of these methods, we always try to avoid the common drawbacks in the employment of non-linear approaches for speech analysis. Some of these drawbacks are listed in [68] as:

- non-linear techniques usually have greater computational burden compared to their linear counterparts;
- the usual lack of closed-form solution, results in iterative procedures with the associated problem of local minima;
- the general complexity and difficulty of analysis, as the use of simple well-established linear tools and techniques like Fourier analysis is no longer valid.

1.2 Summary of contribution

The contributions of this thesis are twofold (list of corresponding publications is provided in section 1.4):

- The first set of contributions is related to the demonstration of the applicability of the MMF to speech analysis. Considering the ongoing developments on successful application of the MMF to analysis of a wide variety of complex signals, our results adds up to establishment of the MMF as an emerging *generic* signal processing framework. Moreover, as the formalism is originally developed for 2-D complex signals, another aspect of the contributions of this work to the MMF would be the adaptation of some of its components to the specific case of the speech signal. In particular, a new definition for a multi-scale functional is proposed that provides better results in case of the speech signal.

- The second set of contributions are those directly related to speech signal analysis, showing the strength of the MMF in localizing *important* events in the signal on different levels:
  - accurate localization of phoneme-level transition fronts. A novel efficient phonetic segmentation algorithm is proposed which shows competitive performance compared to the state-of-the-art, while being more accurate for smaller tolerances of error.
  - providing a parsimonious representation of the speech signal in form of a waveform coder.
  - accurate, efficient and robust localization of Glottal Closure Instants (GCI). The algorithm is as accurate the stat-of-the-art approaches for clean speech and is more robust against different noises (the code for this algorithm will be available on GEOSTAT website).
  - application in sparse linear prediction analysis. A simple closed form soloution is introduced which benefits from stability while providing higher levels of residual sparsity.
  - estimation of multi-pulse excitation sequence of the speech signal in an efficient manner.

In all the above applications, we directly put our developed algorithms in comparison with the state-of-the-art to assess their performance and efficiency and we show that interesting results can be achieved using this novel formalism.

1.3 Organization of the thesis

The first part of this thesis is dedicated to the introduction of the theoretical foundations of the MMF, the presentation of the corresponding computational aspects of the formalism, and also preliminary feasibility studies with regard to its application to the speech signal. In the second part, we present several case studies on the successful application of the MMF to practical speech analysis problems.

The first part starts with a short discussion about the scientific context of this thesis in chapter 2, where we show where our approach stands in the broad context of non-linear speech processing. Chapter 3 presents the basics of the MMF, its components and its advantages over its older counterparts. In Chapter 4 the preliminary experiments are presented to support the idea of applicability of this formalism to speech signal analysis. The experiment is performed over a large database of isolated phonemes so as to separately study whether the formalism is applicable to each phoneme family or not.

The second part (the applications), starts with chapter 5, where the first observations about the usefulness of this formalism to speech analysis are presented. Con-

sequently, a simple algorithm is introduced for *text-independent phonetic segmentation* of the speech signal. The method is shown to have competitive performance compared to the state-of-the-art methods, with interestingly high resolution in detection of phoneme boundaries.

In chapter 6, we first introduce some numerical modifications to the formalism that makes it more adapted to the specific case of the speech signal and then, we proceed to show how the formalism can be used to recognize a subset of points inside the signal (the MSM), from which the whole signal can be reconstructed with good perceptual quality. This subset, which is originally introduced in the formalism as the subset of most informative points of the signal, is then used to develop a novel efficient *waveform coding* technique.

We argue in chapter 7 about the correspondence of this subset to the physical production mechanism of the speech signal and we discuss how the points in this subset are related to the instants of the most significant excitations of the vocal tract system (the GCIs). As the identification of these points is an important topic in many speech processing applications, we develop a novel solution for *GCI detection* problem. The method is shown to have comparable performance to the recent state-of-the-art in terms of reliability while it is more robust against noise. Moreover, the algorithms is more efficient than the fastest available methods.

Consequently, these points (GCIs) are used in chapter 8, to introduce a novel solution to the problem of *sparse Linear Prediction* (LP) analysis, aiming at estimation of LP coefficients such that the resulting residuals are sparse. The approach is based on a weighting of the l_2-norm objective function using the GCI estimates. This provides an efficient closed-form solution for this interesting problem.

The same philosophy of relating the MSM to significant excitations of vocal tract system is followed in chapter 9 for *multi-pulse approximation of speech excitation source*. The latter provides an efficient engine to be used inside classical multi-pulse excitation coder. The resulting coder provides almost the same level of perceptual quality as the classical one, while being much more efficient. Finally we make our conclusions in chapter 10.

1.4 Publications

- **Journal**
  - V. Khanagha, K. Daoudi, *"An Efficient Solution to Sparse Linear Prediction Analysis of Speech"*, EURASIP Journal on Audio, Speech, and Music Processing - Special Issue on Sparse Modeling for Speech and Audio Processing (June 2012).
  - V. Khanagha, K. Daoudi, O. Pont and H. Yahia, *"Non-linear speech representation based on local predictability exponents"*, Elsevier's Neurocomputing Journal, special issue on Non-Linear Speech Signal processing (March 2012).
  - V. Khanagha, K. Daoudi, H. Yahia, O. Pont, *"A novel approach to phonetic segmentation through local singularity analysis of speech", submitted to Elsevier's DSP journals.*
  - V. Khanagha, K. Daoudi, H. Yahia *"Robust detection of glottal closure instants through local singularity analysis of speech", in preparation for IEEE Transactions on Audio, Speech, and Language Processing.*
- **Peer-reviewed conferences/proceedings**
  - V. Khanagha, K. Daoudi, *"Efficient multi-pulse approximation of speech excitation using the most singular manifold"*, Accepted in INTERSPEECH 2012, Portland Oregon, United States.
  - V. Khanagha, H. Yahia, K. Daoudi, O. Pont, *"Reconstruction of Speech Signals from their Unpredictable Points Manifold"*, NOn LInear Speech Processing (NOLISP), November 2011, Las Palmas de Gran Canaria, Spain, proceedings published in: Lecture Notes in Computer Science: Advances in Nonlinear Speech Processing, Springer.
  - V. Khanagha, K. Daoudi, O. Pont, H. Yahia, *"Improving text-independent phonetic segmentation based on the micro-canonical multi-scale formalism"*, International Conference on Acoustics, Speech and Signal Processing (**ICASSP**), May 2011, Prague, Czech Republic.
  - V. Khanagha, K. Daoudi, O. Pont, H. Yahia, *"A novel text-independent phonetic segmentation algorithm based on the micro-canonical multi-scale formalism"*, (INTERSPEECH), September 2010, Makuhari, Japan.
  - V. Khanagha, K. Daoudi, O. Pont, H. Yahia, *"Application of the micro-canonical multi-scale formalism to segmentation of speech signals"*, European Signal Processing Conference (EUSIPCO), August 2010, Aalborg, Denmark.

- **In French**
  - V. Khanagha, K. Daoudi, O. Pont and H. Yahia, *"Une nouvelle approche non-linéaire pour la segmentation phonétique"*, XXIIIe Colloque GRETSI 2011, September 2011, Bordeaux, France.
  - V. Khanagha, K. Daoudi, O. Pont and H. Yahia, *"Segmentation indépendante du texte par le formalisme multiéchelles microcanonique"*, MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication (MAJESTIC), October 2010, Bordeaux, France.
  - V. Khanagha, K. Daoudi, O. Pont and H. Yahia, *"Application du formalisme multiéchelles microcanonique pour la segmentation des signaux de parole"*, CORESA 2010, December 2010, Lyon, France.

Part I

THE MICROCANONICAL MULTISCALE FORMALISM

2

Scientific context

In this chapter we briefly review the evidences regarding the insufficiency of the existing linear techniques in capturing the whole dynamics of the speech signal and also the motivations behind searching for alternative non-linear speech analysis methods. To do so, we first generally introduce in section 2.1, the machinery involved in speech production. We then use that information in section 2.2 to explain the shortcomings of classical linear approaches. Finally in section 2.3, we explain our motivations for employing the particular formalism we have chosen and we specify where in the broad context of non-linear speech processing our approach stands in.

2.1 The production mechanism of the speech signal

Speech signal is produced by elaborate intervening of three main groups of organs [100]:

- the lungs, which are the source of energy for the speech signal.
- the larynx, or "voice box", which houses the vocal folds. The slit-like orifice between the two folds is called the glottis. The muscles in the larynx control the width of the glottis and also the amount of tension in the folds as two important parameters in controlling the pitch and the volume of the sounds [101].
- the vocal tract, including the pharynx plus the oral cavity which is coupled to the nasal cavity. The oral cavity may assume many different shapes, with non-uniform cross-sections, depending on the shape of jaws and the configuration of the articulators (the tongue, teeth and lips) during the speech.

Lungs act as the power supply of the vocal system and expel a burst of DC air, which later experiences a perturbation somewhere in the larynx or the vocal tract. The location where this perturbation happens varies depending the type of sound that is being produced [100]:

- for *voiced* sounds, this perturbation occurs in the larynx when a partial closure of the glottis causes a self-sustained oscillation of the vocal folds. Hence the DC flow of air is converted to quasi periodic glottal pulses.
- for *fricative unvoiced* sounds, passage of the airflow through a quasi-narrow constriction in the oral cavity (generally shaped by the tongue), results in a turbulent flow of air.
- for *plosive unvoiced* sounds, the complete obstruction of the oral cavity (by the lips, tongue or teeth) is followed by a sudden release and hence results in an impulsive release of the air compressed behind the obstruction.

So that, through any one of the above disturbance mechanisms (or a combination of them) the DC airflow is converted to a flow of air with time varying velocity waveform, which later attains its specific phonetic character by traveling in the rapidly time-varying acoustic medium inside the vocal tract. Indeed, the vocal tract may attain a variety of shapes during continuous speech (depending on configuration of the articulators and the shape of its cavities).

This production mechanism is the motivation for the famous source-filter model: a raw excitation source is generated by any one of the above disturbance mechanisms and then, the resulting waveform is colored depending on the shape of the vocal tract system that it experiences [100]. Classically, the effect of the vocal tract system on the excitation waveform is approximated as a linear filter.The excitation source on the other hand, is idealized as periodic puffs for voiced sounds, white noise for fricatives and isolated impulses for plosive sounds. To account for time-varying character of excitation source and the vocal tract, the analysis is performed in windows of 10-30 ms length, during which the corresponding characteristics are assumed invariant.

This linear approximation is the basis for most of the current state-of-the-art speech analysis techniques. An example is the widespread Linear Prediction Analysis (LPA) technique, which is widely used in almost every field of Speech Processing. LPA is based upon consideration of the vocal tract system as an all-pole linear filter. This in turn suggests an Auto Regressive (AR) predictive formulation for speech samples. The LPA thus serves for estimation of parameters of this AR model (i.e. the all-pole filter) representing the spectral shape of vocal tract filter and consequently the estimation of the excitation source would be possible by inverse filtering. The independent functioning of the excitation source and the vocal tract in linear source-filter model is an underlying assumption of many other conventional techniques, even when they involve non-linear mathematical manipulations (as it is the case for homomorphic [100] speech processing).

2.2 Non-linear character of the speech signal

There are many experimental evidences as well as theoretical reasonings regarding the existence of non-linear effects in the production mechanism of speech, which are generally ignored in the main-stream linear approaches. These nonlinear aspects are evidenced in the production process of all different types of sounds produced by human vocal system. They are even acknowledged in classical speech processing literature, just before settling down to their linear approximations.

A typical example of such non-linear phenomena is the existence of turbulent sound source in production process of unvoiced sounds [61, 68]. During the production of these sounds, vocal folds are more tensed and are closer to each other, thus allowing for turbulence to be generated ([100], pp. 64). This turbulence that is produced in the vocal folds is called aspiration and is the source of excitation for whispered and breathy voices. Although in linear techniques the turbulent sound source is accounted for by considering the excitation source as white noise, they still fail to fully characterize the sound as their underlying assumption is to consider the flow as a laminar one [38]. Also for the plosives, whose excitation source is idealized as an impulsive source in linear framework, there exist a time spread and turbulent component in practice ([100], pp. 88). The general existence of small or large degrees of turbulence in speech signal is discussed in [83], using the Navier-Stokes equation of fluid motion.

In case of voiced sounds, [121] reports some evidences regarding their characterization by highly complex airflows like jets and vortices in the vocal tract system. Moreover, it is known that during voicing the vibrations of vocal folds are not exactly periodic as it is idealized in the linear model [84] and their shape changes depending on the amplitude [68]. This prohibits the exact estimation of glottal waveform using the linear inverse filtering approach (while this waveform has an essential role in voicing quality for synthesis applications [36]). These non-linear effects in vocal fold oscillations during voicing are attributed to non-linear feedback via Bernoulli forces [68].

Another major assumption made in linear framework is independent functioning of the vocal tract and the excitation source [36, 38, 68, 84]. When glottis is open there exist a coupling between the two which results in significant changes in formants compared to the closed phase [65]. If there were no interactions, the flow would be proportional to the glottal opening area. But the existence of source-filter interactions, causes the final glottal flow during voicing to be skewed towards right: it slowly increases to a maximum and suddenly ends [96].

The appropriateness of linear systems assumptions [71] are tested in [77, 76] using surrogate data techniques. It is empirically shown and argued (using biomechanical information on speech production) that the mathematical assumptions of LTI system theory cannot represent all the dynamics of the speech. It is then shown

that a stochastic non-linear non-Gaussian model for speech signals offer a simplified but more general mathematical framework for speech (in the sense that more phenomena can be explained with fewer assumptions).

We can thus fairly conclude that the existence of non-linear phenomena is theoretically and experimentally established and so we can accept the fact that the production mechanism of the speech signal is not entirely linear. This has motivated a trend toward exploration of the possibilities to catch this non-linear character aiming both at improvement of conventional linear techniques and also at improvement of scientific understanding of this complex phenomena.

2.3 Speech as a realization of a non-linear dynamical system

There exists a broad spectrum of non-linear methods for speech analysis. Several reviews are provided on this subject in [38, 68, 77, 84] and also some collections of non-linear algorithms are presented in [23, 111, 118, 124]. In this thesis we are interested in an important class among these methods which considers speech as a realization of a non-linear dynamical system and attempts to use the available tools and methods in the study of such systems to catch non-linear features of this signal. Such attempts are motivated by the theoretic and experimental evidences regarding the association of the non-linear aspects in speech production to the turbulent nature of airflow in the vocal tract [61, 83, 121, 75], which in turn justifies the use of methods from the study of chaotic, turbulent systems for speech analysis [38].

In [83], motivated by the considerations on the dynamics of the airflow, it is conjectured that short-term speech sounds contain various degrees of turbulence at time scales below phoneme time scale. The Multi Fractal Dimension (MFD) is thus conceptually equated to the degree of turbulence in a speech sound. Consequently, short-time MFD is shown to be a useful feature for speech-sound classification, segmentation and recognition. Moreover, it is shown that unvoiced fricatives, affricates, stops and some voiced fricatives have a high fractal dimension at all scales, which is consistent with the presence of turbulence phenomena. The MFD value for some voiced fricatives and also vowels is medium-to-high at different scales, which corresponds to turbulence.

It is assumed in [95] that speech production can be regarded as a low-dimensional nonlinear dynamical system $Z(n+1) = H(Z(n))$ and speech signal $s(n)$ as a 1-D time-series resulted from the application of a vector function to the original D_e dimensional dynamic variable $Z(n)$. According to embedding theorem [63], under general assumptions, the lag vector $\mathbf{X}(\mathbf{n}) = [s(n), s(n-T_d, \cdots, s(n-(D_e-1)T_d)]$ has many common aspects with the original phase space of the original (but unknown) phase space $Z(n)$. Hence, several measurements such as correlation dimension and general dimension are tried on the attractor constructed by $\mathbf{X}(\mathbf{n})$ as examples of

invariant quantities. These raw measurements are then used as descriptive feature sets whose values (or their statistical trends) are shown to be [on average] dependent on general characteristics of sounds such as voicing, the manner and place of articulation of broad phoneme classes.

Another characteristic of a dynamical system that might be conserved by the embedding procedure are Lyapanov Exponents (LE). These quantities are often considered as quantitative fingerprints of chaos as they are related to the concept of predictability in a dynamical system. LEs are computed in [66] for isolated phonemes and they are shown to be useful features for phoneme classification.

LEs and correlation dimension are also shown to be valuable tools for voice disorder detection applications. It is discussed in [75] that the classical linear methods can not characterize the whole typography of disordered voice sounds and cannot be used to quantify two main symptoms of voice disorders which do not satisfy near periodicity requirement. Hence, the non-linear dynamical systems theory has been considered in the literature as a candidate for a unified mathematical framework modeling the dynamics seen in all types of disordered vowels. Authors in [75] provide a review on successful application of quantities like LEs and correlation dimension for classification of disordered sounds and also point out some practical limitation in computation of these quantities (for instance, correlation dimension is a quantity which is sensitive to the variance of the signal [85]). They mention that the deterministic non-linear methods are not appropriate for characterization of signals of random and very noisy nature and hence, they introduce an extended non-linear framework covering both deterministic and stochastic components of the speech signal. Recurrence and scaling analysis methods are thus employed using two measures: recurrence period density entropy [63] (measuring the extent of aperiodicity) and detrended fluctuation analysis [55] (which characterizes the self-similarity of the graph of the outcome of a stochastic process, and is used for characterizing the increased breath noise in disordered voices). It is shown that such combination can distinguish healthy from disordered voices with high accuracy [74] with the added benefit of reliability due to reduction of adjustable algorithmic parameters.

In fact, the above experiments altogether reveal the potential of methods related to the characterization of turbulence and the study of chaotic dynamical system in speech analysis. Also, apart from their applicative relevance, they may bring up some interesting indications about the dynamics of the speech signal, if the limitations in their precise estimation can be neglected. The resulted positive value of LEs in [66] for most of the phonemes, which corresponds to chaotic behavior, at least implies the importance of predictability issues in speech analysis. Moreover, the above than 1 value of MFD in most of the cases in [83], at least confirms the existence of meaningful scale-dependent quantities. However, there exist some practical issues which put some limitations on these methods.

For instance, the determination of the appropriate dimension for the phase space and the appropriate time delay for the construction of the phase-space in embedding procedure, which are important factors in proper estimation of dynamically invariant parameters. In case of LEs, their computation is non-trivial for experimental data and particularly for the speech signal, in which the stationary assumption necessitates the procedures to be performed on phone-level which corresponds to very short data lengths. This is the reason why the authors in [16], have used sustained vowels for the computation of local LE (rather than naturally uttered phonemes). In fact, the dynamical models they have used require longer data samples compared to the actual length of ordinary vowels. This problem is addressed in [66], where several dynamical models in phase-space are extensively compared regarding their fidelity in estimating the LEs of a known dynamical system while a very short amount of data is available and finally TSK model [120] is chosen for the computation of LEs. Also, recent works insist on the limitations brought by classical LEs w.r.t. predictability: a Lyapunov exponent is a *global* quantity measuring an average divergence rate. In the general case, there are some fluctuations in finite time, playing an important role in predictability, which lead to the consideration of large deviations [119]. In case of deterministic chaos, Finite Time LE (FTLE) [1] exhibit multi-scale behavior and is related to large deviations in finite time intervals [119]. But FTLE and classical LE only coincide in the limit ($t \to \infty$) and their precise numerical computation can be difficult.

From these considerations it appears that the computation of key quantities related to non-linearity in speech is worth contemplating. In fact, the above references show that a community is developing on these topics. However, the use of classical embedding techniques are computationally challenging in the case of fully developed turbulence; and they usually provide global descriptions, which is suitable for classification applications, but not for geometric point-wise analysis; also the computation of singularity spectrum using known methodologies (Wavelet Transform Modulus Maxima, histogram method, etc) are equivalent to computing thermodynamic averages over microcanonical ensembles [10]. Among the first attempts for geometrization is a paper by She and Leveque [108], in which filaments observed in fully developed turbulence are related to limiting moments ratio $\xi^{(\infty)}$, leading to more accurate formula for the Legendre spectrum (which will be introduced in chapter 3).

As a consequence, the emergence of new computational approaches for accessing quantitatively and robustly to local scaling exponents around each point using specific measures of predictability opens vast areas of research for understanding the geometric multi-scale implications of a complex signal, that is to say, the geometrical interplay between statistical information content and the multi-scale organizations,

1 The FTLE is a scalar value that quantifies the amount of deviation between two particles flowing in a fluid, over a given time interval [106].

predicted by them. The work presented in this thesis is new in the sense that it focuses on the implications contained on accessing localized scaling exponents in the speech signal which can be related to a geometric concept of predictability in the framework of reconstructible systems. To understand the power of the approach, called the MMF, we first focus on a description of the MMF in the following chapter.

3

The Microcanonical Multiscale Formalism

In this thesis, we use a novel formalism called the Microcanonical Multiscale Formalism (the MMF) for the analysis of the speech signal. This formalism has its roots in the study of disordered systems in statistical physics and is related to a precise [quantitative] study of the notion of transition inside a complex system or signal. Statistical physics shows that complexity in a system is intrinsically related to the existence of a hierarchy of multi-scale structures inside the system. A typical example of such multi-scale organization is related to the cascade of energy in the case of fully developed turbulence. The fingerprints of these multi-scale structures are observed in a wide range of natural signals acquired from different complex systems.

In this context, the MMF is an extension of its standard canonical counterpart [6, 42] which provides *global* views upon such complex structures. The particularity of MMF is that it is based on *geometrical* and *local* parameters rather than relying on *global* quantities. Therefore, MMF makes it possible to locally study the dynamics of complex signals from a multi-scale perspective. Meanwhile, rigid mathematical links are made between this geometric analysis of complexity in the MMF and the global statistical view in the canonical framework. So that, MMF provides tools and methods for both geometric and global description of non-linear phenomena in complex signals characterizing their intermittent signature. In other words, it allows the study of local geometrico-statistical properties of complex signals from a multi-scale perspective.

In practice, the MMF is shown to be a valuable approach to model and analyze this multi-scale hierarchy in empirical complex and turbulent systems having corresponding statistical properties at different scales and it has shown outstanding results in a wide range of applications from diverse scientific disciplines [15, 52, 80, 127, 128, 134, 133].

In this chapter, we briefly present the MMF, its distinctive features compared to its predecessors and its capabilities. We provide all the necessary computational details that would form the basis for all the experiments in this thesis. In section 3.1 we introduce classical methods for global study of multi-scale properties of complex systems. Consequently, in section 3.2 MMF is introduced as a geometrical approach

to complexity along with its major components. The numerical details of the formalism is extensively reviewed in section 3.3. Finally the conclusion for the chapter is presented in section 3.4.

3.1 The canonical approach to multi-scale complexity

In the classical canonical formalism, the multi-scale hierarchy is assessed through the application of global statistical descriptors such as structure functions of different physical observables. The origin of this approach can be traced back to the study of highly disordered flows (spatially and temporally) where it was observed that under certain regimes (with low viscosity) the flow can not be characterized by deterministic smooth fluid motions. Instead, as the flow is dominated by inertial forces, a disordered flow of eddies, filaments and other flow instabilities are observed [41]. However, an interesting phenomenon was observed in these flow regimes: certain statistics of the fluid remain intact across several scales. Indeed, it is observed that when a highly non-viscous fluid passes by an obstacle like a cylinder, original deterministic symmetries are broken but they are recovered in a statistical sense over several scales [42, 108].

The persistence of statistical properties across different scales can be explained using the well-known intuition of Kolmogorov about the multiplicative cascade of energy transfer in a range of fine scales (inertial scales). Turbulent flow can be visualized as a cascade of large eddies that successively break-up into smaller-sized ones [41]. In such regimes, dissipation of energy becomes irrelevant and there only exists a mechanical transfer of energy from coarser structures to the similar but finer structures [42, 108]. This mechanism can be multiplicatively written for a general complex signal $\mathbf{s}$ as:

$$|\Gamma_r \mathbf{s}| \doteq \eta_{r/l} |\Gamma_l \mathbf{s}| \tag{3.1}$$

where $r < l$ are scales chosen within the inertial range [1]. Although this multiplicative relationship was initially written in terms of longitudinal velocity increments for a disordered flow, here we have used a more general form, using a general scale-dependent functional Γ_r operating on the complex signal $\mathbf{s}$. The symbol $\doteq$ means that both sides are equally distributed and $\eta_{r/l}$ is the energy injection parameter, which in its simplest form, is assumed to be exclusively dependent on the ratio of scales as $\eta_{r/l} \propto (r/l)$. Taking the p-th moments of both sides of Eq. (3.1) results in the following power-law scaling:

1 the range of scales where energy dissipation is irrelevant and what appears as energy dissipation is actually the flow of energy from coarser scales to the finer ones [42].

$$\langle |\Gamma_r \mathbf{s}|^p \rangle = A_p r^{\tau_p} \tag{3.2}$$

where $A_p = \langle |\Gamma_l \mathbf{s}|^p \rangle l^{-\tau_p}$ is a factor that is independent of the scale r and the exponent of this power law (τ_p) is called the Legendre spectrum. Kolomogrov intuition was that τ_p varies linearly with p: $\tau_p = \delta p$. However, experimental observations show that the Legendre spectrum is not exactly a linear function of p in form of δp. It is actually diverging from this linear form for large values of p and is a curved convex function of p [8].

The deviation of τ_p from the linear scaling with p is due to a common basic feature of complex signals, called intermittency. An intermittent signal displays activity only in a fraction of time and this fraction decreases with the scale under consideration. Indeed, while going toward finer scales, the reproduction of eddies becomes less space filling and hence the exponent of the power-law structure function decreases [42].

As such, the map $p \rightarrow \tau_p$ is an important representation characterizing the intermittent character of complex signals. Indeed, Eq. (3.2) implies that having known τ_p, the p-th order statistics of the physical variable $\mathbf{s}$ at the fine scale r, can be deduced from that of the coarser scale l. Thus, τ_p can uniquely describe the inter-scale organization and the intermittent character of the complex signal. Other than complete characterization of multi-scale hierarchy from a statistical point of view, once this multiplicative relationship is validated, it might be used to obtain a more parsimonious representation, using the fact that if Eq. (3.1) holds for a set of scales, the information in finer scales might be deduced from those of coarser scales.

3.2 The micro-canonical formalism

The exponent of the canonical power-law in Eq. (3.2) describes the intermittent character of a complex signal, only from a global statistical point of view. Indeed, the Legendre spectrum is a global description and provides no information about local complexity. Therefore, it can be only used to recognize the existence of an underlying multi-scale structure, without any information about its geometric organization [2]. Besides, the evaluation of this power-law requires the calculation of p-th order moments of variables (the structure function) which imposes stationarity assumptions. Moreover, the computation of these structure functions is highly demanding in data, which prohibits their use for empirical data in most cases [130].

2 There exist however attempts in the canonical framework to geometrically access these complexities. An example is the Wavelet Transform Modulus Maxima method [81, 82, 90] which is discussed in section 3.3.1.2

The microcanonical framework (the MMF) provides computationally efficient tools for *geometrical* characterization of this inter-scale relationship. MMF provides access to local scaling parameters which provide valuable information about the local dynamics of a complex signal and can be used for precise detection of critical events inside the signal. As such, the micro-canonical formalism not only recognizes the global existence of complex multi-scale structures, but also it shows *locally* where the complexity appears and how it organizes itself. Indeed, as we will see shortly, rigid theoretical links has been made between this geometrical evaluation and the statistical viewpoint in section 3.1. These links also serve to provide theoretical evidence regarding the meaningfulness of such local analysis [43].

MMF does not rely on statistical values for ensemble averages, but rather look at what is going on around any given point. It is based on the computation of a scaling exponent $h(t)$ at every point in a signal domain and out of any stationarity assumption. These exponents are formally defined by the evaluation of the limiting behavior of a multi-scale functional $\Gamma_r(s(t))$ at each point t over a set of fine scales r:

$$\Gamma_r(s(t)) = \alpha(t)\, r^{h(t)} + o\left(r^{h(t)}\right) \qquad r \to 0 \tag{3.3}$$

where $\Gamma_r(s(t))$ can be any multi-scale functional complying with this power-law and the multiplicative factor $\alpha(t)$ generally depends on the chosen Γ_r, but for signals conforming to the multi-scale hierarchy explained in section 3.1, the exponent $h(t)$ is independent of it. The term $o\left(r^{h(t)}\right)$ means that for very small scales the additive terms are negligible compared to the factor and thus $h(t)$ dominantly quantifies the multi-scale behavior of the signal at the time instant t. Indeed, close to a critical point, the details on the microscopic dynamics of the system disappear and the macroscopic characteristics are purely determined by the value of this exponent, called the Singularity Exponent (SE) [113]. A central concern in MMF is the proper choice of the multi-scale functional $\Gamma_r(s(t))$ so as to precisely estimate these exponents. We will address this subject in following sections, but for now let us assume the availability of precise estimates of $h(t)$ and develop the link between this geometric representation and the global one in the canonical formalism.

When correctly defined and estimated, the values of singularity exponents $h(t)$ define a hierarchy of sets having a multi-scale structure closely related to the cascading properties of some random variables associated to the macroscopic description of the system under study, similar to the one observed in the canonical framework. Formally, this hierarchy can be represented by the definition of singularity components F_h as the level-sets of the SEs:

$$F_h = \{t \mid h(t) = h\} \tag{3.4}$$

These level sets, each highlight a set of irregularly spaced points having the same SE values. Consequently, they can be used to decompose the signal into a hierarchy of subsets, the "multi-scale hierarchy". Particularly, they can be used to detect the most informative subset of points called the Most Singular Manifold (MSM) and also, they can be used to provide a global statistical view of the complex system by the use of the so-called singularity spectrum.

3.2.1 The Most Singular Manifold

In the MMF, a particular set of interest is the level set comprising the points having the smallest SE values and provides indications in the acquired signal about the most critical transitions of the associated dynamics [131]. These are the points where sharp and sudden local variations take place and hence, they have the lowest predictability: the degree in which they can be predicted from their neighboring samples is minimal. MSM is formed as the collection of points having the smallest values of SE. In other words, the smaller the h(t) is for a given point, the higher the predictability is in the neighborhood of this point. It has been established that the critical transitions of the system occurs at these points. This property has been successfully used in several applications [127, 128, 134]. The formal definition of MSM reads:

$$\mathcal{F}_\infty = \{t \mid h(t) = h_\infty\}, \qquad h_\infty = \min(h(t)) \tag{3.5}$$

In practice, once the signal is discretized, h_∞ should be defined within a certain quantization range and hence MSM is formed as a set of points where $h(t)$ is below a certain threshold.

The significance of the MSM is particularly demonstrated in the framework of reconstructible systems: it has been shown that, for many natural signals, the whole signal can be reconstructed using only the information carried by the MSM [127, 131]. For example, a reconstruction operator is defined for natural images in [127] which allows very accurate reconstruction of the whole image when applied to its gradient information over the MSM. The reconstruction quality can be further improved, using the Γ_r measure defined in [126] which makes a local evaluation of the reconstruction operator to geometrically quantify the unpredictability of each point.

Although simple, the notion of MSM plays an important role in most of the applications we have developed in this thesis. By those applications, we show how the MSM corresponds to the subset of physically important points in the speech signal.

3.2.2 The singularity spectrum

As the points in each singularity component F_h are irregularly spaced and do not fill their topological space, a non-integer value D(h) can be assigned to each one of them as the Hausdorff dimension of these subsets:

$$D(h) = dim_H F_h \tag{3.6}$$

and the map $h \rightarrow D(h)$ is the so called singularity spectrum. $D(h)$ gives an insight about the hierarchical arrangement and the probability of occurrence of each singularity component. It describes the statistics of change in scales, just like the exponent of canonical power-law τ_p in Eq. (3.2) (the Legendre spectrum). $D(h)$ can also be used to bound completely different physical systems through the concept of universality class. It is known that, different physical systems having similar distributions of singular exponents share common multi-scale properties, even if they are completely different physical systems [131].

Parisi and Frich proved that under some assumptions on the shape of $D(h)$ [130], the Legendre spectrum τ_p can be computed from the singularity spectrum [43]:

$$\tau_p = \inf_{\mathbf{h}} \{ph + d - D(h)\} \tag{3.7}$$

where d stands for the dimension of the embedding space. Hence, as in the canonical framework, the singularity spectrum (if well estimated) could analogously define the multi-scale structure of real world intermittent signals (based on the cascade model of energy dissipations). In the canonical formalism, as there is no easy way to have an estimate of singularity exponents at each point, it is difficult to correctly estimate $D(h)$. Instead, the structure functions in Eq. (3.2) can be used to first estimate τ_p and then $D(h)$ is accessible through the inverse Legendre transform of Eq. (3.7). As the inverse transform, computes the convex hull of the variables, this methods imposes a limiting constraint on $D(h)$ to be a convex function. In Microcanonical framework however, as we have access to precise estimates of $h(t)$ (and hence singularity components can be formed), the frequency of occurrence of particular values of SEs can be used for the estimation of $D(h)$ by the histogram method. Indeed, the empirical histogram of SE ($\rho_r(h)$) at small scale r verifies [130]:

$$\rho_r(h) \propto r^{d-D(h)} + o(r^{d-D(h)}) \tag{3.8}$$

As Eq. (3.8) is a proportionality and not an equality, it can not be directly inverted to assign an explicit value to $D(h)$. One solution is to compute $\rho_r(h)$ at several scales and perform a log-log regression. This method is however highly demanding in data and computations. A simpler alternative is to assume that the dimension $D(h)$ of the most probable singularity component is equal to the dimension of the topological space (d). Consequently, the singularity spectrum at each scale can be estimated as:

$$D(h) = d - \frac{\log \frac{\rho_r(h)}{max(\rho_r(h))}}{\log r} \tag{3.9}$$

The spectrum of singularities, being estimated in this manner for a set of fine scales, can be used to check for persistence of multi-scale properties across scales. Indeed, this is the first check to be verified for any given signal to test whether MMF is applicable to it or not.

3.3 The estimation of singularity exponents

As mentioned earlier, the MMF provides numerically stable methods for estimation of the SEs at each point in the signal domain. It consists of methods which are appropriate for empirical data as they filter all the common artifacts that could arise due to discretization, aliasing, noise, lack of stationarity, correlations, instabilities, and other problems related to the nature of real signals or to the numerical analysis of them.

Several approaches are possible in evaluating the power-law scaling of Eq. (3.3), which are theoretically expected to provide the same estimates. However, they each have their own merits and drawbacks in the way they cope with different real-world situations. These approaches may differ either in the employed multi-scale functional $\Gamma_r(\cdot)$, or in the way the multi-scale behavior is being assessed. In following two subsections, we briefly review all the variants and point out their particular merits and disadvantages.

3.3.1 The choice of $\Gamma_r(\cdot)$

One important factor in precise computation of SEs is the choice of the scale-dependent functional operating on the signal. In a purely turbulent signal, with

no more regular dynamics superimposed, different multi-scale functionals should lead to the same values of SEs [131]. But for practical physical processes, different dynamics might be added to the purely turbulent ones and hence, different strategies might be adopted in the choice of this functional.

3.3.1.1 Linear increments

The simplest choice for the multi-scale functional $\Gamma_r(\cdot)$ for evaluation of power-law scaling in Eq. (3.3) is the absolute value of linear increments:

$$\Gamma_r(s(t)) = |s(t+r) - s(t)| \tag{3.10}$$

A signal which exhibits power-law scaling with this multi-scale functional is called a multi-affine function and the corresponding exponent $h(t)$ is usually called the Hölder exponent. The variations of the velocity field measured in fully developed turbulent flow is an example of empirical signals where multi-affinity is evidenced [7]. However, because of sensitivity and instability of linear increments, it is often difficult to use them to obtain good estimation of exponents for analysis of empirical data. Discretization, noise and long-range correlations hinder the practical calculation of these exponents from Eq. (3.3) [130].

3.3.1.2 Wavelet transform

In practice, many physical signals do not conform to the power-law scaling of multi-affine functions (using Eq. (3.10) as $\Gamma_r(\cdot)$). Indeed for these signals, there may exist some extra contributions in form of long-range correlations superimposed to the pure scaling behavior [131]. These correlations are usually differentiable functions whose Hölder exponent is an integer number. The problem arises when this integer value becomes smaller than some of the exponent in the original hierarchy of singularity components. In this case, some parts of the multi-scale hierarchy will be lost.

As it is believed that these long-range correlations appear as an additive contribution, the wavelet transform can be used to eliminate these terms which mask the pure scaling behavior of the complex signal [131]. It is proven that the continuous wavelet transform of a multi-affine signal scales with the same exponent as the one achieved by Eq. (3.10) (the Hölder exponent). But when additive long-range correlations are present, as they can assume a Taylor expansion around any point t, a proper wavelet function can filter them out. Indeed, if the number of vanishing moments of the wavelet is higher than the highest order of the polynomial in the Taylor expansion, it would be orthogonal to the polynomial function. Conse-

quently, the convolution of the wavelet with the signal will vanish these additive terms which mask the singular behavior [9].

Special care is required in the choice of this wavelet. There actually is a compromise between the number of vanishing moments of the wavelet and the spatial resolution in localization of singularities. Indeed, as the number of vanishing moments increases (by differentiating a given wavelet basis), the number of its zero crossings is also increasing which causes the reduction in its spatial resolution (higher number of zero crosses requires higher number of pixels to approximate the theoretical continuous curve in a discrete form [131]). Usually a wavelet with two vanishing moments is chosen, assuming that the highest order of polynomial contribution is 1.

This has only been used in canonical approach using the structure functions over the skeleton of maxima lines, so as to concentrate on most singular subset of data [11, 81, 82, 90]. It is shown that the local maxima of a wavelet transform can detect the location of irregular structures and thus procedures can be provided to compute local scaling exponents at those isolated singularities [82]. This strategy may suffer from the requirement of isolated singularities [91], as it might not be the case for practical complex signals having dense singularities [130].

3.3.1.3 The Gradient-Modulus Measure

An alternative and more robust definition for the functional Γ_r in Eq. (3.3) is proposed by Turiel et al [131], which is defined from the typical characterization of intermittency in turbulence and has shown very good performance in real world settings: the gradient-modulus measure. This measure on a ball of radius r for a turbulent velocity field, describes the kinetic energy dissipation at scale r. Therefore, it is a quantity linked to the transfer of energy from one scale to another. Thus, the exponent associated to the power law in terms of the scale characterizes the information content and the dynamical transitions of the signal [42, 128].

To reveal the inter-scale power-law correlations in signal $s(t)$, this measure is based on summing all of the variations of the signal around a given point. In this way, the measure is extended over a whole region around the point instead of being simply directional like the linear increments in Eq. (3.10). The summations removes spurious fluctuations due to real-world limitations. Formally, these variations are defined as:

$$d\mu_{\mathcal{D}}(t) = \mathcal{D}s(t)dr \tag{3.11}$$

where $\mathcal{D}$ is an appropriate differential operator like the norm of the gradient $|\nabla s|$. Consequently by integrating all these variations, the final multi-scale measure reads:

$$\Gamma_{\mu_r}(s(t)) = \int_{B_r(t)} d\mu_{\mathcal{D}} \tag{3.12}$$

where $B_r(t)$ stands for the ball of radius r centered at t. The use of the norm of the gradient for $\mathcal{D}$ is motivated in [131] by the special power spectrum scaling of natural images. It helps to avoid the domination of the integral in Eq. (3.12) by the mean value of s on the ball $B_r(t)$, which may be far from zero due to the lack of stationarity. The derivative whitens the signal [131] and hence, the measure of a ball will reflect the actual singular structure around the point. Note that the non-linear operation ($|\cdot|$) has an important role, as without it the sum of variations would be dominated by the finite differences at the largest scale, which is highly affected by long-range correlations [130].

It is proven in [131] that in case of a multi-affine signal, this functional exhibits a similar power-law scaling behavior as that of linear increments in Eq. (3.10). However, the associated exponent has a shift equal to the dimension d of the embedding space:

$$\Gamma_r(s(t)) = \int_{B_r} d\mu_{\mathcal{D}}(t) = \alpha(t)\, r^{d+h(t)} + o\left(r^{h(t)}\right) \qquad r \to 0 \tag{3.13}$$

So that, for a 1-D signal, the exponent of this power-law at any point is equal to the Hölder exponent (the exponent of the power-law when $\Gamma_r(s(t)) = |s(t+r) - s(t)|$), with a shift equal to -1.

However, when facing real world data, this gradient-modulus measure shows more success in revealing the existence of scaling behavior. This is shown in [131], in case of satellite images. There, only less than 25% of points in the image verify the power-law scaling behavior with the functional in Eq. (3.10). Therefore, the signal can not be assumed to be a multi-affine one. Meanwhile, when the functional in Eq. (3.12) is used to check for scaling behavior, more than 95% of the points have shown to conform to the power-law scaling behavior (we will further elaborate in chapter 4, on the set of conditions any given signal must meet so that MMF is applicable to it).

3.3.1.4 The Gradient Modulus Wavelet Projections

One can use wavelet projections of the variations instead of the integration in Eq. (3.12). This results in the Gradient Modulus Wavelet Projections (GMWP) functional used in [130]:

$$\Gamma_r(s(t)) = \mathbb{T}_\Psi[|\nabla s|](r,t) \tag{3.14}$$

where $\mathbb{T}_\Psi[x](r,t)$ stands for the continuous wavelet transform:

$$\mathbb{T}_\Psi[x](r,t) := (\Psi_r * x)(t) \tag{3.15}$$

and $\Psi_r(t)$ is scaled version of wave-like function called mother wavelet (Ψ):

$$\Psi_r(t) := r^{-d}\,\Psi(r/t) \tag{3.16}$$

GMWP shows the same scaling behavior (though the shift d is removed) while benefits from the advantages of using a wavelet transform. The superior performance of this functional is shown in [130], in retrieving the singularity spectrum ($D(h)$) of a synthetic signal (in comparison with the canonical methods). It is also discussed that there is no need for the wavelet to have annihilating moments (contrary to the case in section 3.3.1.3). Consequently, the wavelet may even be a positive wavelet, without zero-crosses that limit the resolution.

Another advantage of using the continuous wavelet transform is that it brings the possibility of computing the transform over a set of non-integer scales in discretized signals. Indeed, the scale variable r in Eq. (3.15) may attain any non-integer value, providing a smooth interpolation scheme for the discrete-time signal.

3.3.1.5 Two-sided variations

In chapter 6, we discuss how the existence of long-range correlations might be an obstacle for precise estimation of SEs of the speech signal. We show that an alternative multi-scale functional does lead to an effectively parsimonious MSM which permits speech reconstruction of high quality.

The new functional is based on two-sided measurement of multi-scale variations of the signal $s(t)$:

$$\mathcal{D}_\tau s(t) = |2s(t) - s(t-\tau) - s(t+\tau)| \tag{3.17}$$

and then, similar to Eq. (3.12), the variations are summed together to form the final multi-scale functional:

$$\Gamma_r(s(t)) = \int_0^r |\mathcal{D}_\tau s(t)|\,d\tau \tag{3.18}$$

The details regarding derivation of this new functional are presented in chapter 6. The resulting measure was used on in chapters 6, 7, 9 to detect the location of

physically important samples in speech signal in applications such as waveform coder, parametric coder and detection of Glottal closure instants.

3.3.2 Estimation of $h(t)$

Once the proper multi-scale functional is chosen, the singularity exponent $h(t)$ can be estimated by the evaluation of Eq. (3.3) over a set of reasonably fine scales. There are several approaches for performing such multi-scale evaluation on which we make a brief review in this section.

3.3.2.1 Punctual estimation

Ideally, if the power-law scaling behavior of Eq. (3.3) holds for a given measure, the measurements at different scales are directly related. In this case, the estimation of SE can be done by punctual evaluation of Eq. (3.3), to solve it for $h(t)$ using only one measurement at the finest scale [130]. For this, we make use of the assumption of the statistical translational invariance [130, 131] of the measure in Eq. (3.12), which states that the first order averages must not be dependnt on scale or time. As such, $\langle|\Gamma_r(t)|\rangle$ is independent of the ratio of scales and is only proportional to the bulk r^d. So we have:

$$\langle|\Gamma_r(t)|\rangle = \alpha_0 r^d \tag{3.19}$$

Also, using Eq. (3.12) we have $\langle|\Gamma_r(t)|\rangle = \langle|\alpha(t)|\rangle r^d \langle|r^{h(t)}|\rangle$. Again, using the statistical translational invariance assumption we have $\langle|r^{h(t)}|\rangle = 1$ and so:

$$\langle|\Gamma_r(t)|\rangle = \langle|\alpha(t)|\rangle r^d \tag{3.20}$$

The above two formulas imply that $\alpha_0 = \langle|\Gamma_r(t)|\rangle = \langle|\alpha(t)|\rangle$. We can make use of this property for making a punctual estimate of $h(t)$. By taking the logarithm of both sides of Eq. (3.3) and after some manipulations we have the following expression for $h(t)$:

$$h(t) = \frac{\log \frac{\Gamma_r(s(t))}{\langle\Gamma_r(s(t))\rangle}}{\log r_0} - \frac{\log \frac{\alpha(t)}{\alpha_0}}{\log r_0} \tag{3.21}$$

As the second term in Eq. (3.21) contains the term $\frac{\alpha(t)}{\alpha_0}$ which should be close to 1 (because $\alpha(t)$ fluctuates around its average α_0), its logarithm can be considered

negligible for very small scales. Thus, the first term can be computed as an approximate of the singularity exponent. The ability of this method in estimation of $h(t)$ is demonstrated in [130] showing that it can give a reasonable estimate of singularity spectrum of synthetic signals with known multi-scale dynamics. However, clearly there will be perturbations due to the second term, specially when in practice, the limited resolution of acquisition devices do not let having access to very fine resolutions.

3.3.2.2 log-log regression

Since at least for the linear increments used in Eq. (3.21), the inter-scale relation of measurements holds only for distributions, such punctual estimation would be unstable and the resulting SE would be perturbed (the discussion about the determination of ideal measure for a given signal, such that the inter scale relationship is maximized can be found in [97]). Hence, it is required to incorporate several scales in the evaluation of the power-law, to perform a stable estimation: this way, the perturbations of the measure at different scales cancel each other and a more stable SE estimation results.

This multi-scale evaluation can be done by performing a $\log - \log$ regression over several scales. It is easy to see that taking the logarithm of both sides of Eq. (3.3) reveals a linear relationship between the logarithm of the multi-scale functional and the logarithm of the scale. So it is possible to estimate the singularity exponent $h(t)$ at each time t by performing a linear regression of the $\Gamma_r(\cdot)$ versus the scales r in a log-log plot. Therefore, Eq. (3.3) is verified for a given signal if we attain acceptable correlation coefficients for such linear regression. So that, the advantage is that the correlation coefficients of the regression provide a clue for verifying whether the scaling behavior in Eq. (3.3) truly holds or not. This would be a starting point for testing the validity of the formalism for analysis of a given empirical signal.

In addition, when a wavelet based measure is used, as the continuous wavelet transforms can be computed over a set of non-integer scales (even for discretized signals), the scale variable r in the regression could be assigned any non-integer value, providing a smooth interpolation scheme for the discrete-time signal.

However, the resulting improvement in estimation is only achieved by compromising the resolution. In fact, usually we have no access to the physical finest scale (due to discretization limitations) and hence the measurements at coarser sampling scales may simply combine information carried by adjacent pixels. Moreover, this regression is computationally costly and only serves to enhance the resolution of less singular structures at the cost of coarsening most singular ones [130].

3.3.2.3 Inter-scale modulations

By using the punctual estimation we preserve the finest accessible resolution in SE estimation. However, in practice we do not have access to the physical finest scale (due to discretization limitations) and hence the resulting estimate might be unstable and perturbed by the additive term due to $\alpha(t)$. This perturbation can be corrected by incorporation of the same measurement at coarser scales. But as in the case of log-log regression, this has the risk of compromising the resolution and the measurements at coarser sampling scales may simply combine the informations carried by adjacent pixels. For some applications, such as the phoentic segmentation algorithm of chapter 5, the resolution is of the highest importance. So we decided to only incorporate the measurement at the second finest scale ($\Gamma_{\mu_{r_1}}$) to regularize $\Gamma_{\mu_{r_0}}$ as:

$$\Gamma_r^{\kappa}(s(t)) = \kappa_t \Gamma_{\mu_{r_0}}(s(t)) \tag{3.22}$$

where the regularization term κ_t is defined as the quotient of the measurements at two scales:

$$\kappa_t = \sqrt{\frac{\Gamma_{\mu_{r_1}}(s(t))}{\Gamma_{\mu_{r_0}}(s(t))}} \tag{3.23}$$

where r_1 is one coarser scale ($r_1 > r_0$). In the ideal case, where both measurements are completely correlated, κ would be equal to one (after normalization of the measurements at the two scales). Otherwise, this term corrects possible perturbations of the measurement made at the finest scale r_0. We thus replace $\Gamma_{\mu_{r_0}}$ in Eq. (3.21) with $\Gamma_r^{\kappa}(s(t))$ to compute $h(t)$.

3.3.2.4 Partial singularities

Another possibility for multi-scale evaluation of $\Gamma_r(\cdot)$ is to use the concept of microcanonical cascades introduced in [97], which associates the power-law scaling of Eq. (3.3) to the existence of an underlying microcanonical cascade process. The latter refers to a cascading process similar to Eq. (3.1), which rather than being valid for distributions, is valid for any given point in the signal. This is being made possible by assuming the existence of inter-scale correlations in form of Eq. (3.3). In such cascade processes, energy or information is transferred between scale levels of the signal. This way, the MSM actually corresponds to the set of points where information concentrates as it transfers across scales and, in that sense, it is a *least predictable/reconstructible manifold*. The cascade variable of this process must follow

an infinitely divisible distribution; a property which permits a simple estimation of the desired scaling exponents, as the sum of a set of *transitional* exponents [97]:

$$h(t) = \frac{1}{k} \sum_{i=1}^{k} h_{r_i}(t) \tag{3.24}$$

where $h_{r_i}(t)$ are the *transitional* exponents, which can be computed by direct evaluation of Eq. (3.3) at each scale, using anyone of the multi-scale functionals in section 3.3.1 as:

$$h_{r_i}(t) = \frac{\log(\Gamma_{r_i}(s(t)))}{\log(r_i f_s)} \tag{3.25}$$

where f_s is the sampling frequency of the signal.

3.4 Conclusion

In this chapter we briefly introduced the theoretical foundations of the Microcanonical Multi-scale Formalism, along with a full presentation of its main components (singularity exponents and the Most Singular Manifld) for geometric characterization of complexity. In the meantime, we pointed out the relationship of the MMF with its predecessors in the study of complex systems. Now we can move on to study the application of the MMF to the specific case of the speech signal. But first, we investigate in next chapter whether the application of the MMF to the speech signal is valid and feasible (from both theoretical and numerical point of views).

4

Validation of the MMF for speech analysis

In this chapter we present the theoretical requirements that any given signal should meet so that the application of MMF is meaningful for its analysis. Other than validation of the applicability of the MMF, these conditions may also serve another objective: as mentioned in chapter 2, the existence of complex phenomena in physical production mechanism of the speech signal is theoretically and experimentally established. As the conditions we will examine in this chapter are actually the fingerprints of the existence of such complex mechanisms in the signal itself, they may be seen as indirect evidences of the complex character of the speech signal, which can be observed through the execution of a set of numerical tests.

We will carefully examine whether the speech signal is a proper candidate for being analyzed under the MMF or not. However, special care must be taken for the specific case of the speech signal, because of its wide range of variability on different levels. Therefore in this chapter, we will first present these conditions in section 4.1 in their general form. We then describe in section 4.2, the details of the experiments we follow for evaluation of these conditions for the specific case of the speech signal: the dataset we have used, the experimental protocol we have followed and the results we have achieved. Finally we make our conclusion in section 4.3.

4.1 Theoretical requirements

As mentioned in chapter 3 section 3.1, the MMF and its older counterparts are primarily developed for description of intermittent character of the velocity field in turbulent flows. So that, the first thing to check is the intermittency of the signal under study. The intermittency property imposes a condition on the spectrum of singularities $D(h)$. Indeed, $D(h)$ is related to the classical Legendre spectrum of Eq. (3.2) through the Legendre transform. As the Legendre spectrum is a convex curve for intermittent signals and this property remains intact under the application of the Legendre transform, the singularity spectrum $D(h)$ must also be a convex function of h.

$D(h)$ is also required to be a description which is independent from scale. As such, it is required that the evaluation of Eq. (3.9) for estimation of $D(h)$ results

in the same shape of $D(h)$ for different fine scales r. Finally, the most important requirement is the existence of precise estimations of singularity exponents for any given point in the signal's domain. This verifies the *local* existence of the scaling behavior and thus the meaningfulness of the estimations made by the methods presented in chapter 3 section 3.3.1.

In summary, other than the intermittency check, the validity of the application of the MMF for the analysis of any given signal, requires the verification of following three conditions:

- verification of Eq. (3.3) with enough accuracy for every point in the signal's domain and for a reasonable set of fine scales. This proves the existence of local scaling behavior and thus, shows that precise estimation of singularity exponents are possible and the resulting estimates are meaningful quantities.
- persistence of estimated $D(h)$ for a set of fine scales, thus verifying the statistical scale-invariance character of the signal.
- D(h) has to be a convex curve so that it could be exclusively associated with τ_p of p-th order structure functions through the Legendre transform [131].

We can now proceed to evaluate these experimental requirements for the specific case of the speech signal.

4.2 Test on the speech signal

For the practical examination of the criteria mentioned in section 4.1 in case of the speech signal, we need to cope with its intrinsic non-stationarity. Speech signal as a whole is formed by concatenation of basic sound units called **phonemes**. As the production mechanisms for any phoneme involves essentially different configurations and processes in human vocal system, the corresponding pressure waveform, i.e. the corresponding speech signal would posses different statistical and characteristically properties. Therefore, it is reasonable to separately investigate the validity of the above conditions for each one of these essentially different signals.

On the other hand, since phonemes are extremely short, it is not possible to perform statistical analysis on their single realizations. Hence, we are obliged to use several realizations of each phoneme and accumulate the measurements in a legitimate way (we will later elaborate on the protocol we have followed for such accumulation and we show how it is justified). As such, we constructed a large database of 3000 isolated phonemes, extracted from the TIMIT [45] database (sampling frequency is equal to 16 kHz). We used the time-aligned phonetic transcriptions provided in this database, to draw isolated phonemes realizations: our database includes 500 realizations of six different phoneme families, each of which

is a representative of a family of phonemes. These families include: vowels, fricatives, stops, semi-vowels and glides, affricates and nasals. The left panels in Fig. 4.1, show one example of waveform shape for any one of these representatives. So we take these 500 realizations of each representative and we accumulate the statistics required for evaluation of the criteria in section 4.1. We will later get back to this subject in section 4.2.3, to explain how such accumulation is justified.

4.2.1 Intermittency test

We first investigate whether the representatives of each phoneme family can be considered as an intermittent signal or not. An intermittent signal displays activity only in a fraction of time: they are close to a reference value for long times and suddenly they undergo sharp, short-lived extreme deviations [42]. Therefore, the simplest manifestation of intermittency can be evidenced by observing a small mode and a slowly decaying tail in the gradient histogram of the signal [98]. The right panels in Fig. 4.1, show histograms of the absolute gradient values for the ensemble of instances of each phoneme representatives, while the middle panels show the histogram of the signal values.

It can be seen in Fig. 4.1 that although the histogram of the signal values may be very different for different phoneme families, the histogram of the absolute gradient values conforms to the intermittency condition: they all have a very small mode close to zero and a slowly decaying tail.

A more formal method for the evaluation of intermittency is introduced in [42], which requires the following flatness function to grow with frequency Ω:

$$F(\Omega) = \frac{\langle (s^{>}_{\Omega}(x))^{2p} \rangle}{\langle (s^{>}_{\Omega}(x))^{p} \rangle^{2}} \tag{4.1}$$

where $s^{>}_{\Omega}(x)$ is the high-passed filtered version of the stationary signal $s(x)$ with the cut-off frequency equal to Ω. In fact, we can rewrite Eq. (4.1) as:

$$F(\Omega) = 1 + \frac{variance(s^{>}_{\Omega}(x)^{p})}{average(s^{>}_{\Omega}(x)^{p})^{2}} \tag{4.2}$$

clearly, $F(\Omega) \geqslant 1$. If the active time of the signal is being reduced with the increase in cut-off frequency of the high-pass filter, we intuitively expect the average to be reduced faster compared to the variance. Hence the flatness function will be growing with cut-off frequency Ω. Also if we use the canonical power-law relationship of Eq. (3.2) for structure functions, we can rewrite this function as:

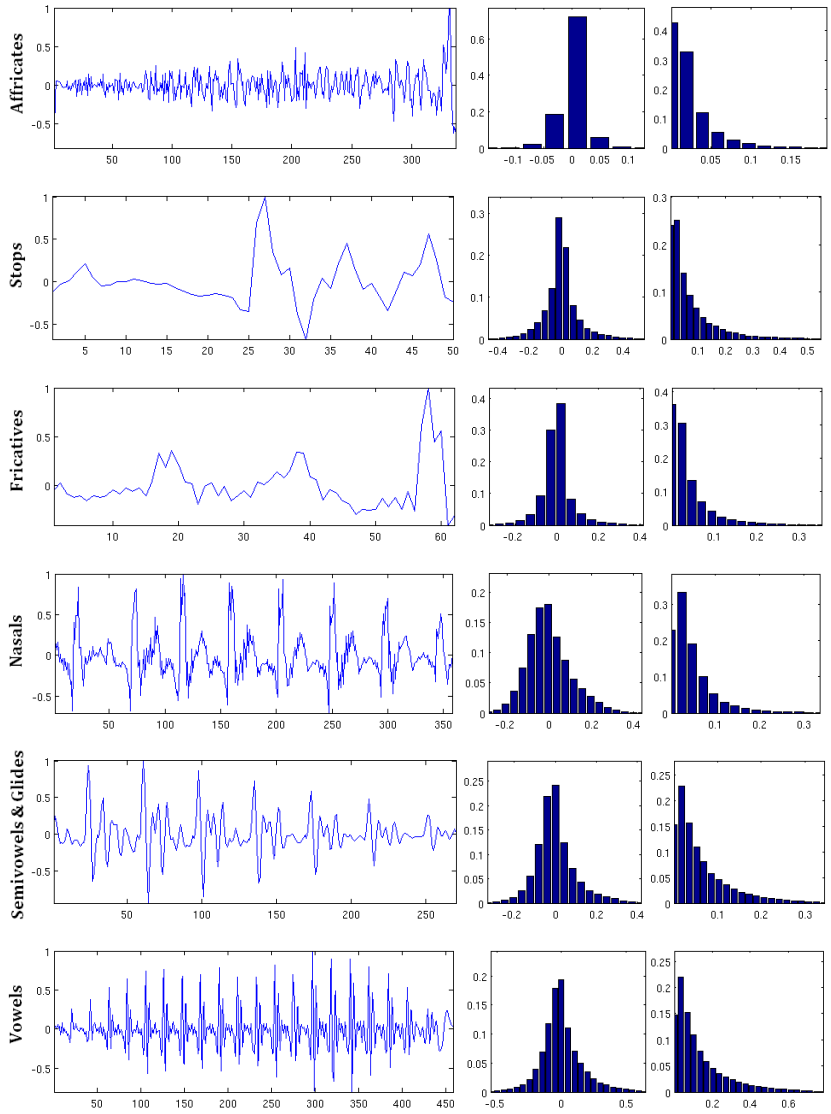


Figure 4.1: **Left panels:** an example waveform shape of each phoneme family, **middle panels:** histograms of the signal values and **right panels:** histograms of the absolute gradient values. The histograms are computed on the ensemble of 500 instances. The bin widths are kept constant for different phoneme families (although different zooms are made to enhance the representations). Phonemes are extracted from the sentences, after normalization to unity of the whole speech signal.

$$F(\Omega) = \frac{\frac{1}{\Omega}^{\tau_{2p}}}{\frac{1}{\Omega}^{2\tau_p}} = \Omega^{2\tau_p - \tau_{2p}} \geqslant 1, \Omega \to 0 \tag{4.3}$$

in the case of $F(\Omega) = 1$ we have the linear relationship $\tau_{2p} = 2\tau_p$, which corresponds to the linear shape of τ_p as in Kolomogrov's universal scaling for non-

intermittent signals. Otherwise, if $F(\Omega) > 1$, $\tau_{2p} > 2\tau_p$ which is the case for intermittent signals. This measure also serves to discriminate Gaussian source: since the Gaussian property is conserved by any linear operation, the filtered version will be still Gaussian, hence the value of $F(\Omega)$ will be constantly equal to 1.

Fig. 4.2 shows the flatness function for each representative phoneme. Since not all of the phonemes are full-band signals, to avoid numerical artifacts, the flatness function is computed only up to the frequency where the filtered signal has a remarkable variance. It can be seen that in the active frequency band of each phoneme, flatness function is roughly an increasing function of frequency. This further supports the possibility of the existence of the intermittent character for the speech signal.

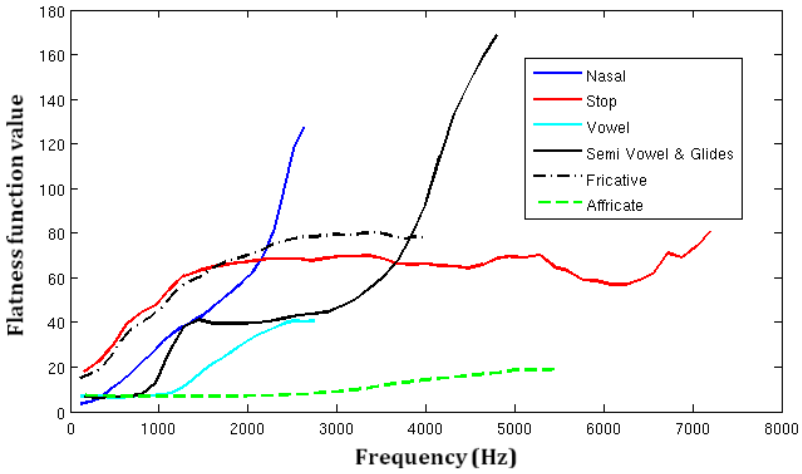


Figure 4.2: The flatness function for different phoneme families. To avoid numerical artifacts, the flatness function is computed only up to the frequency where the filtered signal has a remarkable variance.

4.2.2 Local scaling test

Next, we evaluate the point-wise existence of power-law scaling behavior in form of Eq. (3.3), to check the availability of precise estimations of the singularity exponents. In MMF a simple mechanism is provided for such verification, using the log-log regression method of SE estimation, as explained in section 3.3.2.2. Indeed, if such inter-scale relationship actually exists, the resulting regression coefficient of the log-log regression would attain a very high value (with maximum equal to unity). We thus employ this method for estimation of SEs on the wavelet transform of the Gradient Modulus Wavelet Projections (GMWP) functional (section 3.3.1.4).

The wavelet we use is the Lorentzian wavelet. This wavelet provides an accurate estimation for smaller exponents, at the expense of a saturation of all exponents that are greater than unity [129]. This is desirable because small exponents are the

most informative ones and in the presented case, retrieved exponents are far from the saturation and so it does not appear as an actual limitation. To perform the regression we chose 10 scales which are log-uniformly spaced between 1 and 100 samples (which correspond to the interval from 62.5 μs to 6.25 ms considering that the sampling frequency is 16 Khz). The resulting average regression coefficients are reported in Table 4.1 for each family of phonemes. It must be noted that for this experiment, we perform the regression on the inverse of $\log\Gamma_r$ versus inverse of $\log r$, as we expect the same linear relationship to hold, while the effect of $\log\alpha(t)$ is reduced. It can be observed that for all of the phoneme families the resulting averages are so close to unity, meaning that precise estimations of SEs are available.

We also performed the same procedure over the whole sentences. 500 different speech signals with an approximate length of 3–5 seconds were used for this experiment. We obtained an average regression coefficient of 0.96.

Overall, these experiments show that excellent regression coefficients are obtained using our estimation procedure. This suggests that not only the power-law scaling in Eq. (3.3) is valid for speech signal, but also we achieve very precise estimation of the singularity exponents.

Phoneme type	Vowel	Fricative	Stop	Semi Vowel & Glide	Affricate	Nasal
Phoneme	/aa/	/dh/	/b/	/el/	/ch/	/en/
Average regression Coef.	0.97	0.95	0.99	0.98	0.99	0.99

Table 4.1: The average regression coefficients of the linear regression for 500 sample phonemes.

4.2.3 Test on multi-scale persistence of $D(h)$

Now that the availability of precise SE estimations is verified, we can proceed to evaluate the persistence of singularity spectrum $D(h)$ across different scales. For the estimation of $D(h)$, we use the histogram method of Eq. (3.8). However, single phonemes are too short to be used for meaningful formation of histograms. Therefore, we need to accumulate the histograms of several instances of each phoneme while computing $D(h)$ by the histogram method. The question is, are we allowed to assume that different instances of the same phoneme have the same statistics? To answer this question, we perform a standard statistical hypotheses test, called the Kolomogrov-Simirov(KS) Test [34].

The KS test is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare two data samples, to check wether they are drawn from the same distributions or not. It quantifies a distance between the empirical distribution functions of two samples. The KS statistic is defined as:

$$KS_{D_{n1,n2}} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)| \tag{4.4}$$

where F_{k,n_k} is the cumulative distribution function of the kth data sample and n_k denotes the length of each data sample. The two samples are considered to be emitted from the same distribution at level α if

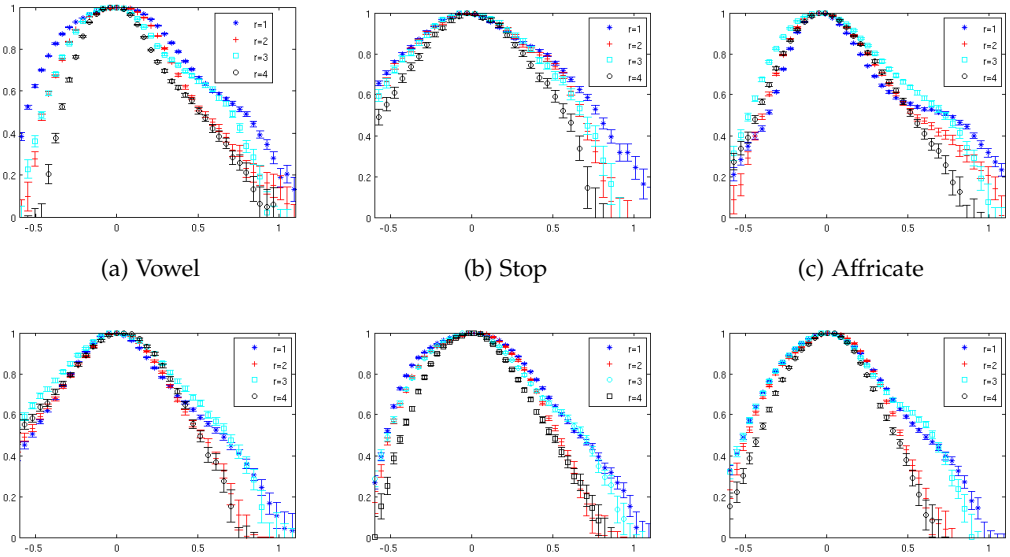
$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} KS_{D_{n_1,n_2}} < K_\alpha \tag{4.5}$$

where K_α could be found from $Pr(K < K_\alpha) = 1 - \alpha$. If we take $\alpha = 0.001$, it corresponds to 0.999 fidelity level and the associate K_α is found to be equal to 1.95 [34]. For each phoneme family, We perform this test on the computed singularity exponents of 500 phoneme pairs of each phoneme family which are chosen randomly from the database. Also, to remove the effect of averages in our comparison (and concentrate the comparison on the *shape* of distributions), we perform the same experiment on the centered version of singularity exponents of phoneme pairs. The resulting average of KS static is reported in table 4.2 for both singularity exponents and also the raw speech signal itself. Observing the results in table 4.2 we can fairly conclude that the *centered* version of singularity exponents have similar statistical properties. In other words, although different instances of a single phoneme may have different averages of singularity exponents, they share similar shape of probability distribution function. Note that this is not the case for the speech signal itself, even when we remove the average.

Table 4.2: The average KS test statistic on singularity exponents, for the representatives of 6 phoneme families. For each representative, 500 pairs are randomly chosen from the TIMIT database. The pairs can be assummed to be drawn from the same distributions if the static is smaller than 1.95 (With 0.999 fidelity level).

		KS test statistic			
		speech signal		singularity exponents	
Phoneme type	phoneme	Original	Centered	Original	Centered
Vowel	/iy/	3.43	3.42	4.62	1.28
Fricative	/sh/	4.25	4.23	6.00	1.43
Stop	/d/	2.2	1.96	3.18	1.76
Semi Vowel & Glides	/el/	3.75	3.71	2.99	2.98
Affricate	/ch/	3.43	3.42	6.25	1.88
Nasal	/m/	2.52	2.48	3.55	1.48

Now we can accumulate the SE histograms of different instances of the same phoneme, in order to compute $D(h)$. To evaluate the persistence of $D(h)$ across different scales, we compute $D(h)$ in four consecutive scales. From the finest possible scale, to four times the finest possible scale. The resulting curves are shown in figure 4.3, along with the corresponding error bars.



(a) Vowel (b) Stop (c) Affricate

(d) Fricative (e) SemiVowel & Glide (f) Nasal

Figure 4.3: The test on persistence of the shape of singularity spectrum across fine scales.

Besides the clear convexity of all of the singularity spectrum for each phoneme family, a remarkable coincidence between spectrums at different scales can be observed in Figure 4.3. Of course, some differences could be observed between $D(h)$ at different scales, but these differences are mainly concentrated on the tails of singularity spectra, corresponding to the less probable values of h. Since these parts are less probable, there exists more uncertainty in their experimentally estimated histograms. Hence the same uncertainty is imposed to the resulting spectrum. For a fair, formal evaluation of the closeness of the spectra in different scale, we must take these uncertainties into account.

A measure of mutual closeness of pairs of singularity spectrum is defined in [130], which is called the Directed Weighted Average Difference (DWAD). Given a pair of singularity spectrum $D_{r_1}(h)$ and $D_{r_2}(h)$, along with their associated uncertainty $b_{r_1}(h)$ and $b_{r_2}(h)$, the DWAD is defined as:

$$\delta D_{r_1,r_2} = \sum_{h_n} \frac{|D_{r_1}(h_n) - D_{r_2}(h_n)|}{b_{r_1}(h_n) b_{r_2}(h_n)} \Big/ \sum_{h_n} \frac{1}{b_{r_1}(h_n) b_{r_2}(h_n)} \tag{4.6}$$

where the uncertainties $b_{r_1}(h)$ and $b_{r_2}(h)$ are the result of the propagation of error, by considering histogram formation as a multinomial stochastic process. $\{h_n\}$ are the points at which $D_{r_1}(h)$ is sampled. As the points $\{h_n\}$ are not the sampling points of $D_{r_2}(h)$ in general, $Dr_2(h)$ and $b_2(h)$ are linearly interpolated to evaluate their value at each h_n. Also, the sum in Eq. (4.6) is restricted to the values of $\{h_n\}$ in the common range of h values for $D_{r_1}(h)$ and $D_{r_2}(h)$. If for a pair of $D(h)$, the mutual distance is smaller than the uncertainty in computation of these $D(h)$, we can conclude that the two spectrums are fairly similar.

The resulting DWADs are reported in Table 4.3. It can be seen in Table 4.3, that for almost all of the cases, the directed distance is less than the average error bar. Hence we can conclude the persistence of singularity spectrum with the change in scale.

4.3 Conclusions

In this chapter we presented a set of four conditions that can be experimentally verified to validate the existence of an underlying multi-scale structure inside any given signal, which in turn validates the use of MMF for the analysis of this signal. Upon the validation of these conditions for speech, these conditions can also be considered as the fingerprints of the complex phenomena in production mechanism of the speech signal that can be directly evaluated using the speech signal itself (other than the evidences which are introduced in chapter 2).

As practical evaluation of some of these conditions requires the availability of large data samples, their precise estimation for speech is circumvented by its rapidly time-varying characteristics. We thus took special care in examination of these conditions for speech signal to make the tests as reliable as possible. The results showed that all of the four conditions are met by the speech signal and at least, none of them is rejected. Nevertheless, the local scaling property was verified in section 4.2.2, out of any stationarity assumption and without any accumulation of statistics. In fact, the latter is the most important property and shows that it is indeed possible to precisely estimate the singularity exponents at each point in the domain of the speech signal as meaningful local variables. We thus proceed to study the usefulness of these local parameters for the analysis of the speech signal.

Table 4.3: The DWAD for pairs of singularity spectrums.

	$r_1 = 1$				$r_1 = 2$				$r_1 = 3$				$r_1 = 4$			
r_2	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Vowel	uncertainty : 0.036				uncertainty : 0.023				uncertainty : 0.026				uncertainty : 0.025			
	-	0.02	0.03	0.06	0.02	-	0.02	0.04	0.02	0.02	-	0.03	0.06	0.04	0.03	-
Stop	uncertainty : 0.043				uncertainty : 0.06				uncertainty : 0.04				uncertainty : 0.05			
	-	0.013	0.016	0.043	0.013	-	0.01	0.03	0.02	0.01	-	0.03	0.03	0.04	0.035	-
Affricate	uncertainty : 0.035				uncertainty : 0.047				uncertainty : 0.028				uncertainty : 0.03			
	-	0.02	0.04	0.02	0.02	-	0.02	0.01	0.04	0.02	-	0.03	0.02	0.01	0.03	-
Fricative	uncertainty : 0.043				uncertainty : 0.035				uncertainty : 0.057				uncertainty : 0.043			
	-	0.02	0.03	0.03	0.02	-	0.03	0.01	0.03	0.03	-	0.02	0.03	0.01	0.02	-
SemiVowel	uncertainty : 0.036				uncertainty : 0.031				uncertainty : 0.032				uncertainty : 0.041			
	-	0.01	0.02	0.04	0.01	-	0.01	0.03	0.02	0.01	-	0.03	0.046	0.03	.03	-
Nasal	uncertainty : 0.043				uncertainty : 0.034				uncertainty : 0.037				uncertainty : 0.033			
	-	0.01	0.01	0.03	0.01	-	0.02	0.03	0.01	0.02	-	0.03	0.03	0.03	0.03	-

Part II

APPLICATIONS

Until now, we have throughly introduced the MMF and its main components: the Singularity Exponents (SE) and the most informative subset (the MSM). In following chapters, we present our exploratory research on discovering the relevance of this formalism for speech analysis, by showing its potential in some attractive applications. We start by studying the local behavior of SEs in chapter 5. We show how they readily convey instructive information about local dynamics of speech that can be used for detection of transition fronts on phoneme level. We thus develop a simple phonetic segmentation algorithm with interestingly high geometrical resolution in detection of phoneme boundaries.

We then proceed to study the significance of the other major component of the formalism, the MSM, with respect to the speech signal. First, we show in chapter 6 that indeed speech can be compactly represented by the MSM. We use a *generic* interpolator which reconstructs speech from its MSM and we use it to develop an efficient *waveform* coder. We then naturally turn our attention to the physical significance of the points in MSM w.r.t speech as a realization of a physical system rather than a generic 1-D signal. In chapter 7, we show how the MSM recognizes the instants of significant excitations of the vocal tract system, called the Glottal Closure Instants (GCI). We develop a robust algorithm for GCI detection and we show its good performance by comparing the detected points to the reference points gathered by precise sensory measurements of the vibratory motion of vocal folds. Consequently, in chapter 8, we present a case study on the application of this GCI detection algorithm to sparse Linear Prediction Analysis (LPA). We show how incorporation of MSM-based GCI detection in optimization procedure of LPA, can provide more realistic estimates of speech's source-filter decomposition. Finally in chapter 9, we proceed to study the significance of MSMs of higher cardinality: rather than restricting the cardinality of the MSM to one sample per pitch period (as for GCI detection), we use MSMs of higher cardinalities to simplify the procedure for estimation of vocal tract's excitation sequence. This is studied in the framework of multi-pulse coding and results in a more efficient estimator of speech excitation sequence compared to the classical techniques. We finally make our conclusions in chapter 10.

5

Text Independent phonetic segmentation using the MMF

We discussed in chapter 3 that according to the MMF, the dynamics of complex signal is related to the existence of an underlying multi-scale hierarchy that explains the inter-scale correlations of the signal under study. The finger-prints of this multi-scale hierarchy was evidenced in chapter 4 for the case of the speech signal. It was also shown that most of the points in the speech signal conform to the local scaling power-law of Eq. (3.3) and hence, we may achieve very precise estimation of the Singularity Exponents (SE) for this signal. This is thus natural to expect that the non-linear dynamics of the speech signal might be locally accessed [to some extent] through the use of this geometrical formalism. In this chapter we present our first observations regarding the informativeness of these exponents about local dynamics of the speech signal. We show how the time evolution of the local distributions of SEs readily contain instructive information regarding the boundaries of phonemes and so, we use this feature to automatically identify the transition fronts between adjacent phonemes. We thus develop an automatic algorithm by employing a classical two-step methodology in change detection to identify phoneme boundaries. Experimental results show that although simple, our algorithm has competitive results compared to the state-of-the-art methods and in particular, it is more accurate for lower tolerances of error.

This chapter is organized as follows: section 5.1 introduces the problem of phonetic segmentation and provides a review of classical segmentation methods. In section 5.2, we present our primary observations regarding the informativeness of SEs about phoneme-level transition fronts of speech, which forms the basis for the automatic phonetic segmentation algorithm we describe in section 5.3: the SE-based measure of change is introduced in 5.3.1 and the two-step approach for decision making is presented in 5.3.3. Consequently, extensive experimental results are provided in section 5.4. We finally make our conclusions in section 5.5.

5.1 Phonetic Segmentation

Most of the recent developments in speech technology (recognizers, synthesizers or coders) strongly rely on corpus-based methodologies which require the availability of the precisely time-aligned labels of speech building units. Also, annotation on the level of individual segments will be useful not only for studying segmental properties of speech (e.g., temporal characteristics, spectral changes within a speech sound), but also for many kinds of tasks associated with prosodic research [78]. The task of time-aligned labeling of speech signal to its smallest unit of sound (phonemes) is called phonetic segmentation.

The most precise method of phonetic segmentation is the manual collection and segmentation of speech corpora. However, manual segmentation has several disadvantages: it is extremely time-consuming and demanding in terms of labeler expertise. Moreover, since the task is inherently subjective and therefore inconsistent and irreproducible, it suffers from both inter-labeler and intra-labeler inconsistency issues [78]. Hence, automatic phonetic segmentation is of great importance and interest.

The most frequent automatic approach is to use an HMM-based phonetic recognizer to adapt it to phonetic segmentation task by letting the recognizer know the phonetic transcription and performing a forced alignment [123]. Although such methods provide very good segmentation performances, they suffer from imposing linguistic constraints to the segmentation algorithm. This makes the algorithm restricted to the database used for training and it can not be applied to different databases of different languages, contexts or accents. Moreover, it is not always possible to have access to the phonetic transcriptions to perform forced time-alignment. Such HMM based methods, or any other approach which rely on an externally supplied transcription of the sentence for determining phoneme boundaries are called Text-Dependent (TD) segmentation methods.

Text Independent (TI) methods on the other hand are exclusively based on the acoustic information contained in the speech signal itself. This information includes either some model-based feature vectors or raw spectral measures. Such methods are suitable for all those applications that may benefit from explicit speech segmentation, when a phonetic transcription is unavailable (as for speech recognition) or inaccurate (as in speech annotation improvement tasks) [35]. The proposed method in this paper, belongs to the TI class of segmentation methods. Hence, we make a brief review on the available methods in this field.

5.1.1 Review of Text Independent Methods

TI methods for phoneme segmentation are usually based on several measures of distortions quantifying the spectral change between consecutive frames. This mea-

sure of distortion or distance can be defined in accordance to a model such as linear predictive coding (model-based) or it can be purely based on raw spectral properties of the signal (model-free) [69]. The points where the change is maximized is then considered as phoneme boundaries. So that, besides the choice of the measure of change, the procedure for selection of the points where distance is maximized is the other main source of variation in TI methods.

Model-based methods accomplish the task of phonetic segmentation through sequential detection of changes, in a model they assume for the signal. For instance, in [5] speech is assumed to be described by a string of homogeneous units, each of which is characterized by statistical AR model whose residuals are uncorrelated zero mean Gaussian sequences. Consequently, three methods are analyzed to detect the change in parameters of this model. For instance, the Generalized Likelihood Ratio (GLR) test as the basic statistical approach for sequential signal segmentation is introduced: assuming that the variance of residuals are constant inside homogeneous units, for any given segment of speech, a likelihood ratio hypothesis test is performed to test whether that segment is better modeled with a single AR model or with two of them separated on an intermediate point. Once the model order, size of analysis window and a threshold for the test are accurately chosen, this test can segment the speech signal into subphonemic units [1] (two units roughly per phoneme).

The above procedure for GLR test is similar to the dynamic windowing technique used in [1] where Bayesian Information Criterion (BIC) is used as a hypothesis test to refine the initially selected candidates. First, a preselection of candidates (of change) is performed by taking the maxima of several distance metrics computed between adjacent windows with fixed width (some of these metrics are introduced in the following). The distance between these windows must be less than their fixed width. A plot of distances is then created and significant local peaks are selected as candidates so as to filter out the points with insignificantly small distance values. In the next step, a second windowing scheme with tunable width is used, where BIC is used to validate or discard the candidates determined in the first step: the sequence of feature vectors (typically Mel-Frequency Cepstral Coefficients or MFCCs) in adjacent speech segments are assumed to follow multivariate Gaussian Distribution (GD), while their concatenation is assumed to obey a third multivariate GD. The problem is to decide whether the data in the large segment fit a single GD better (H_0), or whether a two-segment representation describes it better(H_1). Bayesian model selection is used to evaluate this hypothesis test and each candidate is rejected if H_0 is selected (on the window formed using the two immediate neighboring candidates).

1 It must be noted that a phoneme is actually a linguistic unit and not an acoustic one. As such, an algorithm for detection of acoustic units may not be a phonetic segmentation algorithm: it may happen that a single phoneme contain more than one subphonemic acoustic units.

The model-free methods do not assume a model for the data and they rely on raw measures of spectral change. They then use some heuristics to detect the locations of maximal changes as the phoneme boundaries. The drawback of these model-free distortion measures is that they are strongly dependent on the spectral densities. When mixed-type spectra are present in the signal, large distortion values obtained from these measures may not necessarily correspond to significant changes in the signal [35]. Spectral Variation Function (SVF) is one of the popular model-free distortion measures which quantifies the overall spectral change in magnitude from frame to frame. It is defined as an angle between two normalized Cepstral vectors separated by some frame distance as [40]:

$$SVF(k) = \frac{\hat{C}(k-1) \bullet \hat{C}(k+1)}{\| \hat{C}(k-1) \| \cdot \| \hat{C}(k+1) \|}$$

where $\bullet$ indicates the scalar product and the spectral vectors $\hat{C}(k)$ are the difference between the cepstrum and its average over a multi-frame window. SVF computation gives rise to a curve with peaks corresponding to areas of rapid, intense spectral change. Traditionally, the minima of its second derivative can be used to perform phonetic segmentation [12]. Another variant of SVF is defined in [107] as the norm of the delta-cepstral coefficients for the frame k:

$$SVF_{\Delta cep}(k) = \sqrt{\sum_{m=1}^{D} [\Delta C_m(k)]^2} \tag{5.1}$$

where D denotes the number of coefficients and ΔC_k represents the derivative of the cepstral coefficients for frame k. Similarly, [87] introduces the Delta Cepstal Function (DCF). The difference between the cepstrum around time t is first computed:

$$d_m(k) = C_m(k+1) - C_m(k-1), \quad m = 1, \cdots, D \tag{5.2}$$

Next, d_m is time-normalized over the duration of the signal:

$$\hat{c}_m(k) = d_m(k)/d_m(k_{max}), \qquad k_{max} = \underset{k}{\operatorname{argmax}} \quad |d_m(k)| \tag{5.3}$$

Then the sum of all D coefficients at any frame k are computed:

$$\hat{c}(k) = \sum_{m=1}^{D} \hat{c}_m(k) \tag{5.4}$$

and finally a renormalization step:

$$c(k) = \hat{c}(k)/\hat{c}(k_{max}), \qquad k_{max} = \underset{k}{\operatorname{argmax}} \quad |\hat{c}(k)| \tag{5.5}$$

$c(k)$ is the DCF whose peaks can be taken as phoneme boumdaries. DCF was used along with the SVF of Eq. (5.1) in order to constrain the transitions between phonemes in an HMM phone recognizer.

The Kullback-Leibler (KL) divergence can be used to define a model-free spectral distance measure. In probability theory and information theory, the KL divergence is a non-symmetric measure of the difference between two probability distributions [62]. The following form of KL divergence is given in [35] to measure the discrepancy between the spectral properties of consecutive frames:

$$KL = \int_{-\pi}^{\pi} K\left(\frac{S_1(\omega)}{S_2(\omega)}\right) d\omega \tag{5.6}$$

where $K(x) := x - \log x - 1$ and S_1 and S_2 denote the spectral densities for two adjacent frames. The KL divergence measure may also be applied to the distribution of the spectra or of features. When we consider this distribution to be Gaussian, the KL divergence is easily calculated based on the means and standard deviations of the respective distributions.

Another model-free method is the one introduced in [103] which operates exclusively on speech spectrum. First, the formant levels are equalized by the use of a pre-emphasis filter. Then, FFT analysis is performed on very short windows with a small window shift. The windows are chosen to have small length so as to capture the location of the main vocal tract excitation for voiced sounds. A hyperbolic tangent function is applied to the FFT coefficients to simulate the non-linear sensitivity of human hearing. Next, cross-correlation FFT matrix of the whole signal is computed, with the diagonal being the linear time axis. A special 2D filter composed of a square region and two triangular regions is slid along the diagonal. The result is a frame-by-frame spectral distortion curve. Since this curve may be noisy, it is passed through a minimax filter followed by peak masking. Then a local signal energy criterion is used to modulate the selection of the most significant peaks, which are chosen as the final detected transitions.

In [35] DCF was used as a reference segmentation method. [35] mentions several methods and implements a few as reference, among them SVF, DCF, and KL. The authors introduce their own segmentation algorithm which is compatible with any multidimensional feature encoding scheme. In their work, they consider three encoding schemes: MelBank features, MFCCs, and Log-area Ratios. The latter is based on the area ratio function of partial correlation coefficients drawn from the Linear Predictive model of the signal. The algorithm of [35] proceeds in several stages: consider the encoded speech feature x and x_i the i-th component of feature vector.

First, a change function J_i^a is computed from x_i, using change-assessment window of length a. Peaks of this function are detected according to a relative threshold b. Then these peaks are stored in a matrix $S(i, n)$ where n is the feature frame number. Last, a fitting procedure combines groups of near-simultaneous transitions from the multiple coefficient series into unique indications of phoneme location.

In [33], MFCCs are used as spectral features. The speech signals are first transformed into spectral frames (100 frames per seconds computed over 32 ms Hamming windows) and then transformed into a set of 10 MFCC coefficients (excluding the zero order coefficient that represent the total energy). The total energy coefficient was not used because this analysis focuses on the spectral features alone. Also the dynamic MFCC coefficients were not directly used in this study since the criterion they use for segmentation represents a dynamic measure by itself. This criterion is based on a measure of the spectral rate of change in time, which usually displays peaks at the transition between phones. It can be interpreted as the magnitude of the spectral rate of change. This spectral transition measure (STM), at frame m, can be computed as a mean-squared value:

$$STM(m) = \frac{1}{D}\sum_{i=1}^{D} a_i^2(m) \tag{5.7}$$

where D is the dimension of the spectral feature vector and $a_i(m)$ is the regression coefficient or the rate of change of the spectral feature defined as:

$$a_i(m) = \frac{\sum_{n=-I}^{I} n \times MFCC_i(n+m)}{\sum_{n=-I}^{I} n^2} \tag{5.8}$$

where n represents the frame index and I represents the number of frames (on both sides of the current frame) used to compute these regression coefficients. The authors have used $I = 2$ with a 10ms frame step. This corresponds to a total interval of 40 ms around the current frame at which the STM value is being computed. finally the phone boundaries are detected in two steps: first, all the peaks in Eq. (5.7)are marked as possible phone boundaries. Then the boundaries corresponding to the peaks that are not higher than the adjacent STM values by at least 1% of the highest peak in each sentence are removed. The 1% threshold was determined experimentally. A second criterion for removal of spurious boundaries is to compare the STM peak values with those of the adjacent valleys on both sides. The valleys usually occur at much larger distance than the adjacent frames used in the first part of the post-processing. If the difference between the value of each peak and the values of

its adjacent valleys (on both sides) is not larger than 10% of the peak value, then that peak corresponds to a flat STM region and it is removed from the boundary list. Each of the automatically detected phone boundaries are placed in time at a frame position (multiple of the frame step). Therefore, the reference boundaries are also mapped to this frame position to assess the quality of the method.

As can be seen, most of the above methods are based on heuristic definitions of some measures of change and then, to detect the points where the change is maximized they are highly dependent on tuning of several parameters. The latter may cause the main objective of a TI method to be contradicted: extensive tuning of parameters (sometimes on the whole dataset used for the final test) causes the final algorithm to be over-trained to that database.

5.2 Temporal evolution of singularity exponents

We start the development of our MMF-based TI phonetic segmentation, by mentioning a note on the temporal evolution of the Singularity Exponents (SE). As for the estimation of SEs, we use the multi-scale functional of Eq. (3.10) along with the estimation method described in 3.3.2.3 so as to provides a good balance between geometrical resolution and computational complexity. Indeed, computational efficiency is an important design constraint for phonetic segmentation task as it deals with very large databases of speech signal.

Once the exponents $h(t)$ are computed at each time instant t, we can proceed to study the useful information conveyed by these exponents about the variable temporal dynamics of speech signals. Indeed, different phonemes are basically different signals with different frequency content and statistical properties. Hence it is logical to expect the corresponding SEs to have different behavior inside the boundaries of each phoneme. This can be evidenced in different aspects of SEs. These changes can be demonstrated by the study of changes in distribution of SEs conditioned on time: $\rho(h|t)$. A graphical representation of the latter is shown in Fig. 5.1. In Fig. 5.1–top the original speech signal is shown along with the reference phoneme boundaries extracted from the manual transcription of TIMIT database. Fig. 5.1–middle displays the time evolution of the conditional distribution of singularity exponents. In the vertical axis we show the rank of the singularity exponents in bins of 5 percentiles. Then, at each time instance t, we take a 32ms window centered around t and we accumulate the exponents to the globally computed bins. As we want to represent conditional probability each row is norm-∞ normalized.

It is remarkable in Fig. 5.1 that between adjacent phonemes, there is a change in the position of maxima and in the variabilities of h distribution. Moreover, the distribution alternates from uni-modal to multi-modal, with uni-modal cases centered

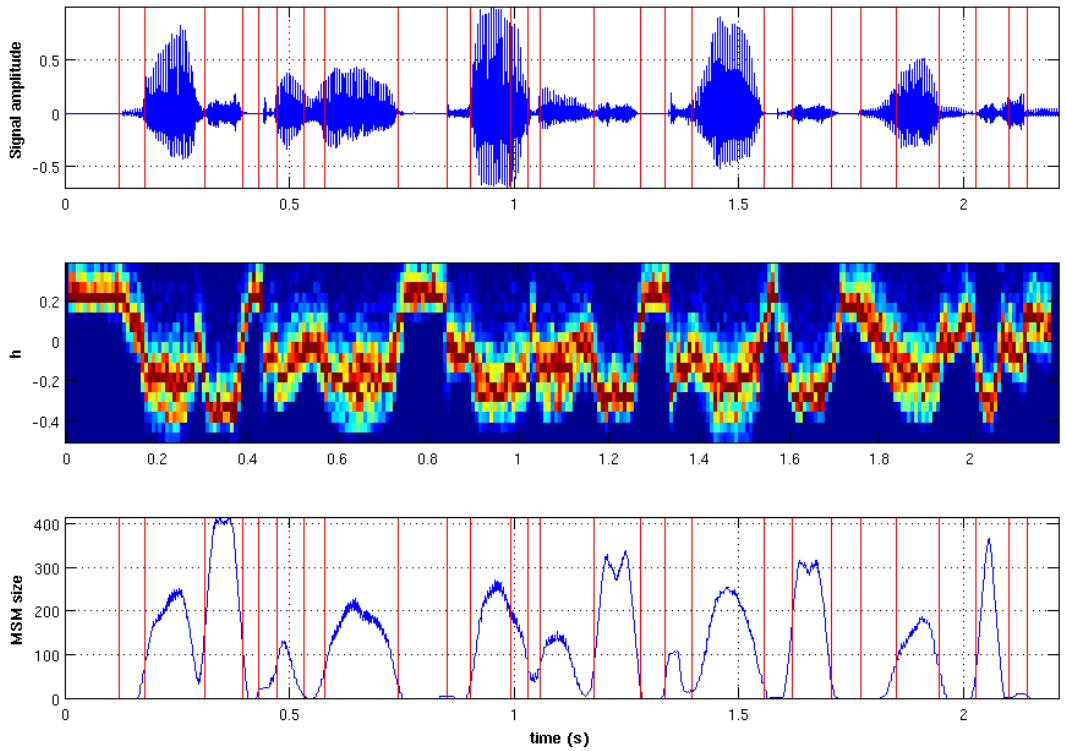


Figure 5.1: **top:** A part of a speech signal from TIMIT database. Manually-positioned phoneme boundaries are marked with vertical red lines. **middle:** The Joint histogram of the distribution of singularity exponents (vertical axis) conditioned to the time window (horizontal axis). Red corresponds to maximum probability and dark blue corresponds to zero probability. The horizontal axis is divided in 32ms bins. The vertical axis is divided in global 5-percentile bins, so that it is proportional to the global rank of the singularity exponents, not to their value. This avoids low-probability distortions. **bottom:** The number of points belonging to the MSM in a sliding window of 32ms length.

at the middle of the global range and multi-modal cases typically with two modes: one at the lowest extreme and one at the middle of the range.

The inter phoneme variability of exponents can be further demonstrated, by the study of the distribution (over time) of the MSM. To do so, we first form the MSM by taking 10% of the points having the lowest SE values. We then count the number of points belonging to the MSM for any given time instance t, in a window of length 32ms around this point. The result is shown in 5.1–bottom. It can be seen that the counter value is almost constant inside each phoneme and this constant value is different for adjacent phonemes (if we neglect the transitional regions on phoneme boundaries where the sliding window partially covers samples belonging to both of the neighboring phonemes).

These observations suggests that SE indeed convey valuable information about local dynamics of the speech which can be even used for phoneme-level classification. However, the fact that there is considerable change in the distribution of SE between neighboring phonemes can be readily used for the phonetic segmentation task.

5.3 MMF based phonetic segmentation

To develop an automatic segmentation method which uses the properties of SEs mentioned in section 5.2, we employ a two step procedure:

1. First, we define a quantity representative of the distinctive properties (section 5.3.1) and we develop a simple and efficient method for detecting changes on this quantity (section 5.3.2).

2. We apply the above method to the signal itself and its low-passed version, to make a preliminary list of phoneme boundary candidates. These candidates are then refined by dynamic windowing and Log-Likelihood Ratio Test (LLRT) to make the final decision (section 5.3.3).

This 2-step approach is similar to that of traditional segmentation methods where there is a boundary pre-selection followed by statistical hypothesis tests to make the final decision [1].

5.3.1 The Accumulative function ACC

To define a simple measure which is a quantitative representative of the changes in distribution of SE between neighboring phonemes, we use the simplest descriptive statistic, which is the average of SE. We consider $h(t)$ as a random variable whose average is changing between adjacent phonemes and we search for the locations of changes in local averages of $h(t)$ as the candidate phoneme boundaries. However, since the SE estimations are available at the finest resolution, and this resolution is useful to be preserved for phonetic segmentation task, we would want to avoid any windowing for the estimation of averages. To do so, we use the primitive of $h(t)$ as the representative quantity. Indeed, inside the boundaries of each phoneme, the slope of this quantity would be an estimate for the local average of the $h(t)$. Formally, the definition reads:

$$ACC(t) = \int_{t_0}^{t} d\tau \, h(\tau) \tag{5.9}$$

The resulting functional is plotted in Fig. 5.2. This time-variable functional is detrended to enhance the presentation of the values. Just as we expected, this new functional reveals the changes in distribution in a more precise way. Indeed, inside each phoneme the functional ACC is almost linear (if we neglect the small scale fluctuations). Moreover, there is a clear change in the slope at the phoneme boundaries. These slope changes are even able to identify the boundaries between extremely short phonemes, such as stops. Extensive observations over different sentences confirms this behavior and thus the strength of the proposed functional in Eq. (5.9). The next step is to develop an automatic method of detection of changes in the slope of the ACC. A very simple solution is to fit a a piecewise linear curve to the ACC and take the break-points as the candidates of change in distribution of SE, i.e. phoneme boundary candidates.

5.3.2 Piece-wise linear approximation of ACC

Assuming that for a speech signal of length N, ACC has K significant break-points, the problem of finding these breakpoints can be formalized as the following optimization problem (note that K is unknown):

- find $LA_m(ACC, 1, N)$ such that m is minimized and $E_m^{1\to N} < \epsilon$.

where $LA_m(ACC, 1, N)$ denotes m-piece linear approximation of the curve ACC between the time indices 1 and N and $E_m^{1\to N}$ is the corresponding mean squared approximation error. This optimization problem, is addressed in [72] for denoising of piece-wise constant functions and it is argued that when $K << N$, a greedy search for jump placements (called the jump penalization method) may give rise to a more efficient and more accurate approximation compared to other solvers. However, the computational complexity of this method is still increasing with K which is not desirable for the task of automatic phonetic segmentation since it deals with large databases of the speech.

We develop a similar but more efficient solver for the above optimization problem which its computational is independent of K. Our optimization method is motivated by the dynamic programming approach for Piece-wise Linear Approximation (PLA) problem described in [58]. There, the problem is solved by recursive solution of some easier sub-problems of the type $LA_2(ACC, 1, N)$, which also suffers from the problem of increasing computational complexity with K (it requires $\log_2 K$ passes of N samples).

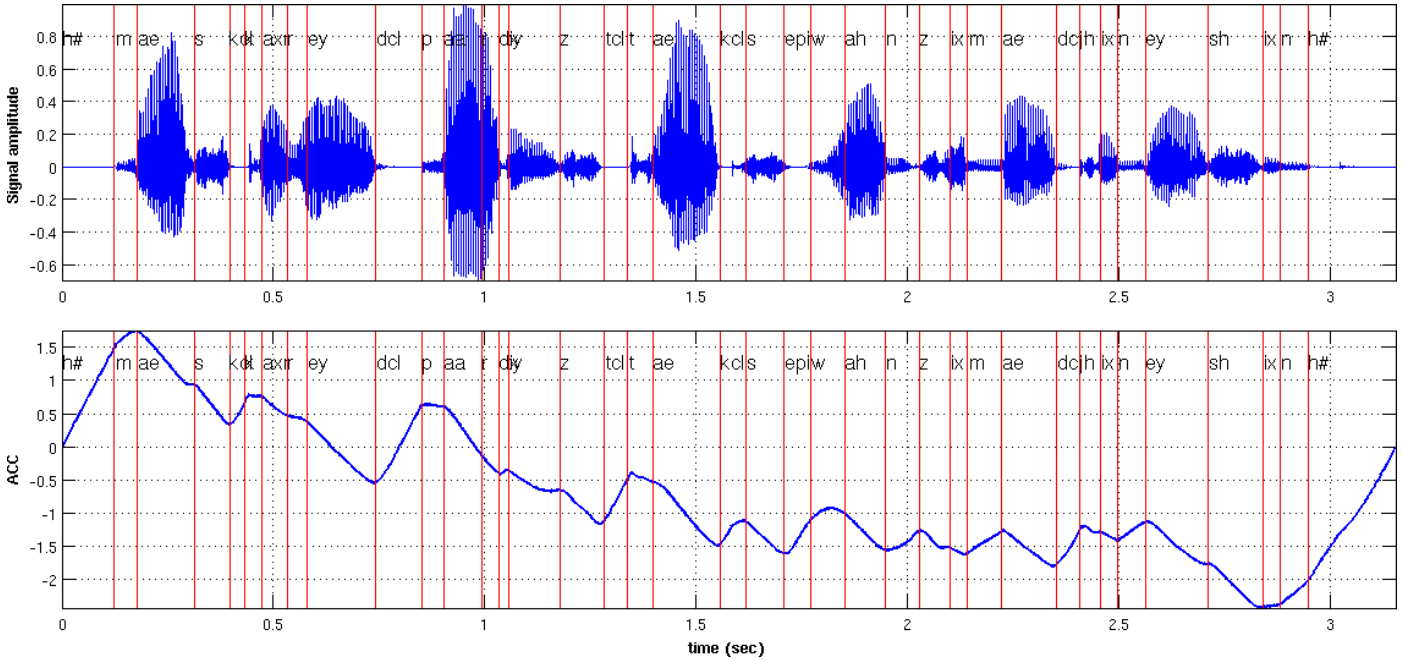


Figure 5.2: **top:** The speech signal "Masquerade parties tax one's imagination" from the TIMIT database and **bottom:** the *ACC* functional of Eq. (5.9). The functional is detrended by subtracting its global mean.

Our efficient implementation includes the replacement of m-point approximation problem $LA_m(ACC, 1, N)$, by sequential solutions of 1-point approximations $LA_1(ACC, t_i, t_{i+1})$, where $i \in (1 \cdots m-1)$ and $t_i < t_{i+1}$. The algorithm starts by $t_1 = 1$, and at each iteration t_{i+1} is determined by the procedure explained in Alg. 1. There, c_i stands for the i-th phoneme boundary candidate. The particularity of the method is that its computational complexity is independent of the number of breakpoints (K) and requires only 2 passes of N data samples. The detected candidates (c_i) are rejected if located in the silence, i.e. where the average power in a 32ms window is less than -30dB.

Algorithm 1 PLA procedure

1: $i \leftarrow 1; t_i \leftarrow 1;$
2: **for** $k_1 \in (1 \cdots N)$ **do**
3: **if** $E_1^{t_i \to k_1} > \epsilon$ **then**
4: **for** $k_2 \in (t_i + 1 \cdots k_1 - 1)$ **do**
5: $E_2^{t_i \to k_1}[k_2] \leftarrow E_1^{t_i+1 \to k_2} + E_1^{k_2+1 \to k_1};$
6: **end for**
7: $t_{i+1} = \underset{k_2}{\operatorname{argmin}}\, E_2^{t_i \to k_1}[k_2];$
8: $c_i \leftarrow t_i;$
9: $i \leftarrow i + 1;$
10: **end if**
11: **end for**
12: ▷ $E_m^{t_i \to t_j}$ is the m-piece PLA error from t_i to t_j.

It is important to note that at each iteration, t_{i+1} is determined through a greedy *minimization* (Alg. 1 line 7) and not through a thresholding operation. This greedy minimization, decreases the impact of the threshold ϵ (Alg. 1 line 3). We will later see in experimental results that the algorithm is not really that sensitive to the selection of the threshold ϵ. The insensitivity to the selection of algorithm parameters is an important property for text-independent phonetic segmentation, where no data is available for training of the parameters.

Alg. 1 is very simple for implementation and is quite fast in practice (as it requires only 2 passes over N data samples). This algorithm is readily a phonetic segmentation algorithm which provides very good performance in detection of phoneme boundaries. This is shown by extensive experiments in section 5.4 (this simple algorithm is denoted by SE-ACC). However, in the following subsection we develop a more accurate segmentation algorithm by using this simple criterion of change inside a classical decision making procedures.

5.3.3 The two-step segmentation algorithm

The piece-wise linear approximation of ACC in section 5.3.2 is already a working phonetic segmentation algorithm. However, by performing detailed analysis about the functionality of such algorithm we found that it can not fully exploit the strength of SE for phone boundary detection. This was motivated by some observations about the behavior of this simple algorithm (error analysis) at some particular phoneme transitions. This led us to propose a two-steps algorithm where we first pre-select candidate boundaries and then use statistical hypothesis test to make the final decision.

By analyzing the performance of the curve fitting method of section 5.3.2, we made two observations. First, we observed that some of the missed boundaries correspond to transitions between fricatives/stops and vowels. We also observed that transitions between speech and low energy segments (such as pauses and epenthetic silence) display strong and easy-to-detect changes in the slopes of ACC. Indeed, as shown in Fig. 5.1–middle, the SEs of low energy segments have high positive values, while they are mostly negative in active speech segments. These observations and the fact that fricatives/stops are essentially high-band signals, motivated us to compute ACC on a low-pass filtered version of the utterance. By doing so fricatives/stops-vowels transitions will be converted into silence-speech transitions which are easier to detect as shown in Fig. 5.3. As for the choice of the cut-off frequency of the low-pass filter, we opted for $f_c = 1800$Hz, as we observed that usually most of the spectral energy of fricatives is located above 2000Hz and, for most stops, the active frequency bands start at 1800Hz.

The second and the more important observation we made is related to the local distribution of SEs. We observed that some missed boundaries correspond to neighboring phonemes which do have a quite distinctive difference in their SE distribution. However, the change in their averages is not strong enough to be translated as a change in the slope of ACC, and thus it is not captured by the simple curve-fitting procedure. It is then natural to think about including a statistical hypothesis test over SE distributions in our segmentation algorithm in order to detect such boundaries. Motivated by these observations, we develop new segmentation algorithm which consists in 2 steps.

1. We use the curve-fitting method in section 5.3.2 to detect the boundaries of the original and filtered signal from their ACC . We gather all the detected boundaries and consider them as candidate boundaries.

2. We make the final decision by performing a dynamic windowing over all the candidates followed by Log Likelihood Ratio Test (LLRT) over SE distributions of the *original* signal.

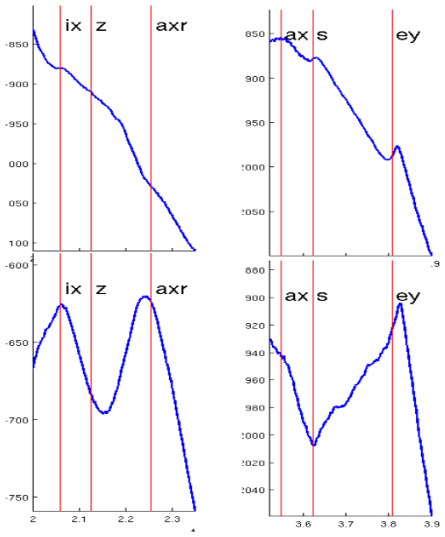


Figure 5.3: **top:** Two examples of ACC for the original signal and **bottom:** The ACC for the low-passed filtered signal. Phoneme boundaries are marked with vertical red lines.

We use a Gaussian hypothesis test in the second step because our purpose is to detect changes in mean and variance. More precisely, for each candidate c_i we consider the large window $Z = [c_{i-1}, c_{i+1}]$ and the two smaller windows $X = [c_{i-1}, c_i]$ and $Y = [c_i, c_{i+1}]$. We then compute LLR statistic to decide between the two hypothesis:

- H_0 : SE of Z are generated by a single Gaussian.
- H_1 : SE of Z are generated by two Gaussians on X and Y.

If H_1 is significantly likelier than H_0 (with an empirical threshold), we select c_i as a boundary. Otherwise, c_i is removed from the candidates list. We emphasize here that SE of the filtered signal are used only in the first step. The final decision is exclusively made upon the information conveyed by SE of the *original* signal. We also emphasize that this new algorithm is still simple and efficient as the original one.

5.4 Experimental results

While presenting the experimental results, we also address a common difficulty in comparing text independent phonetic segmentation methods which is the diversity of evaluation datasets and also incoherencies in performance measures. Indeed, while the literature is quite rich in phonetic segmentation methods, it is relatively poor in material for performance comparison. Most of the papers report either

on undefined subsets of known databases or on personal/unaccessible databases. Moreover, it is often difficult to analyze the reported accuracy scores because of the diversity of measures used. This makes it difficult to make fair comparisons between different TI segmentation algorithms.

We will thus provide an extensive set of comparative results which are easy to interpret and to compare with. We carry out a detailed evaluation on known databases using every possible performance measure and for several degrees of error tolerances.

5.4.1 Experimental setup

Our evaluation is carried out on the *full* Train and Test sets of the TIMIT database which respectively contain 4620 and 1200 sentences uttered by 462 and 120 speakers. We develop our algorithm on a small dataset of 30 sentences extracted from the Train set. Using different sizes of tolerance windows, we provide comparison of segmentation results for 3 methods. In the first one, we give the results reported in [33] while considering the 5ms shift in their comparisons. In the second one, we provide the results using our original SE-based algorithm which is summarized in section 5.3.2 and we call it SE-ACC. In the third one, we present the results obtained using our new algorithm described in section 5.3.3 and we call it SE-LLRT.

We first test the performance of our two algorithms on the whole Train set and we compare the results with the stat-of-the art methods. Then, to compare with [35], and to show that the promising result on Train dataset is not the result of over-trained parameters, we evaluate it on the Test dataset which is balanced for the phonological coverage. Note that we have provided a summary of the algorithms used in [35] and [33] in section 5.1.1.

5.4.2 Performance measures

The basic data required for performane evaluation of a phonetic recognizer includes:

- Total number of detected boundaries (N_T).
- Total number of correctly detected boundaries (N_H).
- Total number of boundaries in the reference transcription (N_R).

Consequently, the segmentation quality can be evaluated and analyzed using following three "partial" scores:

- Hit Rate (HR): the percentage of correctly detected boundaries.

- False Alarm Rate (FA): the percentage of erroneously detected boundaries (with respect to N_T).
- Over Segmentation Rate (OS): the percentage of erroneously detected boundaries (with respect to N_R).

These quantities are defined as:

$$HR = \frac{N_H}{N_R}, \quad OS = \frac{N_T - N_R}{N_R}, \quad FA = \frac{N_T - N_H}{N_T} \tag{5.10}$$

In the ideal case, one desires to achieve HR=100% and zero OS and FA. However, these quantities alone may be misleading, in that higher HR might be achieved with the cost of higher OS or FA. It is argued in [102] that increases in detection detection rates (HR) are often due to increased OS levels and not to algorithmic improvements. In other words, a better HR can be achieved by simply adding random boundaries. Hence, the overall quality of a segmentation method must be evaluated using a global measure which simultaneously takes these scores in to account. One of the well known global measures is the F_1-value:

$$F_1 = \frac{2 \times PCR \times HR}{PCR + HR} \tag{5.11}$$

where $PCR = 1 - FA$ is the *precession rate*. Another global measure, called the R-value, which is supposed to be more accurate than F_1 has been recently proposed in [102]. This measure makes more emphasize on OS by measuring the geometric distance in the performance-plane (HR versus OS plane) from the ideal point of segmentation where HR=100% and OS=0%. A point at which the R-value is equal to 1:

$$R = 1 - \frac{|r_1| + |r_2|}{2}$$
$$r_1 = \sqrt{(1 - HR)^2 + OS^2}, \; r_2 = \frac{HR - OS - 1}{\sqrt{2}} \tag{5.12}$$

5.4.3 Results : Train dataset

Table 5.1 presents HR, FA and OS for the 3 methods. The first observation is that SE-LLRT outperforms SE-ACC for the 3 scores and all tolerance windows. In particular,

a significant improvement is made in FA and OS. This shows that, as expected, some of the insertions introduced by the curve fitting procedure has been corrected by the LLRT. The second observation is that both SE-ACC and SE-LLRT yield considerably much higher accuracy than [33]. In particular, the smaller tolerance window is, the higher relative improvement is. This shows that SE-LLRT is better suited for high precision detection of phoneme boundaries.

To this regard, we can mention another interesting comparison with [70] which is also a sample-based segmentation method as ours. In [70], a total HR of 43.5% is reported noting that 86.8% of detections are located within the first bin of the cumulative histogram of distances from true boundaries. This corresponds to HR=43.5% $\times$ 86.8%=37.75% for 7.5ms tolerance. With SE-LLRT, we obtain HR=44% for the same tolerance which shows that our algorithm is significantly more accurate. More importantly, the algorithm in [70] is supervised (all manual transcriptions are used to train a neural network) while ours is fully unsupervised.

Table 5.1: The comparative table of segmentation results.

tolerance	score	Dusan [33] TIMIT-Train	SE-ACC TIMIT-Train	SE-LLRT TIMIT-Train	SE-LLRT TIMIT-Test
5ms	HR	22.8%	31.7%	32.3%	32.4%
	FA	79.7%	70.2%	68.5%	62.02%
	OS	12.8%	6.4%	2.5%	4.61%
10ms	HR	-	52.8%	53.6%	53.16%
	FA	-	50.4%	47.7%	49.18%
	OS	-	6.4%	2.5%	4.61%
15ms	HR	59.2%	65.5%	66.3%	65.39%
	FA	47.5%	38.4%	35.4%	37.50%
	OS	12.8%	6.4%	2.5%	4.61%
20ms	HR	-	72.4%	72.5%	71.7%
	FA	-	31.94%	29.2%	31.5%
	OS	-	6.42%	2.5%	4.61%
25ms	HR	75.3%	76.2%	76.1%	75.17%
	FA	33.2%	28.3%	25.8%	28.14%
	OS	12.8%	6.4%	2.5%	4.61%
30ms	HR	-	78.8%	80.5%	77.41%
	FA	-	26%	23.7%	26.00%
	OS	-	6.4%	2.5%	4.61%

Table 5.2 presents the performance of each of the 3 methods when evaluated using the global measures F_1 and R. The same observations we made above, still hold for the global performance evaluation. Indeed, SE-LLRT still outperforms SE-ACC for both F_1 and R. Moreover, about 6% (resp. 10%) improvement in R-value and 4% (resp. 10%) in F_1-value is achieved for 25ms of tolerance (resp. 5ms and 15ms). This is a significant gain in accuracy that shows the strength of SEs in revealing the transitions fronts between phonemes.

Table 5.2: The comparative table of global performance measures.

tolerance	score	Dusan [33] TIMIT-Train	SE-ACC TIMIT-Train	SE-LLRT TIMIT-Train	SE-LLRT TIMIT-Test
5ms	R-value	0.29	0.39	0.41	0.41
	F_1-value	0.21	0.31	0.32	0.32
10ms	R-value	-	0.57	0.60	0.58
	F_1-value	-	0.51	0.53	0.52
15ms	R-value	0.60	0.68	0.70	0.69
	F_1-value	0.55	0.63	0.65	0.64
20ms	R-value	-	0.74	0.76	0.74
	F_1-value	-	0.70	0.72	0.70
25ms	R-value	0.73	0.77	0.79	0.77
	F_1-value	0.71	0.74	0.75	0.735
30ms	R-value	-	0.79	0.81	0.79
	F_1-value	-	0.76	0.77	0.76

5.4.4 Results : Test dataset

To the best of our knowledge, for the TEST dataset, no TI method in the literature has reported the detailed performance results. However, to facilitate later comparisons, we have provided the full segmentation results for this dataset, in the last columns of tables 5.1 and 5.2. We compare however to [35] whose results for the Test dataset is reported only for 20ms tolerance (table 5.3). There, extensive experiments are made to select the best encoding scheme and to adjust the three parameters of their algorithm, so as to achieve the best performance possible compared to many traditional methods. It can be seen that the HR of [35] (8-Mel-bank coding scheme) is about 10 % higher than that of ours, while its OS is about 13 % worse. Hence the comparison in terms of HR and OS is a bit difficult. However, in terms of R-value the results are the same.

Table 5.3: Segmentation performance of [35] for 3 different coding schemes.

Coding scheme	(a,b,c)	HR	OS	FA	R-value	F-value
8-Mel-bank	(2,optimal value,5)	82%	18.32%	30.69%	0.743	0.75
5-MFCC		76%	12.31%	31.33%	0.736	0.72
Log Area Ratio		70%	6.31%	34.16%	0.72	0.68

It should be mentioned that the results of [35] is attained by performing an extensive parameter tuning procedure over the whole Test dataset. Quite the contrary, in our case we have applied our algorithm to the Test dataset without any parameter tuning. Our intention was to ensure that the promising performance of the algorithm is not the result of over-training of parameters. Even more, we emphasize an important feature of our algorithm which is its insensitivity to the threshold of the linear curve fitting. Fig. 5.4 displays the R-value for different thresholds. We have used a subset of 30 randomly selected sentences from the Train folder of TIMIT database, to compute these values. One can see that with about 400% change in the value of threshold, that variance of changes in R-value is less than 0.5%. Thus, we can fairly consider that our algorithm is not very sensitive to parameter adjustments. This is a major advantage as most of the TI methods require accurate threshold tuning.

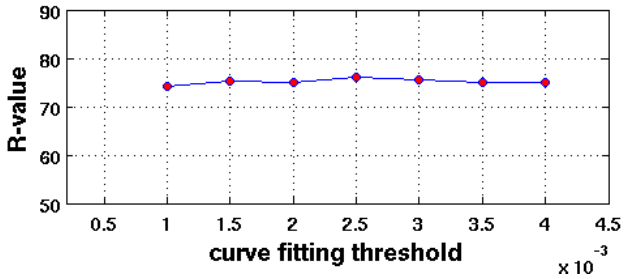


Figure 5.4: The sensitivity of the SE-LLRT to threshold.

5.5 Conclusion

In this chapter, we presented our observation regarding the informativeness of singularity exponents for speech analysis. We exploited the time evolution of local distributions of SEs to analyze local properties of speech signals, in the context of phoneme boundary detection. The interesting results we achieved by a simple anal-

ysis of the singularity exponents in the first run, indicates the very high potential of the formalism. This motivates us to go further in this research, to discover the significance of one other major component of the MMF w.r.t to the speech signal: the MSM which gives access to the subset of least predictable points of the complex signal. This topic will be addressed in remaining chapters of this thesis.

6

Compact representation of speech using the MSM

The encouraging segmentation results we achieved in chapter 5 by a very simple analysis of the Singularity Exponents (SE), motivates us to study the relevance of the other major component of the MMF for non-linear analysis of the speech signal: the Most Singular Manifold (MSM). Indeed, as mentioned in chapter 3 (section 3.2.1), the availability of precise estimates of SEs unlocks the determination a collection of points inside the complex signal which are considered as the least predictable points (the MSM). This leads to the associated compact representation and reconstruction. This chapter presents the very first steps in establishing the links between the MSM and the speech signal. To do so, we make slight modifications to the formalism so as to adapt it to the particularities of the speech signal. Indeed, the complex intertwining of different dynamics in speech (added to purely turbulent descriptions) suggests the definition of appropriate multi-scale functionals that might influence the evaluation of SEs, hence resulting in a more parsimonious MSM. We present a study that comforts these observations: we show that an alternative multi-scale functional does lead to a more parsimonious MSM from which the whole speech signal can be reconstructed with good perceptual quality.

As MSM is composed of a collection of irregularly spaced samples, we use a classical method for the interpolation of irregularly spaced samples, called the Sauer-Allebach algorithm [105], to reconstruct the speech signal from its MSM. We show that by using this generic algorithm [and even by slight violation of its conditions] high quality speech reconstruction can still be achieved from a MSM of low cardinality. This shows that the MSM formed using the new multi-scale functional we define in this chapter, indeed can give access to a subset of potentially interesting points in the domain of speech signal. Finally, in order to show the potential of this parsimonious representation in practical speech processing applications, we quantize and encode the MSM so as to develop a waveform coder which achieves slightly better compression performance compared to the standard G.726 ADPCM [44].

This chapter is organized as follows: section 6.1 provides a brief introduction to the concept of MSM in the framework of reconstructible systems. In section 6.2 we

introduce a new multi-scale functional for the computation of SEs which is more adapted to the particularities of the speech signal and we also present the generic Sauer-Allebach algorithm for signal reconstruction from any subset of irregularly spaced samples (the MSM in our case). In section 6.3, the experimental results are presented, while describing in detail the coder that we have developed using the MSM. Finally we draw our conclusions in section 6.4.

6.1 The MSM in framework of reconsructible systems

As we explained in previous chapters, our strategy in non-linear speech analysis is to relate the non-linear character of speech signal to the concept of predictability [17] but in a *local* manner. To do so, we use a microcanonical approach which relies on the precise evaluation of geometric parameters called Singularity Exponents (SEs) whose values are naturally associated to criticality. Critical transitions in the system are described by the values of SEs, the latter being computed in the speech signal domain itself. This property can be used for geometric localization of the subset of the least predictable points inside the complex signal (the MSM): for a given point, the smaller the value of SE is, the higher predictability is in the neighborhood of this point [127] and hence, this point itself can be considered as the least predictable point. Formal definition of the MSM is given by Eq. (3.5), where it is denoted by $\mathcal{F}_\infty$.

The fact that the MSM maximizes the statistical information in the signal is demonstrated in the framework of reconstructible systems: It has been shown that for many real world signals (natural images in particular), the whole signal can be reconstructed using only the information carried by the MSM [127, 131]. For example, in [127] a formal kernel is presented for accurate reconstruction of natural images from partial information about their gradient (the gradient information restricted to the MSM).

The 2-D reconstruction kernel is motivated in [127] by the well known scaling (with frequency) property of power spectrum in natural images [104]. It acts as an inverse derivative operator in Fourier space which serves for diffusing partial gradient information on the MSM over the whole 2-D space so as to reconstruct the whole image [131]. Having this reconstruction formula in mind, in [126] a new multi-scale functional is defined for the estimation of SEs. This new measure is thus associated to the local degree of predictability at each point: the measure is based on a local evaluation of the reconstruction kernel and can be regarded as a wavelet projection of the gradient which penalizes unpredictability. The resulting MSM is shown to provide higher reconstruction qualities.

In case of the one dimensional speech signal, the direct 1-D adaptation of the aforementioned reconstruction formula reduces to step function interpolation. Also,

the measure defined in [126] simplifies into finite differences for 1-D signals. Consequently, in following sections we first define a multi-scale functional which takes into account the particularities of the speech signal, such as inter-sample correlations.We then introduce and use a generic reconstruction formula to make the reconstruction from the MSM samples of the speech signal. We show that this new functional provide the highest reconstruction quality. We will see that it is indeed possible to use this functional to form a compact MSM from which the whole speech signal is reconstructible with a reasonable perceptual quality.

6.2 SEs and inter-sample dependencies

In a complex system whose dynamics is entirely described by an underlying multi-scale hierarchy, for instance Fully Developed Turbulence (FDT), SEs can be computed indifferently by various multi-scale measures (speed increments, energy dissipation etc.) all giving rise to the same MSMs and reconstruction properties [130]. We made an introduction to some of these measures in chapter 3.

The voice production mechanism however tends to indicate that speech involves the superposition of different dynamics on top of a purely chaotic one, which may mask some of the interesting dynamics. It is known that the speech signal, sampled at the Nyquist rate or faster, exhibits significant correlations between successive samples [24]. Indeed, such inter-sample dependencies imply a certain degree of predictability of each sample from its direct neighboring ones. This predictability, certainly introduces some redundancy to any form of speech representation which does not take these correlations into account. Instead, one can remove the predictable portion of any given sample (from its previous samples) and concentrate on the true information contribution of each sample. This strategy, for instance is central to the Differential Pulse Code Modulation (DPCM) which encodes the difference between samples and their predictions, and thus leads to superior compression performance compared to the simple PCM technique [24].

6.2.1 A new multi-scale functional for the estimation of SEs

Our goal here is to define a multi-scale functional $\Gamma_r(s(t))$ for the specific case of the speech signal, which takes into account the above-mentioned inter-sample dependencies. The simplest solution to remove these redundancies is to work on the differences between successive samples (as in the basic form of DPCM [99]). But the inter-sample correlations are also scale-dependent (within specific limits imposed by the system) because of the wide variation of scales in which they appear. Clearly any single-scale measure will fail in providing a balanced evaluation of predictability about the smoother parts of the speech signal (like voiced parts) and the

noise-like behavior of other parts (like unvoiced speech). Thus, we must provide a functional for cross-scales inter-samples dependencies.

Taking these considerations into account, we define the following multi-scale measurement of variations, operating at the scale r on the signal $s(t)$.

$$\mathcal{D}_\tau s(t) = |2s(t) - s(t-\tau) - s(t+\tau)| \tag{6.1}$$

At the finest scale, this can be seen as the application of the simplest functional introduced in chapter 3, i.e. linear increments, to differences of the signal instead of the signal itself (however, this definition has a multi-scale aspect to it which makes it completely different). Next, similar to Eq. (3.12), the variations are summed together to form the final multi-scale functional:

$$\Gamma_r(s(t)) = \int_0^r |\mathcal{D}_\tau s(t)| d\tau \tag{6.2}$$

For the final estimation of SEs, this $\Gamma_r(s(t))$ must be evaluated in accordance with the power-law of Eq. (3.3) over a set of scales. To do so, we use Eq. (3.24) on the set of four finest scales. The resulting SEs are then being used to form the MSM and evaluate the reconstructibility in the following.

6.2.2 Speech reconstruction from the MSM

We recall that the MSM is composed of a subset of [irregularly spaced] samples in the signal domain having the smallest values of SE. Thus, Shanon's interpolation principle can not be applied for the reconstruction of signal from the MSM. Instead, we make use of a classical method in the interpolation of irregularly spaced samples, called the Sauer-Allebach algorithm [105]. The latter guaranties perfect reconstruction of any band-limited discrete signal $s[k]$ of length N from irregularly spaced samples $s[k^{ir}], k^{ir} = 1 \cdots p < N$, with the condition that the maximal gap between samples must be less than the Nyquist rate. Assuming that signal is band-limited to the frequency of B_ω, the algorithm starts with low-pass filtering of a first approximate of the signal from its subset of irregularly spaced samples. This first approximation can be achieved by a simple interpolation scheme like linear interpolation:

$$s_0[k] = P_{B_\omega}(\mathcal{A}s[k^{ir}]) \tag{6.3}$$

where $\mathcal{A}$ stands for an approximation operator (linear interpolator for instance) and P_{B_ω} denotes the low-pass filtering operation with a filter whose cut-off frequency is equal to B_ω. Next, the following recursion is repeated until convergence:

$$\begin{aligned} e_n[k^{ir}] &= s[k^{ir}] - s_n[k^{ir}] \\ s_{n+1}[k] &= s_n[k] + \lambda P_{B_\omega}(\mathcal{A} e_n[k^{ir}]) \end{aligned} \tag{6.4}$$

It is proven that $s_{n+1}[k]$ converges to the original signal with a geometric convergence rate [39]. However, in case of reconstruction of speech signal from its MSM, there are two issues to overcome before direct application of the Sauer-Allebach algorithm.

The first issue is the result of the time-varying nature of speech's statistical properties which does not permit the determination of a unique B_ω to be used for the whole signal. In fact, the unvoiced parts may have effective frequency content up to 8kHz while for the voiced parts there is no considerable activity for frequencies higher than 4kHz. So the application of the above algorithm is not straightforward. We use a simple frame-based implementation of the interpolation algorithm to account for the time-varying nature of speech. In non-overlapping frames of 20ms length, we first make a simple voiced/unvoiced decision and then, for voiced frames we set $B_\omega = 4$kHz, while for unvoiced frames we put $B_\omega = 8$kHz to maintain the quality of unvoiced frames. As for the voiced/unvoiced decision, we use a very simple reasoning on local cardinality [1] of the MSM: we have observed that the density of MSM points is much more in unvoiced frames. Consequently, for each frame, if the average distance between successive MSM points is more than a specific value (depending on the sampling frequency), we consider the frame as a voiced frame and otherwise, it will be treated as an unvoiced one.

The second issue is that the structure of MSM generally violates the maximal gap condition: quite frequently the distance between two points in the MSM is more than the maximal gap allowed by the Nyquist limit. Of course, since we are violating the maximal gap condition of the Sauer-Allebach reconstruction algorithm, we can not expect perfect reconstruction of the original signal. However, as experimental results will show, a high quality of reconstruction is still achievable.

1 Recall that MSM is generally formed by comparing SE value at each point with a threshold. Consequently, considering the inherent non-stationarity of the speech signal, it is natural to expect different parts of speech to have different number of points below this threshold.

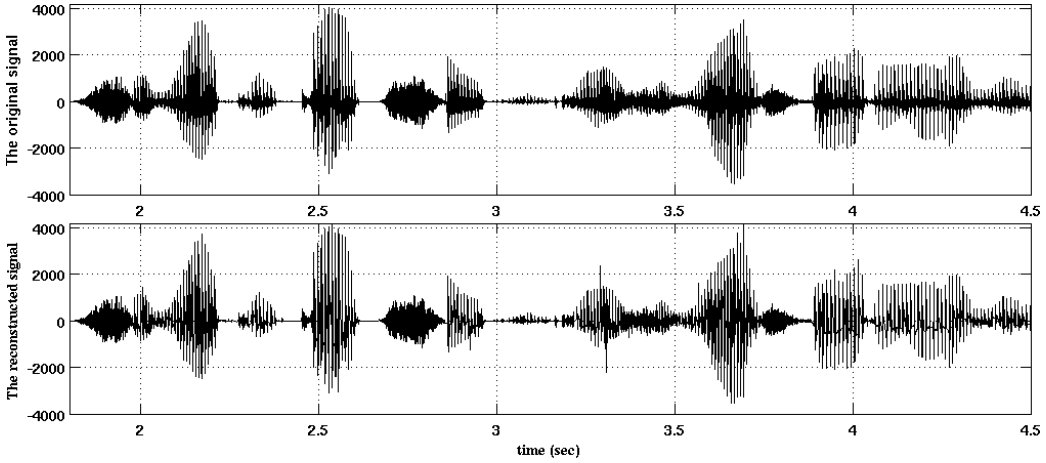


Figure 6.1: Waveforms of [top] the original signal and [bottom] the reconstructed signal from an MSM containing 14.7% of the points.

6.3 Experimental results

Our experiments are carried out on about 2.5 hours of speech signal, composed of 3000 utterances from speakers of different sexes, accentes and ages, which are randomly taken from the TIMIT database.

In our application we compute the $h_{r_i}(t)$ for the four finest scales ($k = 4$) and use them in Eq. (3.24) to compute the singularity exponent. Then, we can use these SEs to form the MSM as explained in section 3.2.1. Practically, one can sort all the samples of a given signal according to their value of SE and take as many points as necessary to achieve a desired level of reconstruction in terms of Mean Squared Error (MSE). However, considering the time-varying statistics of the speech signal, a global formation of MSM would be problematic. To account for such time-varying characteristics, we perform a *local* formation of the MSM in non-overlapping frames of length 20 ms. In each frame, we sort the samples according to their value of $h(t)$ and we take as many samples as necessary to achieve the desired level of MSE from the *local* reconstruction (only in the current frame). We then use the ensemble of these points to form the final MSM. Finally, we use the method explained in section 6.2.2 to reconstruct the whole signal from the MSM. An example of the reconstructed waveform is shown in Figure 6.1-bottom, in the case of an MSM containing 14.7% of the samples.

Figure 6.1-top shows a segment of speech signal from the dataset. The reconstructed waveform, from a MSM containing 14.7% of samples, is shown in Figure 6.1-bottom. It can be seen that the overall shape of the speech signal with its diverse dynamical regimes is preserved. There exist few occasional peaks (such as the one around $t = 3.25$ sec), which are duo to rare events of very large distances between consecutive MSM points (much more than Nyquist limit). However, the fre-

quency of occurrence of these events is too low, such that they have no significant effect on the perceived quality.

We compare the quality of reconstrction for three different sampling methods: the MSM formed using our proposed measure in Eq. (6.2), the MSM formed by the measure of [126] and the subset of equally spaced samples. The same reconstruction method is used for all these three sampling methods as proposed in section 6.2.2. We emphasize that the reconstruction we use in this paper is different from the simple interpolation we use in [64]. The most accurate method for evaluating speech quality is believed to be the subjective listening test [56]. Our informal subjective listening tests confirm that the perceptual quality of the reconstructed signal from an MSM containing 14.7% of the points is quite natural. Also, the quality is clearly better when compared to [126] (with the same size of MSM). In addition, to confirm that such quality can not be achieved by the application of Sauer-Allebach reconstruction algorithm to any subset of points, the results of the reconstruction from the MSM are compared with that of the subset of equally spaced samples (of the same size). In this case, our method gives much better quality.

In order to have an objective evaluation of the relative quality of different reconstructions, we use two objective measure of perceptual quality: the Perceptual Evaluation of Speech Quality (PESQ) as recommended by International Telecommunication Union (ITU) [44] and also a composite objective measure (denoted by Csig) of speech distortion [56]. The latter measure is a combination of several basic perceptual measures, and has shown high correlation with subjective listening test which is believed to be the most accurate method for evaluating speech quality. Both of these objective measures provide a number in the range of 1 (the worst quality) to 5 (the best quality) which is shown to have very high correlation with the average score that would be given by a panel of listeners about the quality of the processed speech signal.

Figure 6.2 shows a comparison of the resulting PESQ, CMOS and SNR versus the size of the subset of samples for all these sampling methods that clearly confirms our subjective informal evaluations. For instance, to achieve the PESQ equals to 3, the proposed method requires 16% of points in the MSM, while 27% of samples are required with the method of [126]. In case of the subset of equally spaced samples, such quality is not achieved with less than 33% of the samples.

6.3.1 A MSM-based speech coder

Now that we have seen the reconstructibility of the speech signal from the MSM, we show how it can be used to build an efficient waveform coder of speech. We emphasize however, that our goal is not to achieve the best coding performances, but rather to demonstrate the potential of the proposed methodology in providing a parsimonious representation using the MSM.

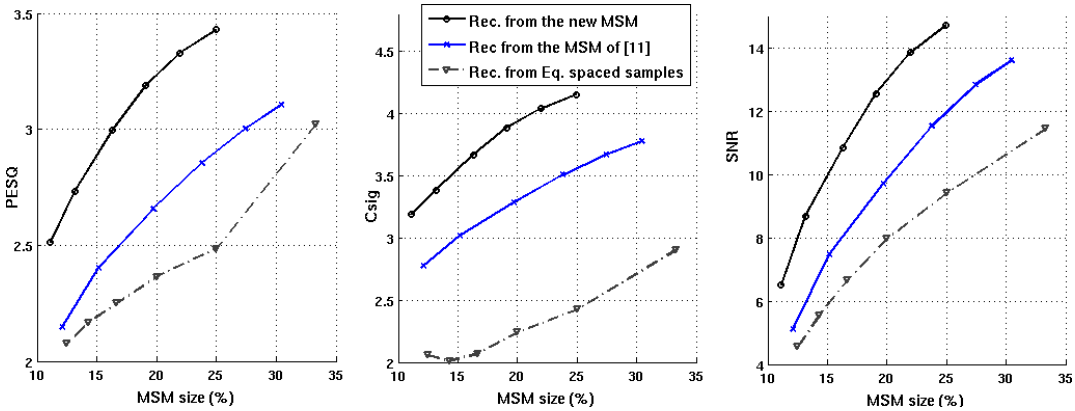


Figure 6.2: Comparison of reconstruction quality for three different sampling methods: the MSM formed using the new measure in Eq. (6.2), the MSM formed by the measure of [126] and the subset of equally spaced samples. The same reconstruction method is used for all these three sampling methods as proposed in section 6.2.2.

In fact, the required data for the reconstruction of the MSM of a signal, which are needed to be encoded in the binary format are:

- **the MSM signal amplitudes:** in order to use fewer bits to encode the MSM signal amplitudes, we encode the gradient values instead. Indeed, the dynamic range of these values is lower compared to the original amplitudes of the signal and hence, they can be quantized with fewer bits. We then use a 6-bit non-uniform quantizer [24] matched to the global histogram of these values. The quantization levels are chosen in a way that the histogram formed by the corresponding bins, has the same number of samples in all bins. This quantization will evidently introduce some quantization error into the gradient values at MSM points that will be propagated by the integration embedded in the approximation part of the reconstruction formula. To avoid this propagation of error, we followed the same principle used in DPCM coding [60], where the gradient value of each sample is replaced with a modified gradient value which is computed off the previously reconstructed sample. Hence, the modified gradient values will be used for the reconstruction using Eq. (6.4).

- **relative position of MSM points on the time axis (Δt):** from our experiments, we observed that the distance between 95% of MSM points is less than 16 samples ($f_s = 16$kHz). Therefore, we dedicate 4-bits to encode the Δt information of successive MSM samples. If, as a result of pauses in the speech signal, the distance is more than 16 samples, a whole 10 bit word would be dedicated to transfer the Δt information. Consequently, there will be about 5% data overhead for the distant MSM samples.

By application of this coder to the full 2.5 hour dataset (converted to $VAF = 0.5$ as classically done), we achieved an avarge perceptual score (Csig) of 3.3 with a

total bit-rate of 18.1 kbps. This is a good coding performance if we compare it with the performance of the G.726 standard of ITU which uses Adaptive DPCM coding and achieves the average PESQ of 3.5 with the bit-rate of 24 kbps [44]. Moreover, we achieved this coding performance using a very simple quantization and we measure the quality on the raw output of the reconstruction without any enhancement. Indeed, the distortions in the reconstructed signal which origins from occasional loss of some intermediate information, may be enhanced with some post-processing. However as our goal was not to achieve the best coding performances, but rather to continue the understanding of the implications in application of the MMF to speech analysis, we confined ourselves to the simplest implementations in this experiment.

6.4 Conclusion

In this chapter we took the very first steps in demonstrating the relevance of the most singular manifold to speech analysis. We showed that a very compact representation of speech signal is possible to achieve using only the information carried by the MSM. We also defined a new multi-scale functional for SE estimation, which is more adapted to particularities of the speech signal and thus provides the highest reconstruction quality. This was done by considering speech as a generic 1-D signal, using a generic reconstruction mechanism. This encouraging results motivates the investigation of physical significance of the MSM: what exactly are the points in the MSM? What is their relationship with the well studied production mechanism of the speech signal? We will address these questions in following chapters and we show that the points belonging to the MSM indeed have important physical significance. This would also provide interesting indications regarding the measure we defined in this chapter.

7

Glottal Closure Instant detection

We saw in chapter 6 that from a MSM of low cardinality, speech signal can be reconstructed with good perceptual quality. This naturally brings up the question about the quiddity of the points in this subset. In this chapter we zoom into the domain of speech signal, to see where exactly the MSM points are located. By doing so, we study the correspondence of the Most Singular Manifold with the physical production mechanism of the speech signal. We show that this subset is indeed related the instants of significant excitations of the vocal tract system. Consequently, we develop an algorithm for automatic detection of these physically interesting instants which are called the Glottal Closure Instants (GCI).

We show that our algorithm has competitive performance to the state-of-the-art methods and effectively over-performs conventional methods in the presence of noise. Indeed, as it is based on both time domain and inter-scale smoothings, it provides higher robustness against some types of noises. In the mean-time, the high geometrical resolution of singularity exponents prevents the accuracy to be compromised. Moreover, the algorithm extracts GCIs directly from the speech signal and does not rely on any model of the speech signal (such as the autoregressive model in linear predictive analysis). We will see in chapter 8 how this property makes the algorithm suitable for the problem of sparse linear prediction analysis.

The structure of this chapter is as follows: section 7.1 presents some background information about the production mechanism of the speech signal and the importance of GCIs. In section 7.2, we briefly review some of the available methods for detection of these important events. Consequently, in section 7.3 we present our preliminary observations regarding the correspondence of the MSM to the GCIs which result in development of a noise robust algorithm for GCI detection in section 7.4. The experimental results are presented in section 7.5. Finally, we draw our conclusions about MSM based GCI detection in section 7.6.

7.1 The significant excitation of the vocal tract

In the classical model of speech production, voiced sound is presented as the output of vocal tract system excited by a quasi-periodic source located in the glottis.

According to the aerodynamic theory of voicing, during the production of a voiced sound, a stream of breath is flowing through the glottis and a push-pull effect is created on the vocal fold tissues that maintains self-sustained oscillation [122]. The push occurs during glottal opening, when the glottis is convergent, whereas the pull occurs during glottal closing, when the glottis is divergent. During glottal closure, the air flow is cut off until breath pressure pushes the folds apart and the flow starts up again, causing the cycles to repeat [122].

In this way, the steady (DC) airflow from the lower respiratory system is converted into a periodic train of flow pulses [115]. The excitation source is hence represented as glottal pulses. However, to a first approximation, the significant excitations of the vocal tract systems (the epochs) can be considered to occur at discrete instants of time (within these pulses) [89]. There can be more than one epoch during a pitch period, but the major excitation usually coincides with the Glottal Closure Instants (the GCIs) [3]. Indeed, when the glottal closing caused the vocal folds to become sufficiently close, the Bernoulli force results in an abrupt closure, which in turn causes an abrupt excitation of vocal tract system [2].

The precise detection of GCIs has found many applications in speech technology. For instance, as the glottal flow is zero immediately after GCIs, the speech signal in this interval represents the force-free response of the vocal tract system [3] and hence, more accurate estimates of vocal tract system can be realized by the analysis of speech signal over this interval (due to the decoupling of the source contribution) [94, 114]. Also, GCIs can be used as pitch marks for pitch synchronous speech processing algorithms, for speech conversion (of pitch and duration) [37], prosody modification [88] and synthesis [31, 117]. GCIs has been also used for speech enhancement in reverberant environments [46], casual-anticasual deconvolution of speech signals [20, 29] and glottal flow estimation [132].

7.2 Review of available methods

The glottal closing (and opening) can be detected accurately and reliably from the contemporaneous Electro-Glotto-Graph (EGG). By a set of skin electrodes on both sides of the larynx, this device provides a non-invasive measurement of the electrical impedance change caused by vocal fold vibration and hence, it can be used for monitoring the vibratory motion of vocal folds. It is known that the GCIs correspond to a rapid decrease in EGG. Consequently, they can detected as the minimum of the differentiated EGG (dEGG) in each pitch period [4].

However, as contemporaneous EGG is not always available, there has been great interest in GCI detection from the speech signal itself and EGG is usually used just as a reference for performance evaluation. Among these algorithms, many of them are based on detection of large values in the residual signal of Linear Predic-

tion (LP) analysis, which is expected to indicate the GCI locations [13]. However, there are practical considerations in epoch extraction from LP residuals as they are vulnerable to noise (and reverberation [46, 92]) and they contain peaks of random polarities [89]. Epoch extraction from LP residuals is extensively studied in [3] and an epoch filtering method is proposed to alleviate the problems in LP based epoch extraction. Still, many of the recent methods use LP residuals for GCI detection as they provide accurate estimates on clean speech [29].

In [89], the impulse like nature of significant excitations is exploited to detect GCIs by confining the analysis around a single frequency. It is argued that an impulsive excitation in the input of vocal tract system, causes discontinuities in the whole frequency range of the output signal. However, the time-varying response of the vocal tract system, makes it difficult to observe these discontinuities directly from the signal. Instead, the effect of these impulses is examined on the output of a narrow band filter centered around a certain frequency. The output of this filter is expected to contain a single central frequency component, while the discontinuities due to impulsive excitations are manifested as deviations from the center frequency. Since the discontinuities due to impulsive excitation are reflected over the whole frequency spectrum (including zero frequency), the authors have opted for zero frequency filtering so as to benefit from the fact that the time-varying characteristics of the vocal tract system is not present at this frequency (as it has resonances at much higher frequencies).

In [110], the properties of the phase spectrum of the speech signal is used for GCI detection. The term phase spectrum refers to the unwrapped phase function (the group delay) of the short time Fourier transform of the signal. It is known that for a minimum phase signal the average slope of the phase spectrum is zero, while for a shifted minimum phase signal, this average will attain a value proportional to the shift. As the impulse response of a minimum phase system (system whose poles and zeros are located within the unit circle) is a minimum phase signal, the average slope of its phase spectrum would depend on the location of excitation impulse. Moreover, note that the phase spectrum of the LP residuals has a similar property. Indeed, as LP residuals are computed by passing the signal through a minimum phase inverse system, the phase slope characteristics of the excitation will not be altered. So finally, in [110], The negative derivative of the phase spectrum of the LP residuals is used to estimate this average and its positive zero crossings are considered as the locations of major excitations, i.e. the GCIs. The advantage of working on LPs is that it minimizes the effects of the position of the analysis window with respect to the impulse response of the vocal tract system, as LP residuals are a first approximation to the excitation signal. The method has performed well, but only for clean speech.

For the the introduction of DYPSA algorithm in [92], it is mentioned that the choice of analysis window size significantly affects the occurrence of zero crossing

in the phase slope function. Ideally, the window should span exactly one impulsive event. When the window is larger, it covers more than one impulsive excitation and hence the zero crossings occur in the mid-way between the two events. The smaller windows causes a raise in false alarm, as spurious zero-crossings would occur in the windows which do not contain any impulsive event. Consequently, the authors use a moderately small window to minimize the risk of missing GCIs. All of these zero-crossings are then taken as candidates (along with additional candidates taken from a procedure called phase slope projection), and then a refined subset is taken as true GCIs by minimizing a cost function using N-best Dynamic Programming. The cost function includes terms considering the spectral quasi stationarity and also the periodic behavior of vocal folds. About the latter, based on the assumption of smooth variations of pitch over short segments, major pitch deviations are penalized heavily (although the method does not require a supplemental pitch estimator). So in effect, for each pitch period, only one of the candidates is picked, which is the one providing the maximum consistency in terms of pitch-period variation. The same idea of dynamic programming is employed in YAGA to refine candidates which are taken by detection of discontinuities in an estimate of voice source signal [2].

Time-scale representation of the voiced speech is employed in [25, 116] for GCI detection. Lines of Maximum Amplitudes (LoMA) across scales in the wavelet transform domain are shown [125] to be organized in *tree* patterns, with exactly one such tree for each pitch period, while the GCI is located at the top of the strongest branch of it. In [25], an algorithm is described for automatic detection of GCIs using LoMA. First a coarse estimation of the fundamental frequency (F_0) is made so as to define the largest scale containing the F_0. Then, following a dyadic wavelet decomposition, LoMA are built according to a local dynamic programming technique. The pitch information is then used to select the scale containing the first harmonic, and thus selecting one [optimal] LoMA per pitch period. The time position along this LoMA is taken as the GCI. Finally, two heuristics are applied to reduce the errors corresponding to detection of more than one GCI per pitch period. The method is shown to compare favorably with EGG data and DYPSA algorithm as a reference.

As mentioned in [32, 29], a class of method use smoothing or measures of energy for GCI detection. The smoothing attenuates the effect of noise, reverberation and vocal tract resonances while preserving the periodicity of the speech signal. The smoothening can be performed on time, or on multiple scale [2, 18]. The drawback is that the accuracy might be compromised. That is why in SEDREAMS method [30], LP residuals are used in conjunction with a smoothing function so as to benefit from accuracy of the residuals in localizing GCIs. The method consists of two steps: first a mean based signal is computed which has the property of oscillating at the local pitch period. Hence, it can be used to locate short regions in each period as expected intervals of GCI presence. The GCIs are then extracted by detecting discontinuities in LP residuals within the determined interval of presence. In this

way, mean based signal provides robustness against noise as it limits the search space within each pitch period and LP residual provides accuracy. This can be considered as a more reliable way of imposing smooth pitch variation constraint to GCI detection, compared to the dynamic programming techniques.

So in conclusion, the success of a GCI detection method lies in the precision of the criterion it uses for localization of GCIs and also on the way of refining true GCIs from spurious candidates, especially in noisy scenarios. Some algorithms use robust estimates of the pitch period so as to restrict their search to one pulse per period (using computationally involved dynamic programming techniques). This adds up to complexity of the algorithm while may sacrificing the accuracy. We propose to use the MSM as a [computationally] simple criterion for localizing the GCIs and we will later explain how we exploit other properties of the singularity exponents to face the problem of noise in a simple but yet robust manner.

7.3 The relevance of MSM to the GCIs

It is shown in [89] that significant impulsive excitations are reflected over the whole speech spectral band. Consequently, excitation impulses would produce *strong* local singularities at different scales of the waveform. This legitimates the use of the multi-scale power-law of Eq. (3.3) to identify and quantify these singularities: it is natural to expect the co-existence of negative transitional SE (Eq. (3.25)) at different scales around these singularities. The summation of these transitional singularities (Eq. (3.24)) would thus result in lower negative values and hence, those points would belong to the MSM (recall that the MSM is defined as the subset of points having the lowest values of SEs).

The relevance of SE values to the instances of significant excitation is illustrated in Fig. 7.1. The top panel shows a part of a voiced sound, while the bottom panel shows the corresponding SE values. The reference GCIs are also shown. It can be seen that $h(t)$ shows a sudden negative peak around GCIs.

Fig.7.2 shows another example which confirms the intuition about the correspondence of the MSM with GCIs. The top panel shows another segment of a voiced sound along with its corresponding pitch marks taken around the GCI points [112]. The bottom panel shows the MSM points of this segment with their value of SE. The MSM is formed as the 5% of samples having lowest value of SE. It can be seen that MSM points are indeed located around the reference GCI. Note also that, around every single GCI, the MSM point with the lowest SE value is the closest one to the GCI mark. This example shows that MSM can indeed identify the locations where significant impulsive excitations occur.

We also made a more objective study, by comparing the MSM against the reference GCIs provided by contemporaneous EGG in the KED database which is free

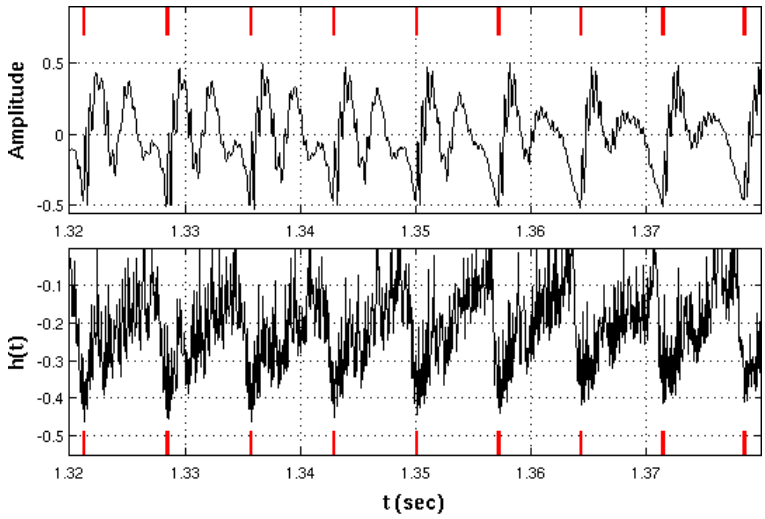


Figure 7.1: **top:** A voiced segment of the speech signal "arctic_a0001" from CMU ARCTIC database [112] and **bottom:** the singularity exponents h(t). The reference pitch marks are represented by vertical red lines.

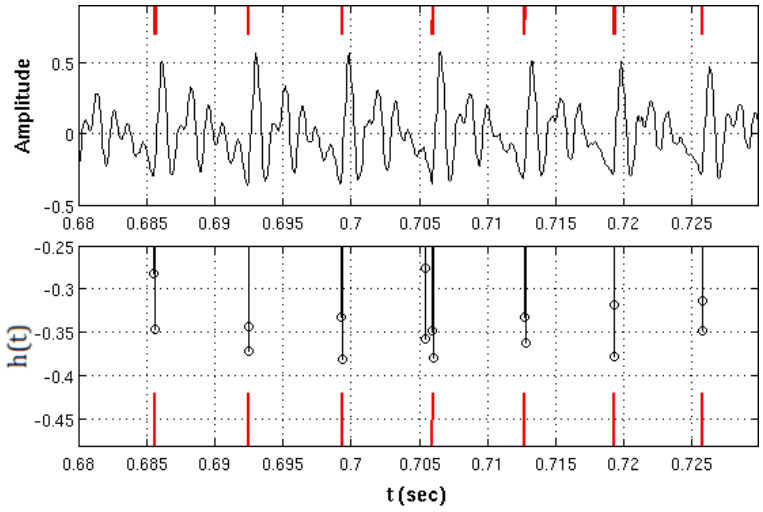


Figure 7.2: **top:** A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from the "CMU ARCTIC" database [112]. **bottom:** MSM samples and their corresponding SE values. The reference pitch marks are represented by vertical red lines.

available on Festvox website [26]. This comparison confirmed this correspondence. We observed that 95% of the points in the MSM (containing 5% of the points having the lowest values of singularity exponents) coincide with the reference GCI points. However, as the 5% density of MSM is not guaranteed to be equal to density of GCIs, to develop an automatic GCI detection algorithm more care must be taken to cope with false alarms and missing GCIs.

7.4 A MSM-based GCI detection algorithm

The preliminary observations presented in section 7.3 showed that MSM effectively coincides with the instants of significant excitations. However, there are some practical considerations in development of an automatic algorithm which must be taken into account. Indeed, practical formation of the MSM requires the specification of a threshold to be applied to singularity exponent values ($h(t)$). However, a global specification of the threshold would be impractical: it may happen that a GCI point does attain a lower h(t) value compared to its surrounding points in one pitch period, but in a larger neighborhood, it may have higher value even compared to non-GCI points. This may specially occur for the starting and ending parts of a voiced phoneme, where the energy of the signal is lower compared to the central parts of phonemes. This case is demonstrated in Fig. 7.3 (around the ime instant of 0.77 sec, there are two GCI points which do have the smallest $h(t)$ in smaller neighborhoods, but in a larger window their $h(t)$ is even larger than non-GCI points). Also, the presence of noise may cause the location of the points having the lowest value of singularity exponent to be slightly changed from desired GCI locations.

To overcome these issues, we first mention that GCIs can be identified using two properties of singularity exponents:

1. In each pitch period, $h(t)$ has the lowest value at the GCI. This is always true for clean speech and hence, the location of *local* minimum in each period, can be taken as the GCI.

2. There is a sharp and clear *level change* before the GCI. Moreover, GCI is the closest point to the instant of level-change. Note that, level change is a relative concept and in our case is independent of the local energy.

Our experiments showed that the first property indeed provided very precise GCI detection in high-energy segments of clean speech, within a single pitch period. However as mentioned before, in a larger window, the GCIs at low-energy parts of speech may attain relatively higher values compared to the non-GCIs belonging to the high-energy parts. Also, the presence of noise may cause a GCI point to attain slightly higher values compared to its immediate neighbors. The second property on the other hand, is a relative quantity itself and local energy have no effect on

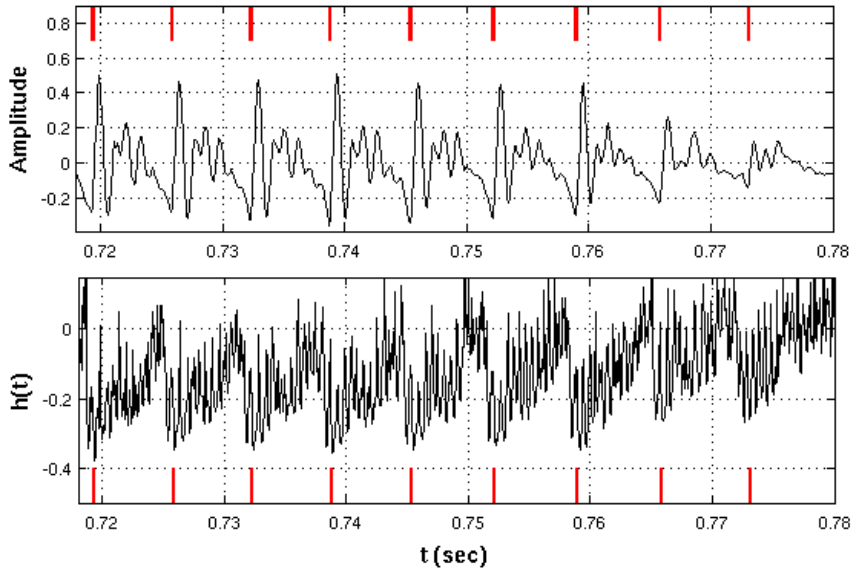


Figure 7.3: **top:** A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from "CMU ARCTIC" database [112]. **bottom:** MSM samples and their corresponding SE values. The reference pitch marks are represented by vertical red lines.

the level-change. Hence, this criterion seems more suitable for GCI detection in segments with lower energy. Also, even the presence of noise would not drastically affect this property. Consequently, we define a new functional to make explicit and easy use of this level-change. We define the level-change functional as:

$$\mathcal{L}_c(t) = \sum_{t-T_L}^{t-\delta t} h(t) - \sum_{t}^{t+T_L} h(t) \tag{7.1}$$

where T_L is a parameter controlling the length of averaging window. Fig. 7.4, illustrates the resulting functional for a segment of voiced speech along with the reference GCIs. It can be seen that indeed, the peaks of $\mathcal{L}_c(t)$ corresponds to the GCIs. Moreover, it oscillates with the pitch period. In that sense, this is similar to the mean-based signal used in [30], with the advantage that the GCI is located on its peak in each period.

Of course, as this functional is obtained through a smoothing procedure, its precision in detection of GCIs may not be competitive with $h(t)$. So we use this $\mathcal{L}_c(t)$ only to limit the search space and we use $h(t)$ itself for final GCI detection. Indeed, the level-change of $h(t)$ occurs once per pitch period. So, in each period, $\mathcal{L}_c(t)$ experiences a peak at GCI which is proceeded by a positive-going zero-crossing and is followed by a negative-going zero-cross (the reason for these zero-cross can be

easily observed in the definition of $\mathcal{L}_c(t)$ in Eq. (7.1) which includes the difference of averages of $h(t)$ on two windows of length T_L, on both sides of any given time instant). As these zero-crossings can be easily detected without any ambiguity, we use them as guiding lines for our algorithm. The final implementation is provided in Algorithm 2.

Algorithm 2 GCI detection procedure

1: Calculate $h(t)$ and $\mathcal{L}_c(t)$.
2: In $\mathcal{L}_c(t)$, for any positive-going zero-cross t_p, find the next negative-going zero-cross t_n.
3: $t_{peak} \leftarrow \underset{t}{\text{argmax}}\ \mathcal{L}_c(t), \quad t \in (t_p, t_n)$.
4: MSM formation: take t_1, t_2, t_3 having the lowest values of $h(t)$ in $t \in (t_p, t_n)$.
5: $t_{msm} \leftarrow \underset{t_i}{\text{argmin}}\ |t_i - t_{peak}|$
6: $t_{gci} \leftarrow 0.5 \times (t_{peak} + t_{msm})$

Note that in step 4 of the algorithm, we take 3 points with the lowest value of singularity exponent so as to cope with noisy scenarios where $h(t)$ at GCI may be slightly higher than one or two of immediate neighbors. That is why the criterion of closeness to the peak of $\mathcal{L}_c(t)$ is used in step 5, to make the final decision. Indeed, $\mathcal{L}_c(t)$ is not simply used for constraining the detection to one detection per period, but rather, as its peak is expected to be located on GCI it is also contributing in increase of accuracy. We emphasize that this algorithm is developed with special care, to avoid any windowing or crucial parameter selection.

7.5 Experimental results

We test our algorithm on CMU ARCTIC databases, which consists of 3 sets of 1150 phonetically balanced sentences, each uttered by a single speaker: BDL (US male), JMK (US male) and SLT (US female) [26]. We also test on the KED Timit database which contains 453 utterances spoken by a US male speaker. All these freely available [26] datasets contain contemporaneous EGG recordings. The reference GCIs are thus taken as the negative peaks of differentiated EGG (manual synchronization of EGG signal and speech recordings are made to compensate for larynx-to-microphone delay). The only parameter to be selected for our algorithm is T_L. The only constraint that we consider for selection of T_L is to keep it smaller than half of the local pitch period for different speakers. On the other hand, the longer it is, the higher the robustness would be against white noise. In the following experiments we have used $T_L = 1.5ms$.

An extensive comparison is made in [32] between stat-of-the-art GCI detection method which shows that for clean speech, SEDREAMS [30] and YAGA [2] have

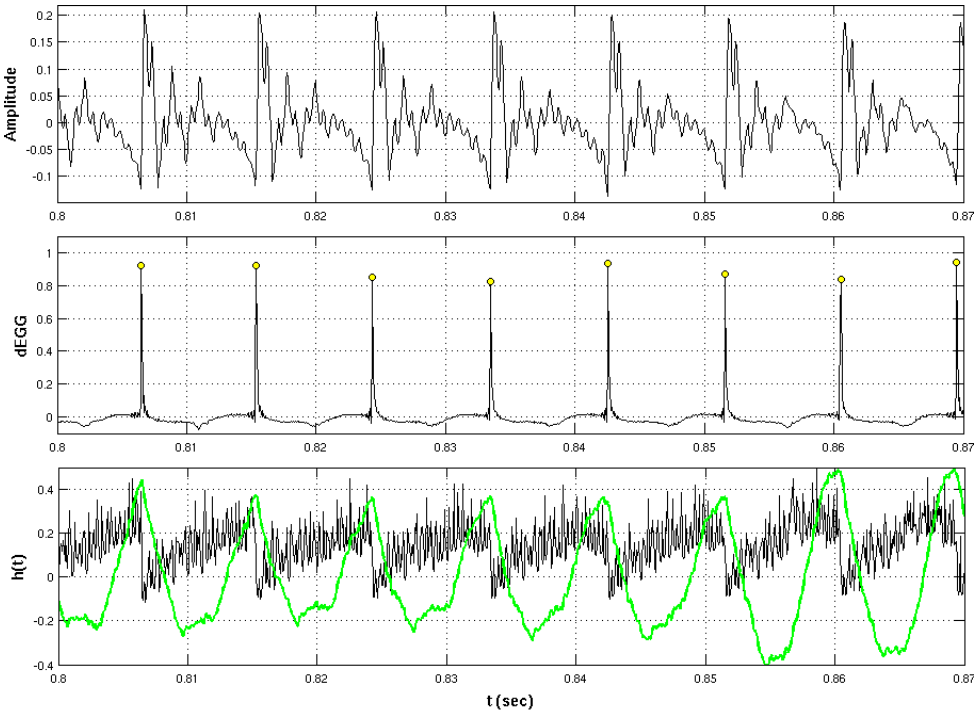


Figure 7.4: **top:** A voiced segment of the speech signal taken from KED database. **middle:** the differenced EGG signal which serves for extraction of reference GCI points. The peaks are marked with yellow circles as the reference GCIs. **bottom:** singularity exponents are shown by black color and the level-change functional $\mathcal{L}_c(t)$ is shown by green color.

the best of performances. In the noisy case however, SEDREAMS significantly overperforms all the other methods. Consequently, we compare our method with SEDREAMS [32] using the implementation that is made available on-line by its author [28]. Note that for each speaker, the EGG signal is synchronized to the speech recordings such that SEDREAMS performance is maximized.

7.5.1 Performance measures

We use the set of performance measures defined in [92] to evaluate the performance of our method. If each reference GCI is denoted by t_{gci_k}, the corresponding larynx cycle can be defined as the range of samples $t \in (\frac{t_{gci_k}+t_{gci_{k-1}}}{2}, \frac{t_{gci_k}+t_{gci_{k+1}}}{2})$. Consequently, two sets of performance measures are defined using the graphical representation in Fig. 7.5. The first set is comprised three measures of the *reliability* of the algorithms:

- Hit Rate (HR): the percentage of larynx cycles for which exactly one GCI is detected.
- Miss Rate (MR): the percentage of larynx cycles for which no GCI is detected.
- False Alarm Rate (FAR): the percentage of larynx cycles for which more than one GCI is detected.

And the second set defines two measures of the *accuracy* of the algorithms:

- Accuracy to $\pm$0.25 ms (A25m): the percentage of larynx cycles for which exactly one GCI is detected and the identification error ζ is less than$\pm$0.25 ms.
- Identification Accuracy (IDA): the standard deviation of identification error ζ (the timing error between the reference GCIs and the detected GCIs in larynx cycles for which exactly one GCI has been detected).

In our experiments, the reference GCIs are extracted from the contemporaneous EGG recordings provided in KED database. We take the significant peaks of dEGG signal as the reference GCIs. For this, manual synchronization is done to compensate for the delay between the EGG recording and the speech signals.

7.5.2 Clean speech

Table 7.1 compares the performance of different GCI detection method for clean speech signals 7.5.1. Overall, it can be seen that SEDREAMS is slightly more reliable, but the accuracy of the two methods are the same. Fig. 7.6 shows histograms of GCI detection timing error for the two algorithms (over the whole four datasets). It can be seen that the distribution of timing error is identification of GCIs are almost the same.

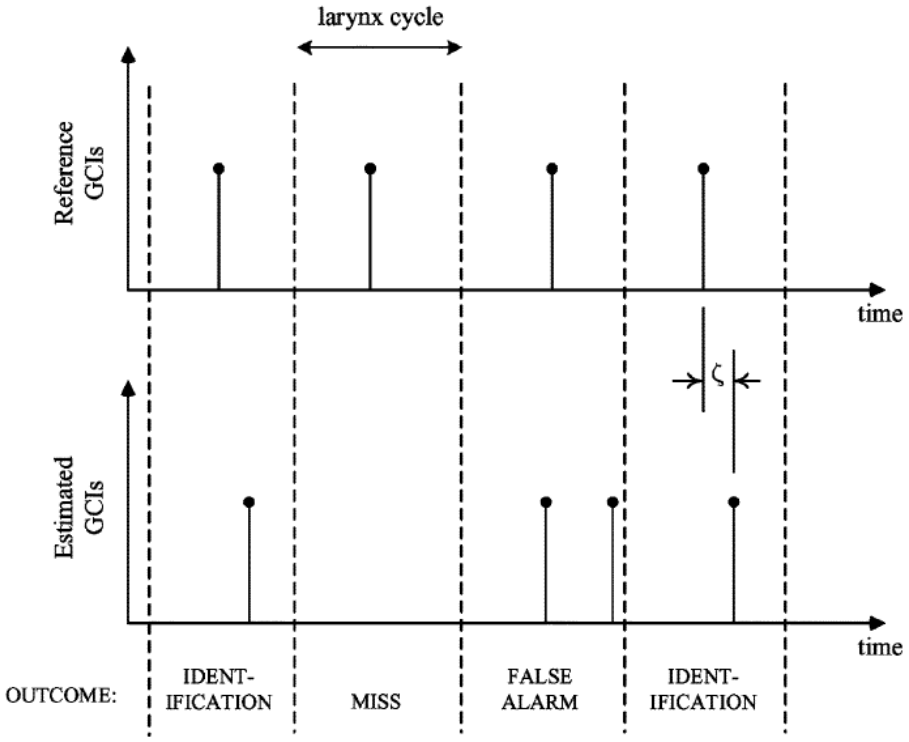


Figure 7.5: Characterization of GCI Estimates showing 4 larynx cycles with examples of each possible outcome from GCI estimation. Identification accuracy is measured by ζ (the graphical representation is taken from [92])

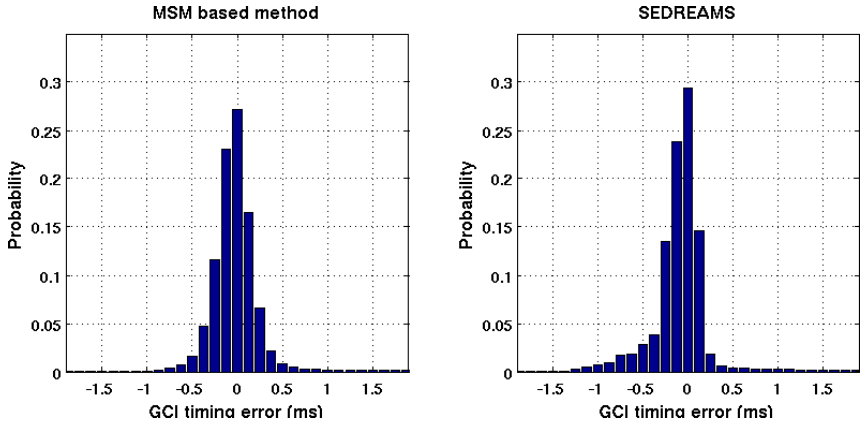


Figure 7.6: Histogram of GCI detection timing error ζ. A reference GCI is considered to be correctly detected when exactly one detection has happened for the corresponding larynx cycle.

Table 7.1: The comparative table of GCI detection performances for clean speech signals.

BDL dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.7	2.7	2.5	0.54	79.5
SEDREAMS	97.4	0.85	1.7	0.38	85.43
JMK dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.9	1.38	3.6	0.55	85.5
SEDREAMS	97.8	0.52	1.6	0.53	78.9
SLT dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.1	4.42	1.4	0.39	80.91
SEDREAMS	98.3	0.02	1.6	0.31	80.25
KED dataset:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	97.4	1.07	1.5	0.39	96.24
SEDREAMS	98.8	0.05	1.14	0.34	94.33
Overall results for four speakers:					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	95.5	2.3	2.2	0.48	82.3
SEDREAMS	98.0	0.4	1.6	0.39	82.5

7.5.3 Noisy speech

To assess the performance of our algorithm in more realistic scenarios, we evaluate its robustness against 14 different types of noises taken from the NOISEX-92 database [93]. We compare the results of our MSM-based algorithm with that of the SEDREAMS method [32], which is shown in [32, 29] to be the most robust method compared to the other state-of-the-art methods.

Fig 7.7 shows the results in presence of different types of noises. To make the comparison easier, we only show two performance measures: the Hit Rate (HR) as a measure of reliability and the Accuracy to ± 0.25 ms as a measure of accuracy. It can be seen that in terms of reliability (Hit Rate), SEDREAMS overperforms in cases of white noise, Babble noise and destroyer engine noise. However, the MSM

based method is more reliable in presence of car interrior noise, factory floor noise, Leopard military car noise and tank noise. For the remaining 7 types of noises, the reliabality of the two methods are quite close, while SEDREAMS shows slightly better results specially for higher SNRs. However, in terms of accuracy, the MSM based methods is showing significantly higher performance for all the 14 types of noises.

SEDREAMS reliabality can be explained by the adaptive control of the window length with a rough estimation of pitch period. This permits the algorithms to smoothen the signal as much as possible. That is why SEDREAMS shows much more reliable results in presence of an uncorrelated noise like white noise. The more accurate result of our MSM-based algorithm compared to SEDREAMS might be explained using the difference between the level-change function in our method and the mean-based signal used in SEDREAMS (both of them are used to constrain the number of detections in each pitch period to one). Apparently both of these functionals serve a similar goal to increase the *reliability* of the algorithms. However, the level-change functional $\mathcal{L}_c(t)$ has two distinctive features that contribute not only to improve the *reliability* of our algorithm but also serves to improve the *accuracy* of it: first, its peak is located on the GCI and hence, it is a smooth (noise-robust) pointer to the GCI (while the mean-based function has no indication about location of GCI). The second difference is that $\mathcal{L}_c(t)$ is a relative quantity which results in its independence from long-range correlations or DC shifts due to changes in energy, or presence of noises like car-noise. It must be always noted however, that the high geometrical resolution of our algorithms is mainly attributed to the geometric resolution of singularity exponents.

7.5.4 Computational complexity

We compare the computational complexity of our algorithm with that of fastSEDREAMS [32], which is shown to be the most efficient algorithm compared to other state-of-the-art algorithms [32, 29]. As the computational complexity of a GCI detection algorithm is highly data-dependent, it is not easy to provide order of computation details [29]. Instead, we use an empirical metric called Relative Computation Time (RCT) and is defined as [32]:

$$\mathrm{RCT}(\%) = 100.\frac{\mathrm{CPU\ time\ (s)}}{\mathrm{Sound\ duration\ (s)}} \tag{7.2}$$

The RCTs for the MATLAB implementation of the two methods are compared in Table 7.2, where the processing times are averaged over the whole database (Note that RCT is a relative quantity that is dependent on the specific processor that is

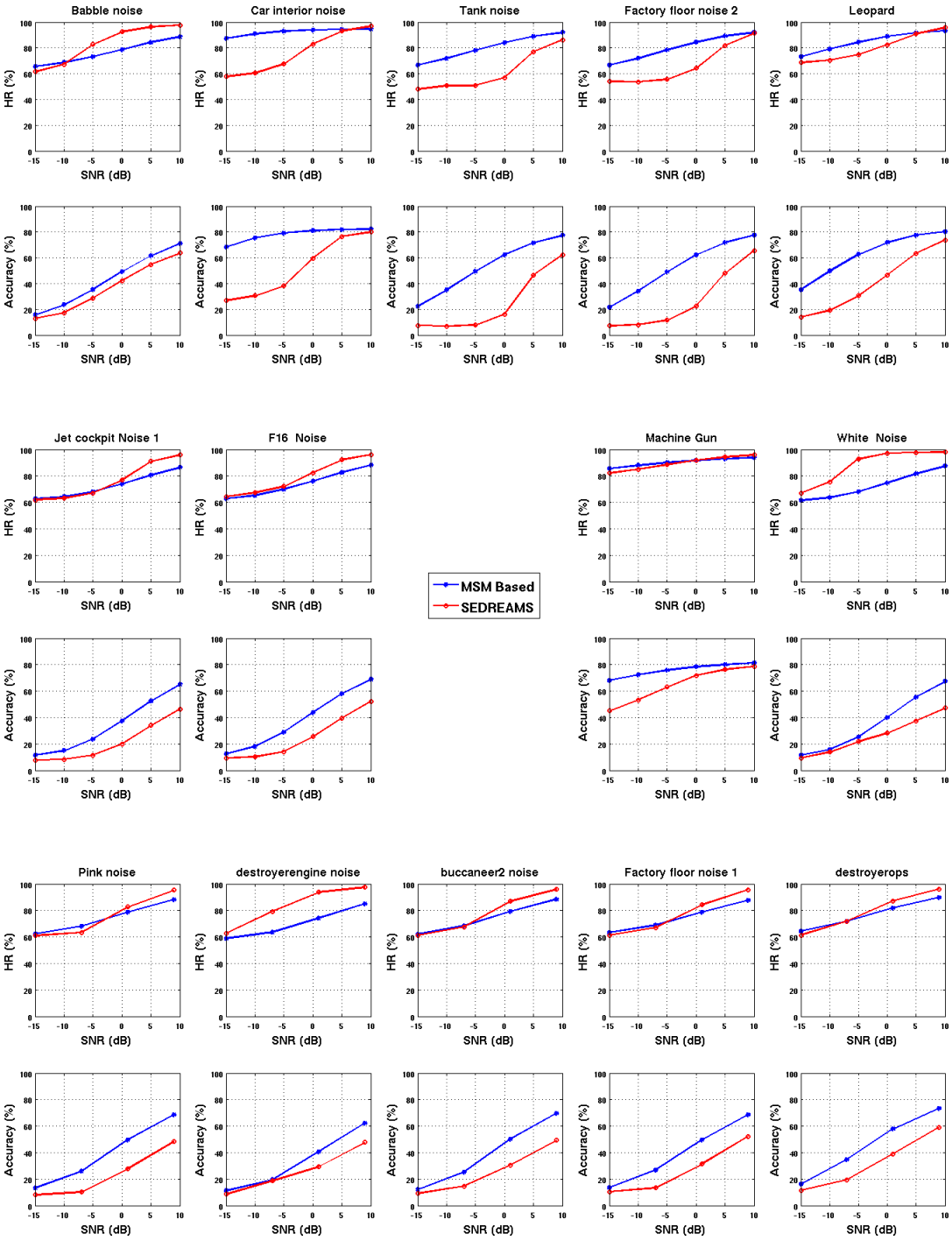


Figure 7.7: Performance comparison in presence of 14 different types of noises taken from NOISEX database [93].

used for the experiment. It can be seen that our MSM based method is almost 30 times faster than SEDREAMS. Also, if we compare with the fast implementation of SEDREAMS [29], the MSM based method is 17 times faster. We use the fast implementation that is provided by the author of SEDREAMS [28]. It must be noted the results of sections 7.5.2 and 7.5.4 are reported using the original implementation and not the fast one. However, the fast implementation is not always as reliable as the original one, specially in the noisy scenarios.

Table 7.2: Comparison between the Relative Computation Time (RCT).

Method	RCT (%)
MSM-based	2.2
SEDREAMS [32]	25.1
fastSEDREAMS [32]	43.8

7.6 Conclusion

In this chapter, we observed that aside from maximization of the *waveform* reconstruction quality, the MSM indeed has significant physical importance as it coincides with the instants of the significant excitation of the vocal tract. We then developed a simple algorithm for detection of these instants which showed competitive performance with the state-of-the-art methods when applied to clean speech. In also showed significantly better robustness against noise. From methodological point of view, our method has this advantage that it is completely model-free and hence, it is not dependent on estimation of parameters of the model. This makes it suitable for the applications where GCIs themselves are used to improve the estimation of model parameters. We show an example of such application in next chapter where MSM-based GCI detection algorithm is used to improve parameter estimations in the framework of linear prediction analysis.

8

Sparse Linear Prediction analysis

We have shown until now that indeed MMF provides precise geometric access to important events of the speech signal: we saw in chapter 7 how MSM corresponds to the instants of significant excitation of the vocal tract called the GCIs. This was demonstrated through comparisons of the resulting GCI estimates with reference data collected by Electro Glotto Graph. In this chapter we present a case study about the practical application of the detected GCIs in main-stream speech technology. We address the problem of sparse Linear Prediction (LP) analysis, which involves the estimation of vocal tract model such that the corresponding LP residuals are as sparse as possible: for voiced sounds, one desires the residual to be zero all the time, except for few impulses at GCIs. Classical LP analysis is based on minimization of l_2-norm of residuals, which fails in providing desired level of sparsity. The standard solutions for this problem are usually complex as they deal with minimization of the l_1-norm of the linear prediction error through complex convex programming optimization techniques. We introduce a simple closed-form solution in this chapter which is based on minimization of weighted l_2-norm of residuals. The weighting function plays the most important role in our solution in maintaining the sparsity of the resulting residuals. We use our MSM-based GCI detector to extract from the speech signal itself, the points having the potential of attaining largest norms of residuals and then we construct the weighting function such that the prediction error is relaxed on these points. Consequently, the weighted l_2-norm objective function can be efficiently minimized by the solution of normal equations of liner least squares problem. The choice of our MSM-based GCI detector is particularly justified, considering the fact that most of successful GCI detection methods use LP residuals for their detection and hence, they cannot be used for constraining the LP problem. Our algorithm is independent of any model that might be assumed for speech signal.

We will see that when compared to classical techniques, our simple algorithm provides better sparseness properties and does not suffer from usual instabilities. We also present an experiment to show how such sparse solution may result in more realistic estimates of the vocal tract by decoupling of the contributions of the excitation source from that of the vocal tract filter. Moreover, to show a potential

application of such sparse representation, we use the resulting linear prediction coefficients inside a multi-pulse synthesizer and show that the corresponding multi-pulse estimate of the excitation source results in slightly better synthesis quality when compared to the classical technique which uses the traditional non-sparse minimum variance synthesizer.

This chapter is organized as follows: In section 8.1 we introduce the general problem of sparse linear prediction. We then provide in section 8.2 the mathematical formulation of the LP analysis problem. In section 8.3 we briefly review the previous work on sparse LP analysis and the numerical motivations behind them. We present our efficient solution in section 8.4. In section 8.5 the experimental results are presented and finally in section 8.6, we draw our conclusion and perspectives.

8.1 Sparse Linear Prediction analysis

Linear Prediction analysis is a ubiquitous analysis technique in current speech technology. The basis of LP analysis is the source-filter production model of speech. For voiced sounds in particular, the filter is assumed to be an all-pole linear filter and the source is considered to be a semi-periodic impulse train which is zero most of the times, i.e., the source is a sparse time series. LP analysis results in the estimation of the all-pole filter parameters representing the spectral shape of the vocal tract. The accuracy of this estimation can be evaluated by observing the extent in which the residuals (the prediction error) of the corresponding prediction filter resemble the hypothesized source of excitation [100] (a perfect impulse train in case of voiced speech). However, it is shown in [100] that even when the vocal tract filter follows an actual all-pole model, this criterion of goodness is not fulfilled by the classical minimum variance predictor. Despite the theoretic physical significance, such sparse representation forms the basis for many applications in speech technology. For instance, a class of efficient parametric speech coders are based on the search for a sparse excitation sequence feeding the LP synthesizer [24].

It is argued in [47] that the reason behind the failure of the classical method in providing such sparse representation is that it relies on the minimization of l_2-norm of prediction error. It is known that the l_2-norm criterion is highly sensitive to the outliers [86], i.e., the points having considerably larger norms of error. Hence, l_2-norm error minimization favors solutions with many small non-zero entries rather than the sparse solutions having the fewest possible non-zero entries [86]. Hence, l_2-norm is not an appropriate objective function for the problems where sparseness constraints are incorporated. Indeed, the ideal solution for sparse residual recovery is to directly minimize the cardinality of this vector, i.e. the l_0-norm of prediction error which yields a combinatorial optimization problem. Instead, to alleviate the exaggerative effect of l_2-norm criterion at points with large norms of error, it is

usual to consider the minimization of l_1-norm as it puts less emphasis on outliers. l_1-norm can be regarded as a convex relaxation of the l_0-norm and its minimization problem can be re-casted into a linear program and solved by convex programming techniques [19].

The l_1-norm minimization of residuals is already proven to be beneficial for speech processing [27, 48, 49]. In [27], the stability issue of l_1-norm linear programming is addressed and a method is introduced for both having an intrinsically stable solution as well as keeping the computational cost down. The approach is based the Burg Method for autoregressive parameters estimation using the least absolute forward-backward error.

In [48], the authors have compared the Burg method with their l_1-norm minimization method using the modern interior points method and shown that the sparseness is not preserved with the Burg method. Later, they have proposed a re-weighted l_1-norm minimization approach in [49], to enhance the sparsity of the residuals and to overcome the mismatch between l_0-norm minimization and l_1-norm minimization while keeping the problem solvable with convex programming tools. Initially the l_1-norm minimization problem is solved using the interior points method and then the resulted residuals are used iteratively, to re-weight the l_1-norm objective function such that less weight is given to the points having larger residual norms. The optimization problem is thus iteratively approaching the solution for the ideal l_0-norm objective function. We also mention that, an interesting review is made in [72, 73], on several solvers for the general problem of mixed l_p-l_0-norm minimization in the context of piece-wise constant function approximation, which indeed their adaptation to the problem of sparse linear prediction analysis can be beneficial (particularly the stepwise jump penalization algorithm, which is shown to be highly efficient and reliable in detection of sparse events).

8.2 Problem formulation

The consideration of the vocal tract filter in the source-filter production model as an all-pole filter results in the well-known autoregressive model for the speech signal $x(n)$:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n) \tag{8.1}$$

where a_k are the prediction coefficients, K is the order of prediction filter and $r(n)$ is the prediction error or the residual. In the ideal case, when the $\{a_k\}$ coefficients are perfectly estimated and the production mechanism verifies the all-pole assumption,

the residual should resemble the hypothesized excitation source. In case of voiced speech, it should be a perfect semi-periodic impulse train which is zero most of the times, i.e., it is a sparse time series. The linear prediction analysis problem of a frame of length N can be written in the general matrix form as the l_p-norm minimization of the residual vector $\mathbf{r}$:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \parallel \mathbf{r} \parallel_p^p, \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a} \tag{8.2}$$

where $\mathbf{a}$ is a vector representing the set $\{a_k\}$ and

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r(N_1) \\ \vdots \\ r(N_2) \end{bmatrix}$$

and $N_1 = 1$ and $N_2 = N + K$ (For $n < 1$ and $n > N$ we put $x(n) = 0$). The l_p-norm is defined as $\parallel \mathbf{x} \parallel_p = (\sum_{k=1}^{N} x(k)^p)^{\frac{1}{p}}$. Depending on the choice of p in Eq. (8.2), the estimated linear prediction coefficients and the resulting residuals would possess different properties.

8.3 Aproaching the l_0-norm

The ideal solution to the LP analysis problem of Eq. (8.2) so as to retrieve the sparse excitation source of voiced sounds, is to directly minimize the number of non-zero elements of the residual vector, i.e. its cardinality or the so-called l_0-norm [22]. As this problem is an N-P hard optimization problem [49], its relaxed but more tractable versions ($p = 1, 2$) are the most widely used.

Setting $p = 2$ results in the classical minimum variance LP analysis problem. Although the latter suggests the highest computational efficiency, it is known that this solution cannot provide the desired level of sparsity, even when the vocal tract filter is truly an all-pole filter [100]. It is known that l_2-norm has an exaggerative effect on the points having larger values of prediction error (the so-called outliers). Consequently, the minimizer puts much effort on forcing down the value of these outliers, with the cost of more non-zero elements. Hence, the resulting residuals are not as sparse as desired.

It is known that this exaggerative effect on the outliers is reduced with the use of l_1-norm and hence, its minimization could be a meliorative strategy w.r.t the minimum variance solution, in that the error on the outliers are less penalized [22]. The solution to the l_1-norm minimization is not as easy as the the classical minimum variance LP analysis problem but it can be solved by recasting the minimization

problem into a linear program [21] and then using convex optimization tools [19]. However, it is argued in [27] that linear programming l_1-norm minimization, suffers from stability and computational issues and instead, an efficient algorithm is introduced, based on a lattice filter structure in which the reflection coefficients are obtained using a Burg method with l_1 criterion and the robustness of the method is shown to be interesting for voiced sound analysis. However, it is shown in [48] that the l_1-norm Burg algorithms behaves somewhere in between the l_2-norm and the l_1-norm minimization. Instead, the authors have shown that enhanced sparsity level can be achieved using modern interior points method [19] of solving the linear program. They have shown interesting results of such analysis and have argued that the added computational burden is negligible considering the consequent simplifications (granted by such a sparse representation) in applications such as open and closed loop pitch analysis and algebraic excitation search.

An iteratively re-weighted l_1-norm minimization approach is consequently proposed by the same authors in [49] to enhance the sparsity of residuals, while keeping the problem solvable by convex techniques. The algorithm starts by plain l_1-norm minimization and then, iteratively, the resulting residuals are used to re-weight the l_1-norm cost function such that the points having larger residuals (outliers) are less penalized and the points having smaller residuals are penalized heavier. Hence, the optimizer encourages small values to become smaller while augmenting the amplitude of outliers [51].

The enhanced sparsity properties of the re-weighted l_1-norm solution compared to the l_1-norm minimization, and also the better performance of the l_1-norm criterion compared to l_2-norm criterion, can be explained numerically with the help of the graphical representation in Fig. 8.1. There, the numerical effect of different residual values on l_p-norm cost functions is graphically depicted. It can be seen that the penalty on outliers is increasing with p. Indeed, as $p \to 0$ the penalty of the corresponding cost function on non-zero values approaches l_0-norm cost function (where any non-zero value is equally penalized and there is no penalization of larger values). This will force the minimization to include as many zeros as possible as their weight is zero. In case of the re-weighted l_1-norm solution [49], any residual is weighted by its inverse at each iteration and hence, the equal penalization of any non-zero value (as in l_0-norm criterion) is achieved. In other words, if a point has a very large (resp. very small) residual, it will be less (resp. much more) penalized in the next iteration and so, the sparsity is enhanced iteratively.

8.4 The weighted l_2-norm solution

We aim at developing an alternative and efficient optimization strategy which approximates the desired sparsity of the residuals. Our approach is based on the min-

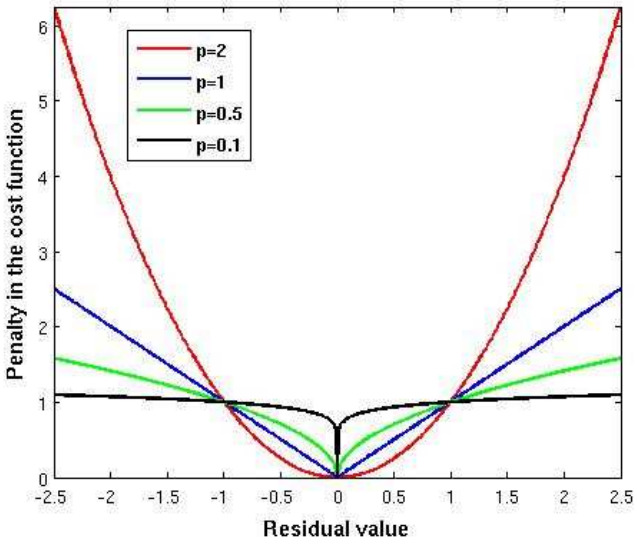


Figure 8.1: Comparison between l_p-norm cost functions for $p \leqslant 2$. The "democratic" l_0-norm cost function is approached as $p \to 0$. The term "democratic" refers to the fact that l_0-norm weights all the nonzero coefficients equally [22].

imization of a weighted version of the l_2-norm criterion. The weighting function plays the key role in maintaining the sparsity of the residuals. Other than pure numerical motivations on de-emphasizing the exaggerative effect of l_2-norm on outliers (as discussed in section 8.3), the design of this function is motivated by the physical production process of the speech signal. We extract from the speech signal itself, the points which are physically susceptible of attaining larger values of residuals and we construct the weighting function such that the error at those outliers is less penalized.

The outliers of LP residuals have an important physical interpretation as their time-pattern follows the pitch period of the speech signal. In other words, they follow the physical excitation source of the vocal tract system, which is a sparse sequence of glottal pulses separated by the pitch period. Indeed, the impulse-like nature of this excitation source is reflected as effective discontinuities in the residual signal [89]: when no significant excitation is presented at the input of the vocal tract system, its output is resonating freely according to the hypothesized all-pole model, and hence, it is maximally predictable by the parameters of the filter. On the other hand, the predictability is minimized when the significant excitations take place and hence, the output signal would be under the influence of both the excitation source and the vocal tract filter. Consequently, LP residual contains clear peaks (outliers) around the instants of significant excitations of the vocal tract system. Hence, if we have a-priori knowledge about these instants we can use this knowledge to impose constraints on the LP analysis problem, so as to relax the prediction error at those

points. By doing so, we ensure that if any enhancement is achieved in the sparsity level of residuals, it also corresponds to the physical sparse source of excitation.

The instants of significant excitations of vocal tract are called the Glottal Closure Instants (GCI) [89]. We use our GCI detection algorithm chapter 7 to find GCI locations and then, the weighting function is constructed such that less emphasize is given to the GCI points. As such, the exaggerative effect on the outliers of the residuals is canceled. We can now proceed to formalize the proposed solution.

8.4.1 Optimization algorithm

We opt for l_2-norm cost function to preserve computational efficiency and then we cope with its exaggerative effect on outliers, by careful down-weighting of the cost function at those points. Formally, we define following optimization problem for the recovery of sparse residuals:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \quad \sum_{k=1}^{N} w(k)(r(k)^2) \tag{8.3}$$

where $w(\cdot)$ is the weighting function. Once $w(\cdot)$ is properly defined, the solution to Eq. (8.3) is straight-forward. Indeed, setting the derivative of the cost function to zero results in a set of normal equations which can be solved as in the classical l_2-norm approach:

$$\hat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{p} \tag{8.4}$$

while in our case, $\mathbf{R} = (\mathbf{W} \odot \mathbf{X})\mathbf{X}^T$, $\mathbf{p} = \mathbf{w} \odot (\mathbf{X}^T\mathbf{x})$, $\odot$ denotes the element-wise product of the two matrices and:

$$\mathbf{w} = \begin{bmatrix} w(N_1) \\ \vdots \\ w(N_2) \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w(N_1) & \cdots & w(N_1) \\ \vdots & & \vdots \\ w(N_2) & \cdots & w(N_2) \end{bmatrix}$$

It is interesting to mention that our experiments show that as long as the smoothness of the $w(\cdot)$ is maintained the stability of the solution is preserved. Indeed, the special form of the input vector $\mathbf{X}$ in Eq. (8.2), is the one used in autocorrelation formulation of LP analysis using l_2-norm minimization. It is proven that autocorrelation formulation always results in a minimum-phase estimate of the all-pole filter, even if the real vocal tract filter is not minimum phase [100]. As our formulation is similar to the autocorrelation formulation, we can fairly expect the same

behavior (though we don't have a theoretical proof). This is indeed beneficial, as having a non-minimum phase spectral estimate results in saturations during synthesis applications. Our experiments show that such saturation indeed never happens. This is an interesting advantage of our method compared to l_1-norm minimization methods which do not guaranty a minimum phase solution, unless if additional constraints are imposed to the problem [48].

8.4.2 The weighting function

The weighting function is expected to provide the lowest weights at the GCI points and to give equal weights (of one) to the remaining points. To put a smoothly decaying down-weighting around GCI points and to have a controllable region of tolerance around them, a natural choice is to use a Gaussian function. We thus define the final weighting function as:

$$w(n) = 1 - \sum_{k=1}^{N_{gci}} g(n - T_k) \tag{8.5}$$

where $T_k, k = 1 \cdots N_{gci}$ denotes the detected GCI points and $g(\cdot)$ is a Gaussian function ($g(x) = \kappa e^{(\frac{x}{\sigma})^2}$). The parameter σ allows the control of the width of the region of tolerance and κ allows the control of the amount of down-weighting on GCI locations. Fig. 8.2 shows a frame of voiced sound along with the GCI points and the weighting function of Eq. (8.5). It can be seen that this weighting function puts the lowest weights around the GCI locations (i.e. the expected outliers) and equally weights the remaining points. Numerically speaking, the minimizer is free to pick the largest residual values for the outliers and it concentrates on minimizing the error on the remaining points (hence the sparsity is granted as explained in section 8.3). This can also be explained with regard to physical production mechanism of the speech signal: as the coupling of excitation source and vocal tract filter is maximized on GCIs, such weighting function assists the minimizer to exclude the points on which the coupling is maximized and concentrate its effort on speech samples where the source contribution is minimized. Such decoupling is investigated in the context of Glottal volume velocity estimation by closed phase inverse filtering techniques [2]. There, the whole time interval on which the glottis is expected to be open is localized and discarded from the analysis frame. Consequently, these methods require the availability of both GCI and Glottal Opening Instants (GOI). However, the determination of GOIs is much more difficult than GCI detection [2]. Moreover, as the analysis window is strictly limited to the closed phase [132], another practical issue may arise: this time-frame might be too short (for high-pitched voices for instance) such that the analysis becomes ineffective [2].

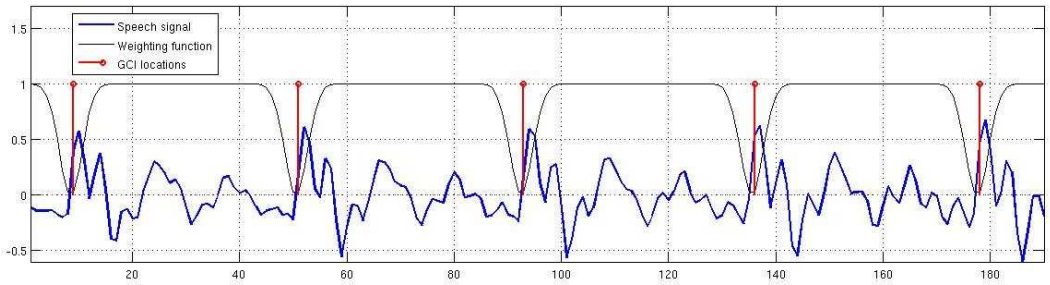


Figure 8.2: A frame of a voiced sound along with the detected GCI locations and the constructed weighting function (with $\sigma = 50$ and $\kappa = 1$).

8.5 Experimental results

We first show the ability of our approach in retrieving sparse residuals for stationary voiced signals and also we show he it can provide a better estimation of the all-pole vocal-tract filter parameters. We then show how such sparse modeling can enhance the performance of a multi-pulse excitation estimation. All the results presented in this section are obtained using the following set of parameters for $w(\cdot)$: $\kappa = 0.9$ and $\sigma = 50$. The choice of the parameters was obtained using a small development set (of few voiced frames) taken from the TIMIT database [45].

8.5.1 Voiced sound analysis

We compare the performance of our weighted-l_2-norm solution with that of the classic l_2-norm minimization and also the l_1-norm minimization via convex programming. For minimization of the l_1-norm, we use the publicly available l_1-magic toolbox [21] which uses the primal-dual interior points optimization [19]. Fig. 8.3 shows the residuals obtained for all these different optimization strategies. It is clear that the weighted-l_2 and also l_1-norm criteria achieve higher level of sparsity compared to the classic l_2-norm criterion. Moreover, a closer look reveals that our weighted-l_2-norm solution shows better sparsity properties compared to the l_1-norm minimization: in the former, each positive peak of residuals is followed by a single negative peak (of almost the same amplitude) while for the latter, any positive peak is surrounded by two negative peaks of smaller (but yet significant) values.

This comparison can be further formalized by using a quantitative measure of sparsity. There exists plenty of such measures on which a review is provided in [57]. Among them, we use the kurtosis, as it satisfies three of the most important properties that are intuitively expected from a measure of sparsity: scale invariance, rising tide and Robin Hood [57]. Kurtosis is a measure of peakedness of a distribution and higher values of kurtosis implies higher level of sparsity. Table 8.1 shows the

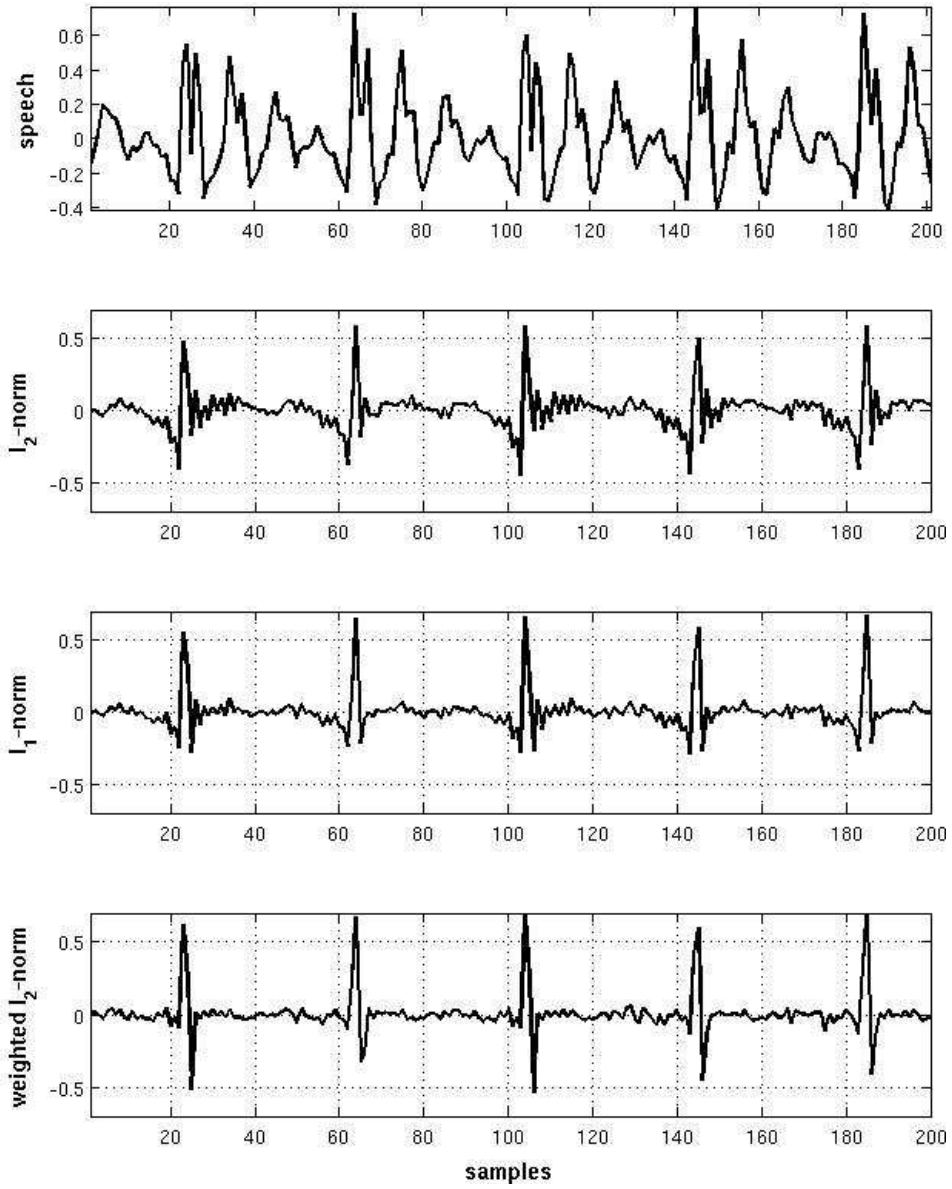


Figure 8.3: The residuals of the LP analysis obtained from different optimization strategies. The prediction order is K = 13 and the frame length is N = 160.

kurtosis of the residuals obtained from the three optimization strategies, averaged over 20 randomly selected sentences of both male and female speakers taken from TIMIT database. From the table, it is clear our method achieves the highest level of sparsity as it obtains the highest values of the kurtosis.

Table 8.1: Quantitative comparison of the level of sparsity of different LP analysis strategies.

Method	kurtosis on the whole sentence	kurtosis on voiced parts
l_2-norm	51.7	39.7
l_1-norm	81.9	65.9
weighted-l_2-norm	85.4	69.1

8.5.2 Estimation of the all-pole vocal-tract filter

We also investigate the ability of our method in estimation of the all-pole filter parameters. To do so, we generate a synthetic speech signal by exciting a known all-pole system with a periodic sequence of impulses (at known locations). We then estimate these parameters from the synthetic signal by LP analysis using our method and the classical l_2-norm method. Fig. 8.4 shows the frequency response of the resulting estimates along with the frequency-domain representation of the synthetic excitation source. It can be seen that for the l_2-norm minimizer, there is a clear shift in the peaks of the estimated filter towards the harmonics of the excitation source. Specifically, the first spectral peak is shifted toward the forth harmonic of the excitation source. Indeed, the effort of l_2-norm minimizer in reducing great errors (the outliers due to the excitation source), has caused the estimated filter to be influenced by the excitation source. However, our weighted-l_2-norm minimization makes a very well estimation of the original all-pole filter and there is no shift in the spectral peaks. This shows that our method effectively decouples the contributions of the excitation source and the all-pole filter (as the source contribution is de-emphasized by the weighting function).

8.5.3 Multi-pulse excitation estimation

The sparseness of the excitation source is a fundamental assumption in the framework of Linear Predictive Coding (LPC) where the speech is synthesized by feeding the estimated all-pole filter by an estimate of the excitation source. The coding gain is achieved by considering a sparse representation for the excitation source. In the popular Multi-Pulse Excitation (MPE) method [14, 109], the synthesis filter is estimated through the classic l_2-norm minimization and then a sparse multi-pulse

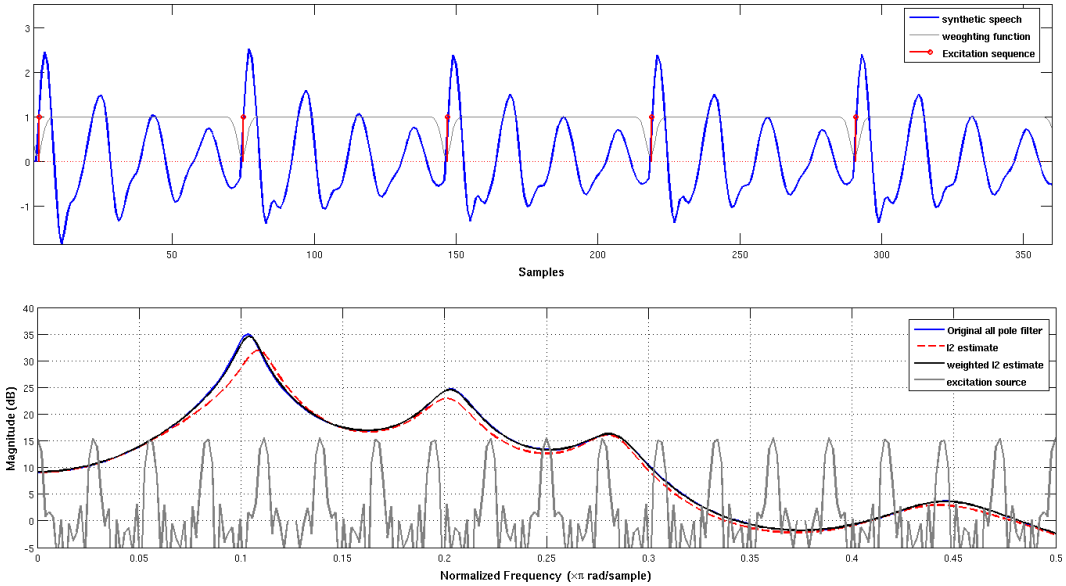


Figure 8.4: **top:** Synthetic speech signal, **bottom:** frequency response of the filters obtained with l_2-norm and weighted-l_2-norm minimization (prediction order $K = 13$). Note that only the first half of the frequency axis is shown so as to enhance the presentation.

excitation sequence is extracted through an iterative Analysis-by-Synthesis procedure. However, as discussed in previous sections this synthesizer is not intrinsically a sparse one. Hence, it would be logical to expect that the employment of an intrinsically sparse synthesis filter, such as the one developed in this chapter, would enhance the quality of the synthesized speech using the corresponding multi-pulse estimate. Consequently, we compare the performance of the classical MPE synthesizer which uses minimum variance LPC synthesizer with the one whose synthesizer is obtained trough our weighted l_2-norm minimization procedure. We emphasize that we follow exactly the same procedure for estimation of multipulse coders for both synthesizers, as in the classical MPE implementation in [109] (iterative minimization of perceptually weighted error of reconstruction).

We have tried to follow the same experimental protocol as in [50]. That is, we evaluate our method using about 1 hour of clean speech signal randomly chosen from the TIMIT database (re-sampled to 8 kHz) uttered by speakers of different genders, accents and ages which provides enough diversity in the characteristics of the analyzed signals. 13 prediction coefficients are computed for frames of 20ms (N=160) and the search for the multi-pulse sequence (10 pulses per frame) is performed as explained in [109]. We evaluate the quality of reconstructed speech in terms of SNR and the PESQ measure [59] which provides a score of perceptual quality in the range of 1 (the worst quality) to 5 (the best quality). The results are

shown in Table 8.2, which shows that our method achieves slightly higher coding quality than the classical MPE synthesizer.

Table 8.2: The quality of Multi-pulse excitation coding using two different synthesizer filters. The sparse excitation source is constructed by taking 10 pulses per 20ms.

Method	PESQ	SNR
MPE + l_2-norm	3.3	9.5 dB
MPE + weighted-l_2-norm	3.4	10.2 dB

Finally, we emphasize that the superior performance of our weighted-l_2-norm solution in retrieving sparse residuals in section 8.5.2, plus the slight improvement of the coding quality in section 8.5.3, was achieved with roughly the same computational complexity as the classical l_2-norm minimization (if we neglect the computational burden of the GCI detector). This is a great advantage compared to the computationally demanding l_1-norm minimization via convex programming (as in [48] or in [49] where multiple re-weighted l_1-norm problems are solved) which also suffers from instability issues. Moreover, another important feature of our solution is that, during the coding experiment we observed that by using the Gaussian shape for the weighting function, the solution is always stable and it does not meet the instability issues as l_1-norm minimization.

8.6 Conclusion

In this chapter we presented a practical application of our MSM-based GCI detection algorithm. We used it to constrain the cost function of LP analysis and we showed that by doing so, the resulting residuals are sparser and hence, they are closer to the hypothesized sparse source of excitation for voiced sounds. Our simple and efficient algorithm has better performance compared to the standard, complex methods which also suffer from instability problems. We mention that the use of MSM based algorithm is particularly justified for this problem, as it extracts GCIs directly from the speech signal without relying on the LP model of speech signal. Moreover, its computational complexity is low which adds up to the simplicity of the solution provided in this chapter. A final note can be made about the better estimation of the vocal tract filter parameters using this method. Indeed, this solution can be seen as [partial] relaxation of one major assumption in linear framework, which is the independent functioning of the excitation source and the vocal tract filter. The down-weighting of the objective function of the optimization problem, can indeed be considered as ignoring the time-instants where the maximal cou-

pling of the source and the vocal filter occurs, and hence, providing a more reliable estimation of filter parameters by inverse filtering.

9

Efficient multipulse approximation of speech excitation using the MSM

In past two chapters, we used the notion of Most Singular Manifold (MSM) to detect the instants of significant excitation of vocal tract in each pitch-period which are called the GCIs. So that, the cardinality of the MSM was restricted to one sample per pitch period. In this chapter we proceed to study the significance of MSMs of higher cardinalities, in the framework of multi-pulse source coding.

Multi-pulse source coding has been widely used and studied within the framework of Linear Predictive Coding (LPC). It consists in finding a sparse representation of the excitation source (or residual) which yields a source-filter reconstruction with high perceptual quality. The MultiPulse Excitation (MPE) method [14, 109] is the first and one of the most popular techniques to achieve this goal. MPE provides a sparse excitation sequence through an iterative Analysis-by-Synthesis procedure to find the position and amplitudes of the excitation source in two stages: first the location of pulses are estimated one at a time by minimization of perceptually wieghted reconstruction error [14]. In the second stage, the amplitude of these pulses are jointly re-optimized to find the optimal pulse values [109].

Using the MSM, we propose a novel approach to find the locations of the multi-pulse sequence that approximates the speech source excitation. We consider locations of MSM points as the locations of excitation impulses and then, the amplitude of these impulses are computed using the second stage of the classical Multi-pulse Excitation (MPE) coder by minimization of the spectrally weighted mean squared error of reconstruction. The multi pulse sequence is then fed to the classical LPC synthesizer to reconstruct speech. The resulting MSM-based algorithm is shown to be significantly more efficient than MPE. We show that MSM provides a good approximation to the locations of the sparse multi-pulse excitations and compare it to the standard MPE method [109] and a recent Compressed Sensing (CS) based approach [50]. The results show that our MSM approach yields similar performances than MPE while it is much faster. They also show that our approach outperforms the CS method, which has roughly the same computational burden as the MPE, when number of pulses per speech frame is low.

The chapter is organized as follows. We first briefly introduce the concept of multi-pulse coding in the framework of parametric speech coding, in section 9.1. In section 9.2 we present the MSM-based algorithm to approximate the multi-pulse source excitation. In section 9.3 the experimental results are presented. Finally, in section 9.4, we draw our conclusion and perspectives.

9.1 Multi-Pulse Excitation coding

The Multi-Pulse Excitation (MPE) coder belongs to parametric class of speech coders, which are motivated by the source-filter production model of the speech signal and are composed of two components:

- A synthesizer filter which resembles the vocal tract filter. This can be the all-pole filter whose parameters (poles) are found through Linear Prediction Analysis (LPA).
- An excitation source for the synthesis filter. This source must be compatible with the assumed source of excitation for voiced and unvoiced sounds as mentioned in chapter 2.

The very basic form of such parametric coder was the LPC coders which use the idealized model of excitation: a train of impulses (separated by pitch period) for voiced sounds and white noise for unvoiced sounds. This idealized (simplified) model of source however fails to provide natural sounding speech, even at high bit-rates [14]. Indeed, the excitation source is not that simple (for instance, the impulse train for voiced sounds are only *quasi*-periodic and the distance between them may [slightly] differ in each glottal cycle). Moreover, the distinction between voiced and unvoiced regimes are not that evident (there might be mixed regimes). Consequently, alternative models of excitation sources got attention of the research community.

The MPE coder was the first parametric coder which achieved the goal of producing natural sounding speech at reasonably low bit-rates, using a simple multi-pulse model of excitation source. The motivation for the method is the evidences regarding the existence of secondary excitation instances (other than the significant ones at GCIs) [54], which may have considerable effect in production of speech sounds. Hence, in MPE coders K impulses are considered for each frame of N samples. K is a parameter that can be tuned according to the requirements: it must be high enough to provide high quality synthesis and in the mean time, it must be low enough to provide reasonable compression gain.

The problem is to find the locations and the amplitudes of these K impulses. The optimum solution is to check for all possible combinations of locations (all the $\binom{N}{K}$ possibilities), and choose the combination which minimizes the error of reconstruction. As this optimum solution is not feasible for efficient hardware implemen-

tations, a sub-optimum solution is considered in MPE: an Analysis-by-Synthesis (AbS) scheme in two stages is employed. In the first one, pulses are added iteratively one at a time by minimizing Perceptually Weighted Mean Squared Error (PWMSE) between the original and reconstructed signal.

The computation required in this first stage is K searches of order N, where K is the number of desired pulses and N is the number of signal samples. In the second stage, once the locations of all pulses are found, their amplitudes are jointly re-optimized such that the PWMSE is minimized [109].

9.2 MSM multi-pulse approximation of source excitation

We saw in chapter 7 that a MSM containing 5% of samples having lowest values of SE, coincide with 95% of the GCIs in a large database of speech signals. There, we needed to limit the density of MSM to one sample per pitch period, so that the GCIs can be reliably estimated. Here, we are interested to study the additional points added to the MSM when its cardinality is not restricted to one point per pitch period. This is true that GCIs are the moments of the most significant excitations of the vocal tract, but this is also known that there may be other excitations present in each pitch period, which are not as significant as those at the GCIs, but still they have some impact on the subtle nuances of the speech signal. It is logical to expect these additional samples in larger MSMs, to correspond to these secondary important excitations.

We study this correspondence in the framework of multi-pulse coder, by using the MSM as an estimate of the locations of multi-pulse excitation sequence of the synthesizer. In our MSM based algorithm, we replace the first stage of MPE, i.e., the iterative search to find the pulse locations, by the following procedure: we form the MSM by taking 2K samples having the lowest SE values. Then, assuming that the pulses are located on the MSM grid, we find their amplitudes using the same joint optimization as in the second stage of the MPE by minimization of the PWMSE. Finally, we choose the K impulses with the highest amplitude as the excitation sequence. Clearly, our approach is computationally more efficient than MPE since the whole first stage of the classical MPE (K searches of order N) is replaced by a simple sorting of SE values to form the MSM.

As for the estimation of singularity exponents, we use the multi-scale functional of Eq. (3.18) along with the estimation method described in 3.3.2.4 so as to have a good estimate of exponents with reasonable computational burden.

9.3 Experimental results

We have tried to follow the same experimental protocol as in [50]. That is, we evaluate our method using about 1 hour of clean speech signal randomly chosen from the TIMIT database (re-sampled to 8 kHz) uttered by speakers of different genders, accents and ages which provides enough diversity in the characteristics of the analyzed signals. 10 prediction coefficients are computed for frames of 20ms (N=160) and the search for impulses are performed separately in each subframes of 10 ms (following the procedure explained in [109]). We use the four finest scales to estimate SE, $j = 4$ in Eq. (3.24). No long-term pitch prediction is performed. All the results we show are without quantization and are almost the same with different randomizations. We compare the performance of our method with the classical MPE [109] and the CS-based method [50].

Before starting this comparison, we first show an example of the reconstruction quality using the MSM method. Fig. 9.1 shows a stationary voiced sound, the MSM excitation sequence with 7 (and also 14) pulses per 20 ms and the corresponding reconstructions. This example shows clearly that our method can indeed yield good reconstruction quality of voiced speech even when using only few pulses.

Fig. 9.2 shows the average normalized reconstruction error ($\bar{e}_N = \frac{\|s-\hat{s}\|_2}{\|s\|_2}$) of the MSM and MPE methods for different number K of pulses per 20ms of speech. This results shows that our method achieves a satisfactory performance but is still less accurate than MPE in terms of mean square error (mse). Still, our method surprisingly outperforms the CS one [50] in terms of mse when $K < 10$ (see Fig.1 in [50]). For instance, for K=8, we achieve $\bar{e}_N = 0.55$ while CS method gives $\bar{e}_N \approx 0.68$. For $K = 10$, which is the typical operating point of a multipulse coder, our method and the CS method perform almost the same. Meanwhile, the CS method has roughly the same computational complexity as the classical MPE, and hence, our method is also much more efficient than the CS one. The computational processing times are compared in Table 9.1, in terms of the average empirical Relative Computation Time:

$$\mathrm{RCT}(\%) = 100.\frac{\mathrm{CPU\ time\ (s)}}{\mathrm{Sound\ duration\ (s)}} \tag{9.1}$$

On the other hand, mse is not the best way to assess the perceptual quality of reconstructed speech. First, our informal subjective listening test showed that the perceptual quality of our method is indeed very close to that of MPE. Especially for K=20, both methods provide almost the same perceptual quality. Second, we evaluated the perceptual quality of reconstructed speech from MSM and MPE using the composite measure of speech quality CMOS [56]. This measure is a combination of PESQ, Cepsterum distance measure, LLR and Itakura-Saito distance. It provides

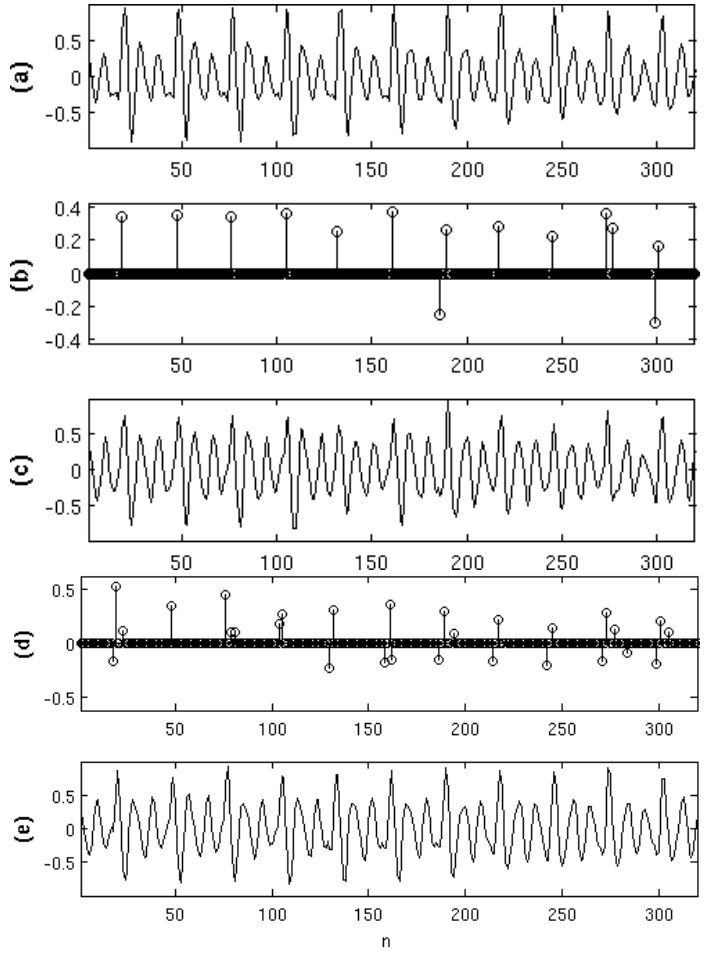


Figure 9.1: (a) a 40 ms segment of stationary voiced speech, (b) the MSM excitation sequence using 7 pulses per 20 ms, (c) the corresponding reconstructed signal, (d) the MSM excitation sequence using 14 pulses per 20 ms and (e) the corresponding reconstructed signal.

a score of perceptual quality in the range of 1 (the worst quality) to 5 (the best quality). The results are shown in Table 9.1. This comparison confirms our informal listening tests. Indeed, the perceptual quality (measured in terms of CMOS) of the MSM and MPE methods are roughly the same.

In summary, all these results suggest that the MSM method achieves similar perceptual quality of reconstruction as MPE [109], with much higher computational efficiency. They also suggest that our method outperforms the recent CS-based method [50] when $K < 10$ in terms of both mean squared error and efficiency. It is noteworthy that there exist more efficient implementations of the MPE, such as the Regular-pulse excitation method [67] to be considered for a fair comparison of computational efficiencies. However, we emphasize that in this work our goal was

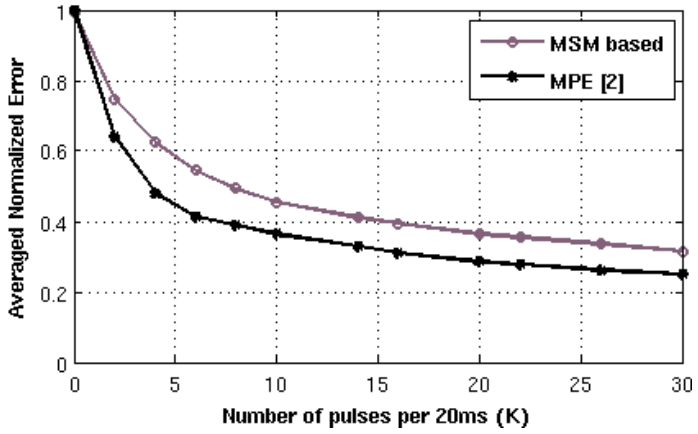


Figure 9.2: The average normalized reconstruction error, averaged over 1 hour of speech signals from TIMIT database.

to assess the significance of the points in larger MSMs (compared to the one in chapter 7 which gives exclusive access to GCIs) w.r.t the speech signal.

Table 9.1: Comparison between the average perceptual quality of reconstruction and the average Relative Computation Time.

Method	K	CMOS	RCT (%)
MSM	10	4.0	9.6
	20	4.2	21.7
MPE [109]	10	4.1	71.9
	20	4.2	143.7

9.4 Conclusion

In this chapter, we used the MSM to efficiently determine the locations of the multipulse excitation sequence of the speech signal. The amplitudes of this sequence are found using the second stage of MPE by minimization of the perceptually weighted error of reconstruction. The resulting algorithm is significantly more efficient than the classical MPE, while the experimental results showed that the MSM algorithm achieves similar perceptual quality as MPE and outperforms the recent CS method in terms of mean squared error when mumber of impulses is smaller than 10. It must be noted our goal was not to design a new coder, but it was rather to show the significance of non-GCI points in a MSM of higher cardinality (compared to what is used in previous two chapters). This has important indications in proving

the ability of singularity exponents in hierarchical classification of points. We will make this point clear in our overall conclusions in the last chapter.

10

Conclusions and perspectives

This thesis presented an exploratory research on the application of a novel non-linear multi-scale formalism to speech analysis, called the MMF. In this final chapter, we will summarize briefly the results of the thesis and draw conclusions about the work, pointing out future research directions that these results suggest.

10.1 Summary and discussion

We have presented in this thesis, our step by step progress in the study of the relevance of the MMF to speech analysis. At the very first glance, MMF seemed as a suitable candidate for such analysis as it relies on geometric local parameters that gives precise access to underlying multi-scale organizations of a complex signal, without relying on stationarity assumptions. This is an important feature of the formalism that implies the relaxation of one limiting assumption of many other linear and non-linear speech analysis approach: the time invariance.

We first ran a set of preliminary experiments investigating the validity of general theoretic conditions for applicability of the formalism to speech analysis. Despite some practical issues hindering the exact evaluation of these conditions, we observed that we can confidently conclude the existence of local scaling behavior for most of the points in domain of the speech signal and hence, precise and meaningful estimates of Singularity Exponents (SE), as key parameters characterizing multi-scale complexities can be achieved.

Consequently, the study of the applicability of the MMF to speech analysis was initiated. To that aim, our strategy in this thesis was to enrich our understanding about the MMF as applied to the speech signal, through the development of practical speech analysis applications. We started by simple adaptation of the formalism to the case of the speech signal as a general 1-D signal, and through the implementation of those applications we updated some of the computational details of the formalism according to particularities of this signal.

The first observation regarding the informativeness of the SEs to speech analysis was made on the time-evolution of local distribution of SEs, which led us to develop a very simple but yet highly accurate phonetic segmentation algorithm. The partic-

ular property of the algorithm was its interestingly high resolution in detection of phoneme boundaries. This accuracy can actually be attributed to the geometric nature of the formalism.

This first successful experiment motivated us to study the use of another major component of the MMF: the MSM that is shown to be the most informative subset of points in the domain of complex signals. It has been indeed shown that a universal reconstruction kernel can reconstruct the whole complex signal using its gradient field restricted to the MSM. However, the reconstruction kernel is originally developed for 2-D signals (using known spectral properties of natural images) and its direct adaptation to the case of 1-D signals results in a simple integration over the gradient field restricted to the MSM.

It is noteworthy that the most powerful definition of SEs in case of natural images is derived by local evaluation of the above mentioned kernel. This in turn related the resulting SE to the local concept of predictability. However, as this kernel reduces to an integration operation for 1-D signal and hence, its local evaluation for estimation of SEs simplifies to evaluation of finite differences. So, the proper definition of SEs for speech analysis was [and still is] an open question, on which we provided a pragmatic answer in chapter 6. We first started the study of the MSM in case of the speech signal, by viewing it as a general 1-D signal: to assess the reconstructability of the signal from the MSM, we employed a generic classical reconstruction technique that can reconstruct any band-limited signal from a set of irregularly spaced samples (just like the MSM). We thus used several definitions of SEs (using different definitions of scale-dependent functional Γ_r) to form the MSM and evaluate the reconstruction quality using this generic reconstructor. We then introduced a new scale-dependent functional for SE estimation (the two-sided variations functional), which resulted in the highest quality of reconstruction (compared to the other functionals). We used this functional in the remaining experiments of this thesis.

Our next step was to study the potential physical meaning of the MSM. By comparing the MSM with locations of the GCIs extracted from Electro Glotto Graph signal, we observed that the MSM can give access to the time instants of significant excitations of vocal tract. Consequently, we developed a GCI detector using the MSM with quite interesting results: In clean speech the algorithm was almost as reliable as the state-of-the-art, while being more accurate. More importantly, the algorithm showed higher robustness in the presence of different types of noises.

As a direct application of our GCI detection algorithm, we then tackled the problem of sparse Linear Prediction Analysis (LPA). The aim of sparse LPA is to achieve a spectral representation whose residuals are sparse. Our strategy was to down-weight the l_2-norm of the LP error on the GCIs such that the optimizer can relax the error at those points and concentrate on minimizing the error on the rest of points. This provided an efficient closed-form solution for the interesting problem of sparse LPA. Apart from numerical motivations for this solution, this work can be seen as

an effort to overcome one major non-valid assumption in classical LPA framework: the independent functioning of the vocal tract and the excitation source. With the proposed strategy we discard the time instants where the coupling of the source and the filter is maximized (the GCIs where the significant excitations of the vocal tract takes place). It is noteworthy that our GCI algorithm has two properties that makes it practically suitable to be used for this weighted l_2-norm solution. Firstly, the computational burden of the algorithm is low (this justifies the use of the whole solution instead of the computationally complex l_1-norm minimizers) and secondly, it does not rely on any model for the speech signal and it extracts the residuals directly from the signal itself (some GCI detection methods are relying themselves on the residuals, and hence are not suitable for modifying the residuals).

We finally studied the use of larger MSMs to estimate the excitation source of the speech signal. With the GCI algorithm the size of the MSM is restricted to one location per glottal cycle. However, in the context of multi-pulse coding, it is required to find more than one location per cycle so as to reconstruct speech signal with good perceptual quality. We thus used MSM to determine the locations of the excitation sequence. We then used the minimization of spectrally weighted reconstruction error (as in the classical approach) so as to find the value of pulse amplitudes. We showed that we can achieve almost the same level of perceptual quality as of the classical techniques but with less computational complexity. The good perceptual quality may indicate that the value of SE indeed provides a hierarchical ranking of speech samples: the smallest one correspond to the GCIs, while the secondary smallest ones correspond to the secondary significant excitations of the vocal tract. Care must be taken however, as the high perceptual quality of MPE synthesizer might be attributed to the objective minimization of perceptually weighted error of reconstruction.

Overall, in this thesis we have demonstrated that MMF provides precise access to important events in the domain of speech signal at different levels: phoneme boundaries, the GCIs and the locations of excitation sequence of the speech signal. We have shown the usefulness of two major components of the MMF (the SEs and the MSM) to speech analysis. Our approach regarding the other important aspect of the MMF, which is the reconstruction of the signal from its MSM was rather pragmatic: as the direct adaptation of the 2-D reconstruction kernel looses information in case of the speech signal, we used two alternatives. First we employed a generic reconstructor called the Sauber-Allebach algorithm and then, for the MPE coding example in chapter 9 we used classical LP based synthesizer.

10.2 Perspectives

A number of potentially promising directions of research can be taken from the results of this thesis. In short term, some improvement or more sophisticated extensions can be readily anticipated to the applications that have been developed in their simplest form in this thesis.

- the high geometrical resolution of the text-independent phonetic segmentation algorithm in chapter 5, suggests the possibility of adapting it to the case of *text-dependent* phonetic segmentation, using statistical models of the speech signal;

- the GCI detector introduced in chapter 7 is particularly an efficient one, with enough reliability. This suggests the investigation of its applicability in pitch-synchronous speech processing problems. This may serve as more concrete (and subjective) demonstration of the ability of the algorithm in detecting meaningful events in the speech signal;

- the efficient GCI detector of chapter 7 together with the closed-form stable solution for sparse linear prediction analysis in chapter 8, open the way to investigate the usefulness of such sparse representation, in any high-level practical speech processing application such as speech recognition, speaker identification, synthesis. In particular, the accurate estimates of the vocal tract filter may have very high potential in text-to-speech synthesis applications, where proper modeling of excitation source and synthesizer filter is a serious concern. The first interesting subject to investigate is the possible improvement of synthesis quality using the residuals of this sparse representation to model the excitation source for HMM-based speech synthesis. This synthesizer, in its basic form, uses a very simple model of excitation source: one pulse per period. This is the highest level of sparsity and is too far from reality with the classical minimum variance modeling. So it is reasonable to expect that the use of a representation whose residuals are effectively sparser, may improve the synthesis quality;

- another interesting possible subject may arise by considering the results of chapters 8 and 9 as a whole. In chapter 8, we showed how weighting of the objective function on GCIs may assist the optimizer to relax the prediction error at those points and to focus on minimizing the error at remaining time instants. In chapter 9, we showed how the MSM can provide direct access to a subset of points which are interesting for multi-pulse coding of speech. However, we used the classical synthesizer for reconstruction of the speech, which is found through the minimization of l_2-norm error. One can fairly expect that a similar weighting of the l_2-norm error, but on the MSM (not

just on the GCIs) may improve quality. Of course there are practical issues to overcome, such as instabilities or the proper design of weighting function. Such combination would result in a unified coding framework based on the MMF;

Apart from the perspectives drawn from the applications introduced in this thesis, a number of applications might be considered on which the MMF may potentially give rise to appreciable results:

- the development of a new feature set for all classification problems in the context of statistical speech processing (recognition, identification and synthesis tasks). Our preliminary observations show that such development is indeed possible. For instance, we have observed that the time evolution of time-conditioned histogram of singularity exponents (the one shown in Figure 5.1 in chapter 5) exhibits consistent patterns for different realizations of the same phoneme. As such, this histogram can be readily considered for all classification tasks;

- the successful detection of the GCIs and also the multi-pulse excitation sequence, also suggests the investigation for detection of Glottal Opening Instants using the MMF, which are much more difficult to detect than GCIs.

All these potential extensions to the present work, may add up to substantiate the establishment of the MMF as a powerful tool for the analysis of the speech signal. However, another important track for future research is to take a similar strategy such as the one taken for 2-D natural images to attain more powerful definitions of SEs. As mentioned in section 10.1, our strategy regarding the reconstruction of speech from its MSM was only pragmatic. This actually provided us with interesting results. However, noting that the most powerful definition of SEs is realized in case of the 2-D signals by local evaluation of an appropriate reconstruction kernel, one can fairly expect that the definition of appropriate recunstruction kernels in the case of speech signals, and the consequent definition of the corresponding scaling exponents, would greatly contribute to improvement of the results of this thesis.

A

Résumé français

Cette thèse présente une recherche exploratoire sur l'application du Formalisme Microcanonique Multiéchelles (FMM) à l'analyse de la parole. Dérivé de principes issus de physique statistique, le FMM permet une analyse géométrique précise de la dynamique non linéaire des signaux complexes. Il est fondé sur l'estimation des paramètres géométriques locaux, appelés les Exposants de Singularité (EdS), qui quantifient le degré de prédictibilité à chaque point du domaine du signal. S'ils sont correctement définis les EdS peuvent fournir des informations précieuses sur la dynamique locale de signaux complexes et ils ont ainsi été utilisés dans plusieurs applications allant de la représentation des signaux à l'inférence ou la prédiction.

Nous démontrons la pertinence du FMM en analyse de la parole et développons plusieurs applications qui montrent le potentiel et l'efficacité de ce formalisme dans ce domaine. Ainsi, dans cette thèse, nous introduisons: un nouvel algorithme performant pour la segmentation phonétique indépendante du texte, un nouveau codeur du signal de parole, un algorithme robuste pour la détection précise des instants de fermeture glottale, un algorithme rapide pour l'analyse par prédiction linéaire parcimonieuse et une solution efficace pour l'approximation multipulse du signal source d'excitation.

a.1 Caractère non linéaire du signal de parole

Il est classiquement supposé que le signal de parole est produit par un modèle source-filtre: une source d'excitation est d'abord générée par un mécanisme de perturbation d'air qui varie en fonction du type de la voix. Cette source d'excitation est filtré ensuite par un filtre dont les caractéristiques dépendent de la forme du conduit vocal [101]. Les méthodes d'analyse linéaire simplifient l'effet du système du conduit vocale en le considérant comme un filtre linéaire (opération de convoloution). D'ailleurs, la source d'excitation est idéalisée comme des impulsions périodiques pour les sons voisés, le bruit blanc pour les fricatives et les impulsions isolées pour plosives. Ce modèle simplifié est le fondement de la plupart des méthodes courantes de traitement de la parole.

Cependant, il y a beaucoup de preuves théoriques et expérimentales concernant l'existence d'effets non linéaires dans le mécanisme de production de la parole qui sont généralement ignorées dans les approches linéaires:

- la source d'excitation des sons non voisés est turbulente [61, 68], mais les techniques linéaires la considerent comme écoulement *laminaire* d'air [38].
- la source d'excitation des plosives possède un écart de temps et une composante turbulente [100], mais elle est idéalisé comme une source impulsive.
- il y a des évidences concernant le caractérisation de sons voisés par des débits d'air très complexes comme les jets et tourbillons dans le système du conduit vocal [121]. En outre, les vibrations des cordes vocales ne sont pas exactement périodique tel qu'il est idéalisé dans le modèle linéaire [84].
- Il est supposé dans le cadre linéaire que le conduit vocal et la source d'excitation fonctionnent de façon indépendante [36, 38, 68, 84]. Mais en pratique, lorsque la glotte est ouverte, il existe un couplage entre les deux qui se traduit par des changements significatifs des formants.

La pertinence des hypothèses linéaires sont testés dans [76, 77] et il est démontré que les hypothèses mathématiques de la théorie des systèmes linéaires invariants dans le temps ne sont pas entièrement conformes à la dynamique de la parole. Le fait que le mécanisme de production du signal de parole n'est pas entièrement linéaire a motivé une tendance vers une exploration de ce caractère non-linéaire.

Il y a une vaste gamme de méthodes non linéaires pour analyse de la parole [38, 68, 77, 84, 23, 111, 118, 124]. Parmi eux, nous nous intéressons aux methodes qui associent ces effets non-linéaires à la nature turbulente du flux d'air dans le conduit vocal et utilisent des outils et méthodes dans l'étude des systèmes chaotiques et turbulents. Cependant, ces méthodes sont souvent mathématiquement complexe et ils sont parfois fondées sur des hypothèses irréalistes comme la stationnarité. En outre, ils s'appuient généralement sur des mesures globales et ne parviennent pas à fournir un accès local à la dynamique du signal de parole. Dans cette thèse, nous utilisons un nouveau formalisme (le FMM) qui permet une analyse précise de la dynamique locale des signaux complexes.

a.2 Formalisme Microcanonique Multiéchelles

Le FMM est basé sur le calcul des exposants d'échelle locaux d'un signal donné dont leur distribution est la quantité clé définissant sa dynamique intermittente. Ces exposants, appelés les Exposants de Singularité (EdS), sont utiles pour l'étude des propriétés géométriques des signaux, et ils ont été utilisés dans une grande variété d'applications allant de la compression de données à l'inférence et la prédiction [15,

52, 80, 127, 128, 134, 133]. Ces exposants sont définis par l'évaluation d'une loi de puissance multi-échelles. Étant donné un signal s, pour au moins une fonctionnelle Γ_r dépendante de l'échelle r, la relation suivante doit être valide à tout instant t [131]:

$$\Gamma_r(s(t)) = \alpha(t)\, r^{h(t)} + o\left(r^{h(t)}\right) \qquad r \to 0 \tag{A.1}$$

où $h(t)$ est l'EdS du signal s à l'instant t. L'objectif consiste à faire un bon choix de Γ_r de sorte qu'une estimation précise des EdS est atteint. Si la fonctionnelle est choisie comme l'incrément linéaire, $\Gamma_r(s(t)) = |s(t+r) - s(t)|$, les exposants qui en résultent sont les exposants de Hölder. Cependant, il est typiquement difficile d'obtenir une bonne estimation de ces exposants pour les données empiriques (en raison de la discrétisation, du bruit et des corrélations à longue portée). Turiel et al. [131] ont proposé un choix plus robuste pour Γ_r défini à partir des caractérisations typiques de l'intermittence en turbulence: $\Gamma_{\mu_r}(s(t)) = \int_{B_r(t)} \mathcal{D}s(t)dr$ ou $\mathcal{D}$ est un opérateur différentiel approprié comme la norme du gradient $|\nabla s|$. Basé sur la même fonctionnelle, dans cette thèse nous proposons d'utiliser $\mathcal{D}_\tau s(t) = |2s(t) - s(t-\tau) - s(t+\tau)|$ pour le cas du signal de parole. En pratique, cette choix de Γ_r fournit les meilleurs résultats dans la plupart des applications dans cette thèse. Après le choix de Γ_r, l'estimation peut se faire par plusieurs méthodes introduites dans la thèse.

Une fois définies et estimées avec précision, les valeurs des EdS fournissent des informations précieuses sur la dynamique locale des signaux complexes et ils peuvent quantifier le degré de régularité ou bien le degré de prédictabilité: plus h(t) est petit moins le système est régulier (prédictible) en t. Ainsi, un deuxième élément important du formalisme, appelé la Variété la Plus Singulière (VPS), est défini comme l'ensemble des points du signal dont ses EdS sont plus petit à une certaine précision numérique. Il a été établi que les transitions critiques du système se produisent en ces points: dans le cas d'images naturelles, il a été montré que la totalité du signal peut être reconstruit uniquement à partir des informations contenues dans la VPS.

a.3 Applications

Dans cette thèse, nous demontrons d'abord que le formalisme est valide pour l'analyse du signal de parole, puis nous procédons à l'étude de sa pertinence pour des applications réalistes de traitement de la parole.

a.3.1 Segmentation phonétique indépendante du texte

Nous commençons par l'étude de la distribution locale des EdS. Nous observons qu'il ya une différence significative dans ces distributions entre les phonèmes voisins. Nous proposons donc d'utiliser la primitive des EdS ($ACC(t) = \int_0^t d\tau\, h(\tau)$) comme

une représentation quantitative de ces changements. La fonctionnelle qui en résulte est une courbe linéaire par morceaux, avec des changements abrupts de pente aux frontières des phonèmes. Pour le développement d'un algorithme automatique de segmentation, nous ajustons une courbe linéaire par morceaux à ACC et identifions les points de rupture comme les frontières entre phonèmes.

Nous montrons que cet algorithme simple est significativement plus précis qu'une méthode de l'état de l'art [33]. Ensuite, en effectuant une analyse d'erreur de cette implémentation très simple, nous développons un algorithme en deux étapes: dans la première étape, nous utilisons notre algorithme original pour détecter les frontières dans le signal original et sa version filtrée. Nous recueillons toutes les frontières détectées et les considérons comme des candidats. Dans la deuxième étape, nous prenons la décision finale en effectuant un fenêtrage dynamique sur ces candidats suivi d'un test de rapport de vraisemblance sur les distributions des EdS du signal original. Ce nouveaux algorithme surpasse notre méthode originale et il est aussi nettement plus précis que [33].

a.3.2 Codage en forme d'onde

Le résultat encourageant de la segmentation phonetique obtenu par une simple analyse des EdS nous motive à étudier la pertinence de l'autre composante du FMM pour l'analyse de la parole: la VPS. Nous commençons cette étude en évaluant la constructibilité du signal de parole à partir de la VPS. Dans le cas des images naturelles un noyau 2D de reconstruction est défini qui peut reconstituer l'image eniere à partir de l'informations de gradient dans la VPS. Toutefois, l'adaptation directe de ce noyau 2D pour un signal 1D (comme la parole) résulte a l'interpolation arrondie (aussi appelée interpolation par plus proche voisin) qui n'est pas approprié pour la reproduction de la dynamique complexe de la parole. Au lieu de cela, nous utilisons l'algorithme Sauer-Allebach [105], qui est classiquement utilisé pour la reconstruction d'un signal à partir d'un sous-ensemble de ses échantillons irrégulièrement espacées (ce qui est le cas de la VPS).

Nous comparons la qualité de perception de la parole reconstruite à partir de VPS différentes (qui sont générées en utilisant des Γ_r différents). Nous montrons que la Γ_r nouvellement définie obtient la meilleure qualité. Par conséquent, par une quantification de 10 bits de la VPS, nous développons un codeur en forme d'onde de la parole.

a.3.3 Détection des instants de fermeture glottale

Nous étudions la signification physique des points dans la VPS en évaluant leur correspondance aux Instants de Fermeture Glottale (IFG). Les IFG forment l'essentiel

de l'excitation du conduit vocal durant la production des sons voisées. La détection de ces instants a en effet de nombreuses applications dans la technologie vocale.

La première observation montre que la VPS a une correspondance intéressante aux IFG: les plus faibles valeurs des EdS (la VPS) en zone voisée sont directement liées aux IFG. Par conséquent, nous développons un algorithme automatique pour détecter les IFG à partir des EdS. Nous introduisons une nouvelle fonction de lissage qui donne une localisation brute des IFG. Nous utilisons cette dernière pour limiter l'espace de recherche et après nous employons les EdS eux-mêmes pour faire la détection finale. En évaluant notre algorithme sur des bases de données bien connues, nous comparons notre algorithme à une méthode de l'état de l'art [32]. Nous montrons que notre algorithme est plus efficace, et plus robuste en présence de bruit.

a.3.4 Analyse par Prédiction Linéaire parcimonieuse

Pour démontrer un exemple d'application de notre algorithme de détection des IFG, nous considérons le problème de l'Analyse par Prédiction Linéaire (APL) «parcimonieuse». L'objectif est d'estimer les paramètres du filtre de prédiction tels que les résidus qui en résultent (l'erreur de prédiction) soient aussi parcimonieux que possible (ayants le plus petit nombre d'éléments non-nuls).

La solution classique d'APL consiste à minimiser la norme l_2 des résidus. Toutefois, cette methode préfère une solution avec une variance minimale des résidus plutôt qu'une solution avec des résidus parcimonieux. En effet, la norme l_2 amplifie dramatiquement les erreurs le plus grands et elle donc préfère une solution avec les erreurs le plus petites plutôt que celles avec les erreurs le plus parcimonieuses. Cependant, comme la résolution de problemes d'optimization de norme l_2 est facile et stable, cette solution a été utilisée dans la plupart des applications. Afin d'obtenir un soloution plus parcimonieuse, on peut minimiser la norme l_1 qui est plus tolérable vers les grandes valeurs d'erreur. Mais la minimisation de la norme l_1 est plus complexe et peut provoquer des problèmes d'instabilité.

Nous proposons de minimiser une version pondérée de la norme l_2. La fonction de pondération est construite de telle façon a porter moins l'accent sur les points susceptibles de produire les grandes valeurs d'erreur (les IFG). Ce choix est motivé par le fait que les erreurs le plus grandes sont censés être situé aux IFG. Ainsi, nous utilisont notre algorithme de détections des IFG pour trouver ces points et nous les attribuons les poids les plus petits. En tant que tel nous avons atteint une solution plus parcimonieuse tout en bénéficiant de la simplicité et de la stabilité de minimisation de la norme l_2.

a.3.5 Approximation multipulse de l'excitation

Durant ces deux dernières applications, nous avons utilisé la notion de VPS pour détecter les instants d'excitation significative de l'appareil vocal dans chaque période d'hauteur (les IFG). Donc la cardinalité de la VPS était égal à un échantillon par période d'hauteur. Nous procédons à étudier la signification des VPS avec une plus grande cardinalité dans le cadre du codage paramétrique de parole (le modèle source-filtre de codage de la parole).

Le problème consiste à trouver des emplacements et des valeurs d'une séquence d'excitation parcimonieuse, avec laquelle on obtient une reconstruction source-filtre de bonne qualité de perception. Dans la solution classique, la séquence d'excitation est trouvée à partir d'un processus itératif d'analyse-par-synthèse de deux étapes: tout d'abord, les emplacements des impulsions sont trouvés un par un, par la minimisation de l'erreur de reconstruction. Dans la deuxième étape, les amplitudes de ces impulsions sont optimisés conjointement pour trouver leur valeurs optimales. En utilisant la VPS, nous proposons une nouvelle approche pour trouver l'emplacement de la séquence d'excitation. Nous prenons les emplacements des points dans la VPS comme les emplacements des impulsions d'excitation. Ensuite, les amplitudes de ces dernières sont calculées en utilisant la deuxième étape de la solution classique. Notre algorithme est nettement plus rapide que la solution classique tout en offrant une qualité de perception identique de reconstruction.

a.4 Conclusions

Dans cette thèse, nous avons demontré le potentiel du FMM pour l'analyse de la parole en développant plusieurs applications intéressantes et en les comparant avec les méthodes de l'état de l'art. Par rapport à d'autres méthodes non linéaires, les algorithmes que nous avons développés sont tous plus simples et plus efficaces (même par rapport aux techniques linéaires). Les résultats encourageants que nous avons obtenus dans cette thèse motivent une étude plus approfondie de ce formalisme pour d'autres applications de traitement de la parole.

B

Appendix A: Phonemic symbols

	Symbol	example word	DESCRIPTION
STOPS:	/b/	bee	voiced
	/d/	day	voiced
	/dx/	dirty	voiced
	/g/	gain	voiced
	/p/	pea	unvoiced
	/t/	tea	unvoiced
	/k/	key	unvoiced
	/q/	bat	unvoiced
FRICATIVES:	/s/	sea	unvoiced
	/sh/	she	unvoiced
	/z/	zone	voiced
	/zh/	vision	voiced
	/f/	fin	unvoiced
	/th/	thin	unvoiced
	/dh/	that	voiced
	/v/	van	voiced
NASALS:	/m/	mom	voiced
	/n/	noon	voiced
	/ng/	sing	voiced
	/em/	bottom	voiced
	/en/	button	voiced
	/eng/	washington	voiced
	/nx/	winner	voiced

Table B.1: Phonemic symbols used in TIMIT

	Symbol	example word	DESCRIPTION
VOWELS:	/iy/	beet	voiced
	/ih/	bit	voiced
	/eh/	bet	voiced
	/ey/	bait	voiced
	/ae/	bat	voiced
	/aa/	bott	voiced
	/aw/	bout	voiced
	/ay/	bite	voiced
	/ah/	but	voiced
	/ao/	bought	voiced
	/ow/	boat	voiced
	/uh/	book	voiced
	/uw/	boot	voiced
	/ux/	toot	voiced
	/er/	bird	voiced
	/ax/	about	voiced
	/ix/	debit	voiced
	/axr/	butter	voiced
	/ax-h/	suspect	voiced
AFFRICATES:	/jh/	joke	voiced
	/ch/	choke	unvoiced
SEMIVOWELS	/l/	lay	voiced
& GLIDES:	/r/	ray	voiced
	/w/	way	voiced
	/y/	yacht	voiced
	/hh/	hay	voiced
	/hv/	ahead	voiced
	/el/	bottle	voiced
OTHERS:	/pau/	-	pause
	/epi/	-	epnthetic silence
	/h#/	-	begin/end marker (non-speech events)
	/1/	-	primary stress marker
	/2/	-	secondary stress marker

Table B.2: Phonemic symbols used in TIMIT

Bibliography

[1] G. Almpanidis, M. Kotti, and C. Kotropoulos. Robust detection of phone boundaries using model selection criteria with few observations. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:287–298, 2009.

[2] M.R.P Thomas amd J. Gudnason and P.A. Naylor. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 20 (1):82–97, 2012.

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

[4] A. Krishnamurthy and. Two channel speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34 (4):730–743, 1986.

[5] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1):29 –40, jan 1988. ISSN 0096-3518. doi: 10.1109/29.1486.

[6] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, multifractales et turbulence*. Diderot Editeur, Paris, France, 1995.

[7] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J.F. Muzy. *Ondelettes, Multifractales et Turbulences: de l'ADN aux Croissances Cristallines*. Diderot Editeur, Arts et Sciences, Paris, 1995.

[8] A. Arneodo, C. Baudet, F. Belin, R. Benzi, B. Castaing, B. Chabaud, R. Chavarria, S. Ciliberto, R. Camussi, F. Chillà, B. Dubrulle, Y. Gagne, B. Hebral, J. Herweijer, M. Marchand, J. Maurer, J. F. Muzy, A. Naert, A. Noullez, J. Peinke, F. Roux, P. Tabeling, W. van de Water, and H. Willaime. Structure functions in turbulence, in various flow configurations, at reynolds number between 30 and 5000, using extended self-similarity. *EPL (Europhysics Letters)*, 34(6):411, 1996.

[9] A. Arneodo, E. Bacry, S. Jaffard, and J. F. Muzy. Oscillating singularities on cantor sets: A grand-canonical multifractal formalism. *Journal of Statistical Physics*, 87(1-2):179–209, 1997. doi: 110.1007/BF02181485.

[10] A. Arneodo, E. Bacry, and J. F. Muzy. *The thermodynamics of fractals revisited with wavelets*. Cambridge University Press, 1999.

[11] A. Arneodo, B. Audit, P. Kestener, and S. Roux. Pwavelet-based multifractal analysis. *Scholarpedia*, 3(3):4103, 2008.

[12] M.M. Artimy, W. Robertson, and W.J. Phillips. Automatic detection of acoustic sub-word boundaries for single digit recognition. In *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, volume 2, pages 751–754 vol.2, 1999. doi: 10.1109/CCECE.1999.808034.

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

[14] B. S. Atal and J. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1982.

[15] Hicham Badri. Computer graphics effects from the framework of reconstructible systems. Master's thesis, Rabat faculty of science-INRIA Bordeaux Sud-Ouest, 2012.

[16] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *Speech and Audio Processing, IEEE Transactions on*, 7:1 – 17, Jan 1999.

[17] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani. Predictability: a way to characterize complexity. *Physics Reports*, 356(6):367–474, 2002. doi:10.1016/S0370-1573(01)00025-4.

[18] Aicha Bouzid and Noureddine Ellouze. Glottal opening instant detection from speech signal. In *EUropean Signal Processing COnference (EUSIPCO)*, 2004.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[20] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA Voice Quality: Functions, Analysis and Synthesis*, 2003.

[21] E. J. Candès and J. Romberg. l1-magic : Recovery of sparse signals via convex programming. 2005.

[22] E. J. Candès and M. B. Wakin. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

[23] G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro. *Nonlinear Speech Modeling and Applications: Advanced Lectures and Revised Selected Papers*. Lecture notes in computer science: Tutorial. Springer, 2005. ISBN 9783540274414.

[24] WAI C. CHU. *Speech coding algorithms: foundation and evolution of standardized coders*. Wiley-Interscience, 2003.

[25] C. d'Alessandro and N. Sturmel. Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36:601–622, 2011. ISSN 0256-2499. doi: 10.1007/s12046-011-0040-6.

[26] KED TIMIT database. [Online], http://festvox.org/.

[27] Etienne Denoel and Jean Philippe Solvay. Linear prediction of speech with a least absolute error criterion. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 33:1397–1403, 1985.

[28] T. drugman. Gloat toolbox. [Online], http://tcts.fpms.ac.be/ drugman/.

[29] T. Drugman. *Advances in Glottal Analysis and its Applications*. PhD thesis, University of Mons, 2011.

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

[31] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Interspeech conference*, 2010.

[32] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994 –1006, march 2012. ISSN 1558-7916. doi: 10.1109/TASL.2011.2170835.

[33] S. Dusan and L. Rabiner. On the relation between maximum spectral transition positions and phone boundaries. *Proceddings of INTERSPEECH/ICSLP 2006*, pages 645–648, 2006.

[34] W.T Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet. *Statistical methods in experimental physics*. North-Holland Pub. Co., 1971. ISBN 0-444-10117-9.

[35] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In Chollet G. et al et al, editor, *Lecture Notes in Computer Science: Nonlinear Speech Modeling*, pages 261–290. Springer Verlag, 2005.

[36] A. Esposito and M. Marinaro. Nonlinear speech modeling and applications, chapter: Some notes on nonlinearities of speech. pages 1–14. Springer-Verlag, 2005. ISBN 3-540-27441-3, 978-3-540-27441-4.

[37] T. Ewender and B. Pfister. Accurate pitch marking for prosodic modification of speech segments. In *Proceedings of INTERSPEECH*, 2010.

[38] M. Faundez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, and J. Schoentgen. Nonlinear speech processing: Overview and applications. *Control and intelligent systems*, 30:1–10, 2002.

[39] H. G. Feichtinger and K. Grochenig. *Theory and practice of irregular sampling*, chapter Wavelets Mathematics and Applications. CRC-Press, 1993.

[40] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg. Segment based variable frame rate speech analysis and recognition using a spectral variation function. In *ICSLP*. ISCA, 1992.

[41] C. Foias, G. C. Rota, O. Manley, R. Rosa, and R. Temam. *Navier-Stokes equations and turbulence*. Cambridge University Press, Cambridge, 2001.

[42] U. Frisch. *Turbulence: The legacy of A.N. Kolmogorov*. Cambridge University Press, 1995.

[43] U. Frisch and G. Parisi. On the singularity structure of fully developed turbulence. In M. Gil, R. Benzi, and G. Parisi, editors, *Turbulence and Predictability in Geophysical Fluid Dynamics and ClimateDynamics*, pages 84–88. Elsevier, 1985.

[44] G726. G726 recommendation: 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation, international telecommunication union, 1990.

[45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical report, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[46] N.D. Gaubitch and P.A. Naylor. Spatiotemporal averagingmethod for enhancement of reverberant speech. In *15th International IEEE Conference on Digital Signal Processing*, 2007.

[47] D. Giacobello. *Sparsity in Linear Predictive Coding of Speech*. PhD thesis, Multimedia Information and Signal Processing, Department of Electronic Systems, Aalborg University, 2010.

[48] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. Sparse linear predictors for speech processing. In *Proceedings of the INTERSPEECH*, 2009.

[49] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen. Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[50] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen. Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction. *IEEE Signal Processing Letters*, 17, 2010.

[51] D. Giacobello, M. G. Christensen, M. N. Murth, Søren Holdt Jensen, and Fello Marc Moonen. Sparse linear prediction and its applications to speech processing. *IEEE Transactions on Audio, Speech and Language Processing*, 20:1644–1657, 2012.

[52] J. Grazzini, A. Turiel, H. Yahia, and I. Herlin. Edge-preserving smoothing of high-resolution images with a partial multifractal reconstruction scheme. In *International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2004.

[53] J. Grazzini, A. Turiel, and H. Yahia. Multifractal Formalism for Remote Sensing: A Contribution to the Description and the Understanding of Meteorological Phenomena in Satellite Images. In Miroslav M. Novak, editor, *Complexus Mundi. Emergent Patterns in Nature*, pages 247–256. World Scientific Publishing Co. Pte. Ltd., 2006. ISBN 981-256-666-X.

[54] J. Holmes. Formant excitation before and after glottal closure. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 39–42, apr 1976. doi: 10.1109/ICASSP.1976.1170095.

[55] K. Hu, P.C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley. Effect of trends on detrended fluctuation analysis. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64(1 Pt 1), July 2001.

[56] Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio Speech Language Processing*, 16:229 – 238, 2008.

[57] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55:4723–4740, 2009.

[58] I. Ihm and B. Naylor. *Scientific visualization of physical phenomena*, chapter 32: Piecewise linear approximations of digitized space curves with applications. Springer-Verlag New York, 1991.

[59] ITU-T Recommendation P.862 : Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

[60] N.S Jayant. Digital coding of speech waveforms: Pcm, dpcm, and dm quantizers. *Proceedings of the IEEE*, 62, Issue 5:611 – 632, 1974.

[61] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, pages 358–386. The Denver Center for the Performing Arts, 1983.

[62] Y. Kakihara. *Abstract methods in information theory*. Multivariate Analysis. World Scientific, Singapore, 1999.

[63] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge nonlinear science series. Cambridge University Press, 2003. ISBN 9780521529020.

[64] Vahid Khanagha, Hussein Yahia, Khalid Daoudi, Oriol Pont, and Antonio Turiel. Reconstruction of speech signals from their unpredictable points manifold. In *NOLISP*, Lecture Notes in Computer Science. Springer, 2012.

[65] T. Koizumi, S. Taniguchi, and S. Hiromitsu. Glottal source - vocal tract interaction. *Journal of Acoustic Society of America*, 78 (5):1541–1547, 1985.

[66] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6):1098–1109, Jan. 2005.

[67] P. Kroon, E. Deprettere, and R. Sluyter. Regular-pulse excitation–a novel approach to effective and efficient multipulse coding of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 34(5):1054 – 1063, oct 1986. ISSN 0096-3518. doi: 10.1109/TASSP.1986.1164946.

[68] Gernot Kubin. *Speech coding and synthesis*, chapter Chapter 16: Nonlinear processing of speech, pages 557–610. Elsevier, 1995.

[69] Ta-Hsin Li and J.D. Gibson. Speech analysis and segmentation by parametric filtering. *IEEE Transactions on Speech and Audio Processing*, 4(3):203 –213, may 1996. ISSN 1063-6676. doi: 10.1109/89.496216.

[70] Y. Lin, Y. Wang, and Yuan-Fu Liao. Phone boundary detection using sample-based acoustic parameters. *Proceedings of INTERSPEECH*, 2010.

[71] M. A. Little. Mathematical foundations of nonlinear, non-gaussian, and time-varying digital speech signal processing. In *NOLISP*, pages 9–16, 2011.

[72] M. A. Little and N.S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals: Part i – background theory. *Proceedings of the Royal Society A*, pages 3088 – 3114, 2011. doi: 10.1098/rspa.2010.0671.

[73] M. A. Little and N.S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals: Part ii – new methods. *Proceedings of the Royal Society A*, pages 3088 – 3114, 2011. doi: 10.1098/rspa.2010.0674.

[74] M. A. Little, P. McSharry, I. Moroz, and S. Roberts. Nonlinear, biophysically-informed speech pathology detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, page II, may 2006. doi: 10.1109/ICASSP.2006.1660534.

[75] M. A. Little, P. McSharry, S. Roberts, D. Costello, and I. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, 2007. ISSN 1475-925X. doi: 10.1186/1475-925X-6-23.

[76] M. A. Little, P. E McSharry, I.M. Moroz, and S.J. Roberts. Testing the assumptions of linear prediction analysis in normal vowels. *Journal of the Acoustical Society of America*, 119:549–558, January 2006. doi: 10.1121/1.2141266.

[77] M.A. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, Oxford University, 2007.

[78] P. Machač and R. Skarnitzl. *Principles of Phonetic Segmentation*. Edition erudica. Epocha, 2009. ISBN 9788074250323.

[79] S. K. Maji, O.l Pont, H. Yahia, and J. Sudre. Inferring information across scales in acquired complex signals. In *European Conference on Complex Systems (ECCS)*, 2012.

[80] S. K. Maji, H. M. Yahia, O. Pont, J. Sudre, T. Fusco, and V. Michau. Towards multiscale reconstruction of perturbated phase from hartmann-shack acquisitions. In *AHS*, pages 77–84, 2012.

[81] S. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier Science, 1999. ISBN 9780124666061.

[82] S. Mallat and W. Liang Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38:617–643, 1992.

[83] P. Maragos and A. Potamianos. Fractal dimensions of speech sounds: Computation and application to automatic speech recognition. *Journal of Acoustic Society of America*, 105:1925–1932, March 1999.

[84] S. McLaughlin and P. Maragos. *Nonlinear methods for speech analysis and synthesis, in Advances in Nonlinear Signal and Image Processing*. Hindawi Publ. Corp., 2006.

[85] PE. McSharry, LA. Smith, and L. Tarassenkeo. Prediction of epileptic seizures: are nonlinear methods relevant? *Nature Medicine*, 9(3):241–242, 2009.

[86] D. Meng, Q. Zhao, and Z. Xu. Improved robustness of sparse pca by l1-norm maximization. *Pattern Recognition Elsevier*, 45:487–497, 2012.

[87] C.D. Mitchell, M.P. Harper, and L.H. Jamieson. Using explicit segmentation to improve hmm phone recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 229–232 vol.1, may 1995. doi: 10.1109/ICASSP.1995.479406.

[88] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 (5-6):453 – 467, 1990. ISSN 0167-6393. doi: 10.1016/0167-6393(90)90021-Z.

[89] K.S.R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:1602–1613, 2008.

[90] J. F. Muzy, E. Bacry, and A. Arneodo. Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev. E*, 47:875–884, Feb 1993. doi: 10.1103/PhysRevE.47.875.

[91] J.F. Muzy, E. Bacry, and A. Arneodo. Wavelets and multifractal formalism for singular signals: Application to turbulence data. *Physical Review Letters*, 67: 3515–3518, Dec 1991. doi: 10.1103/PhysRevLett.67.3515. URL http://link.aps.org/doi/10.1103/PhysRevLett.67.3515.

[92] P. A. Naylor. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15 (1):34–43, 2007.

[93] Noisex-92. [Online], www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[94] E. N. Pinson. Pitch synchronous time domain estimation of formant frequencies and bandwidths. *Journal of the Acoustical Society of America*, 35 (8):1264–1273, 1963.

[95] V. Pitsikalis and P. Maragos. Analysis and classification of speech signals by generalized fractal dimension features. *Speech Communication*, 51, Issue 12: 1206–1223, December 2009.

[96] M. D. Plumpe. Modeling of the glottal flow derivative waveform with application to speaker identification. Master's thesis, Massachusetts Institute of Technology, 1997.

[97] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Description, modeling and forecasting of data with optimal wavelets. *Journal of Economic Interaction and Coordination*, 4(1), June 2009. ISSN 1860-711X (Print) 1860-7128 (Online). doi: 10.1007/s11403-009-0046-x.

[98] O. Pont, A. Turiel, and C.J. Perez-Vicente. Empirical evidences of a common multifractal signature in economic, biological and physical systems. *Physica A*, 388(10):2025–2035, May 2009.

[99] J. G. Proakis and D. K. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, 1995.

[100] T. F. Quatieri. *Discret-time speech signal processing principles and practice*. Prentice-Hall, 2001.

[101] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[102] O. J. Rasanen, U. K. Laine, and T. Altosaar. An improved speech segmentation quality measure: the R-value. *Proceedings of INTERSPEECH*, 2009.

[103] O. J Räsänen, U. K Laine, and T. Altosaar. Blind segmentation of speech using non-linear filtering methods. In Ivo Ipsic, editor, *Speech Technologies*. InTech, 2011. doi: 10.5772/16433.

[104] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994. doi: 10.1103/PhysRevLett.73.814.

[105] K. Sauer and J Allebach. Iterative reconstruction of bandlimited images from nonuniformly spaced samples. *IEEE Transactions on Circuits and Systems*, 34 Issue:12:1497 – 1506, Dec 1987.

[106] S. C. Shadden. Lagrangian Coherent Structures Tutorial[Online], http://mmae.iit.edu/shadden/LCS-tutorial/, 2009.

[107] M. Sharma and R. J. Mammone. "blind" speech segmentation: Automatic segmentation of speech without linguistic knowledge. In *Spoken Language Processing, 4th International Conference on*, pages 1237–1240. ISCA, October 1996.

[108] Z. She and E. Leveque. Universal scaling laws in fully developed turbulence. *Physical Review Letters*, 72(3):336–339, January 1994.

[109] S. Singhal and B.S. Atal. Amplitude optimization and pitch prediction in multipulse coders. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:317 – 327, 1989.

[110] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 3 (5):325–333, 1995.

[111] J. Sole-Casals and V. Zaiats. *Advances in Nonlinear Speech Processing: International Conference on Nonlinear Speech Processing, NOLISP 2009, Vic, Spain, June 25-27, 2009, Revised Selected Papers*. Lecture Notes in Artificial Intelligence. Springer, 2010. ISBN 9783642115080.

[112] CMU ARCTIC speech synthesis databases. [Online], http://festvox.org/cmu_arctic.

[113] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomena (International Series of Monographs on Physics)*. Oxford University Press, USA, July 1987. ISBN 0195053168.

[114] K. Steiglitz and B. Dickinson. The use of time-domain selection for improved linear prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 25 (1):34–39, 1977.

[115] B. H. Story. An overview of the physiology, physics, and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

[116] N. Sturmel, C. d'Alessandro, and F. Rigaud. Glottal closure instant detection using lines of maximum amplitudes (loma) of thewavelet transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4517–4520, april 2009. doi: 10.1109/ICASSP.2009.4960634.

[117] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9 (1):21–29, 2001.

[118] Y. Stylianou, M. Faundez-Zanuy, and A. Eposito. *Progress in Nonlinear Speech Processing*. Lecture Notes in Computer Science. Springer, 2007. ISBN 9783540715030.

[119] J. Tailleur. *Grandes déviations, physique statistique et systèmes dynamiques*. PhD thesis, Universitè Pierre et Marie Curie - Paris VI, 2007.

[120] T. Takagi and M. SUGENO M. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, 15:116–132, 1985.

[121] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*. NATO Advanced Study Institute Series D, 1989.

[122] I. R. Titze. *The Myoelastic Aerodynamic Theory of Phonation*. The national center for voice & speech, 2007.

[123] D. Torre-Toledano, L.A. Hernandez-Gomez, and L. Villarrubia-Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625, 2003.

[124] C.M. Travieso-González and J. Alonso-Hernández. *Advances in Nonlinear Speech Processing: International Conference on Nonlinear Speech Processing, NoLISP 2011, Las Palmas de Gran Canaria, Spain, November 7-9, 2011, Proceedings*. Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence. Springer, 2011. ISBN 9783642250194.

[125] V.N. Tuan and C. d'Alessandro. Robust glottal closure detection using the wavelet transform. In *in Proceedings of the European Conference on Speech Technology*, pages 2805–2808, 1999.

[126] A. Turiel. Method and system for the singularity analysis of digital signals, patent registered under number pct/es2008/070195, 2008.

[127] A. Turiel and A. del Pozo. Reconstructing images from their most singular fractal manifold. *IEEE Transactions on Image Processing*, 11:345–350, 2002.

[128] A. Turiel and N. Parga. The multi-fractal structure of contrast changes in natural images: from sharp edges to textures. *Neural Computation*, 12:763–793, 2000.

[129] A. Turiel and C. Pérez-Vicente. Multifractal measures: definition, description, synthesis and analysis. a detailed study. In J.-P. Nadal, A. Turiel, and H. Yahia, editors, *Proceedings of the "Journées d'étude sur les méthodes pour les signaux complexes en traitement d'image"*, pages 41–57, Rocquencourt, 2004. INRIA.

[130] A. Turiel, C.J. Pérez-Vicente, and J. Grazzini. Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study. *Journal of Computational Physics, Volume 216, Issue 1, p. 362-390.*, 216:362–390, 2006.

[131] A. Turiel, H. Yahia, and C.J Pérez-Vicente. Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis. *Journal of Physics A: Mathematical and Theoretical*, 41:015501, 2008.

[132] D. Wong, J. Markel, and A. Jr. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):350–355, aug 1979. ISSN 0096-3518. doi: 10.1109/TASSP.1979.1163260.

[133] H. Yahia, J. Sudre, C. Pottier, and V. Garçon. Motion analysis in oceanographic satellite images using multiscale methods and the energy cascade. *Journal of Pattern Recognition*, 2010, to appear. doi:10.1016/j.patcog.2010.04.011.

[134] H. Yahia, J. Sudre, V. Garçon, and C. Pottier. High-resolution ocean dynamics from microcanonical formulations in non linear complex signal analysis. In *AGU FALL MEETING*, San Francisco, États-Unis, December 2011. American Geophysical Union.