Exploiting regularities in natural acoustical scenes for monaural audio signal estimation, decomposition, restoration and modification

音環境に内在する規則性に基づくモノラル音響信号の推定・分解・復元・加工に関する研究

–

Exploitation de régularités dans les scènes acoustiques naturelles pour l'estimation, la décomposition, la restauration et la modification de signaux audio monocanal

Jonathan Le Roux

ルルー ジョナトン

THE UNIVERSITY OF TOKYO
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
DEPARTMENT OF INFORMATION PHYSICS AND COMPUTING
東京大学 大学院情報理工学系研究科
システム情報学専攻

Ph.D. Thesis
博士論文

submitted by
Jonathan LE ROUX

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Science and Technology

Exploiting regularities in natural acoustical scenes for monaural audio signal estimation, decomposition, restoration and modification

(音環境に内在する規則性に基づくモノラル音響信号の推定・分解・復元・加工に関する研究)

defended on January 29, 2009
in front of the committee composed of

Shigeki SAGAYAMA	University of Tokyo	Thesis Supervisor
Alain de CHEVEIGNÉ	École Normale Supérieure	Co-supervisor
Shigeru ANDÔ	University of Tokyo	Examiner
Keikichi HIROSE	University of Tokyo	Examiner
Nobutaka ONO	University of Tokyo	Examiner
Susumu TACHI	University of Tokyo	Examiner

UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE
ÉCOLE DOCTORALE EDITE

THÈSE DE DOCTORAT

spécialité
INFORMATIQUE, TÉLÉCOMMUNICATIONS ET ÉLECTRONIQUE

présentée par
Jonathan LE ROUX

pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE

Exploitation de régularités dans les scènes acoustiques naturelles pour l'estimation, la décomposition, la restauration et la modification de signaux audio monocanal

(Exploiting regularities in natural acoustical scenes for monaural audio signal estimation, decomposition, restoration and modification)

soutenue le 12 mars 2009
devant le jury composé de

Alain de CHEVEIGNÉ	École Normale Supérieure	Directeur de Thèse
Shigeki SAGAYAMA	University of Tokyo	Co-Directeur
Dan P. W. ELLIS	Columbia University	Rapporteur
Mark D. PLUMBLEY	Queen Mary University	Rapporteur
Laurent DAUDET	Université Paris VI	Examinateur
Gaël RICHARD	TELECOM ParisTech	Examinateur
Emmanuel VINCENT	INRIA - IRISA	Examinateur
Jean-Luc ZARADER	Université Paris VI	Examinateur

Sagayama/Ono Laboratory
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
113-8656 Tokyo (Japan)

and

Laboratoire de Psychologie de la Perception (UMR 8158)
Équipe Audition
École Normale Supérieure
29, rue d'Ulm
75005 Paris (France)

Acknowledgements

The long and winding road that leads to the doors of the Ph.D. sometimes disappeared, and I am extremely grateful both to the many people who let me know the way, and to the not less many who often helped me temporarily forget about it altogether.

I would first like to thank my thesis advisors, Professor Shigeki Sagayama at the University of Tokyo and Dr. Alain de Cheveigné at the École Normale Supérieure. Professor Sagayama has taught me many things, from technical topics to the way to conduct and present one's research, and continuously encouraged and helped me to participate in national and international conferences, introducing me to many researchers and enabling me to widen my research horizon. Most importantly, he has created and constantly fueled with ideas and energy the wonderful research environment that constitutes the Sagayama/Ono Laboratory in which I spent most of the past four years. Alain de Cheveigné has been extremely present and dedicated from my very first steps in audio engineering five years ago, and helped and advised me on a daily basis on each and every aspect of the Ph.D. process, which is all the more remarkable knowing that we spent most of that time about 10000 km away and with a 7 to 9 hour time difference from each other. His careful supervision was definitely the key element which made me keep sight of what had to be done and when it had to be done.

I would like to thank as well sincerely Professor Nobutaka Ono, whom I truly consider as my (unofficial) third thesis advisor. Professor Ono has both an extremely deep understanding of all the technical details involved in my work and a very clear view of the broader picture. The countless discussions we had were the roots to many of the ideas exposed in this thesis. Watching him explain utterly abstract concepts very simply by drawing figures in the air with his hands certainly counts among the most scientifically enjoyable and exciting moments of my Ph.D.

I am very grateful to the members of my thesis committees (yes, with an "s", as there were two) to have accepted to take some of their very precious time to review my work and attend the defenses (yes again, as there were three): Professor Shigeru Andô, Professor Keikichi Hirose and Professor Susumu Tachi, of the University of Tokyo, for the defenses in Tokyo, Professor Dan P. W. Ellis, of Columbia University, Professor Mark D. Plumbley, of Queen Mary University of London, Dr. Laurent Daudet, of Université Paris 6, Professor

Gaël Richard, of Telecom ParisTech, Dr. Emmanuel Vincent, of INRIA-IRISA, and Professor Jean-Luc Zarader, of Université Paris 6, for the defense in Paris. It is an immense honor for me to have such a gathering of renowned scientists as my committee(s). I am especially thankful to Dan and Mark for accepting the most important and time-demanding roles of thesis Readers, and I am happy to be able to present my work under their auspices. I would also like to thank specially Emmanuel for the many comments he gave me on early versions of some key parts of the thesis during his stay in our lab at the University of Tokyo. His presence in the committee has a particular significance as Emmanuel and I were actually classmates ten years ago in our first year in the math department of the École Normale Supérieure. Obviously he was much faster than me in both starting and finishing his Ph.D.

I would not be anywhere close to finishing this work or even getting it started if it was not for Dr. Hirokazu Kameoka, my former colleague at the Sagayama/Ono Lab, two years above in the Doctoral program. I have been blessed with thesis advisors all along, and he can definitely be counted as one of them. He is an incredibly brilliant young researcher coming up with groundbreaking ideas on a daily basis, and I am convinced that he will become one of the most prominent faces worldwide in the field of audio signal processing in the near future. My work on HTC in Chapters 3 and 4 is a direct collaboration with him, and we spent days and nights discussing on most of the other problems investigated in this thesis. He is now one of my best friends in Japan (and the other half of our Manzai duo), and I am looking forward to being his colleague once again at NTT CS Labs during my post-doc.

I would also like to sincerely thank Professor Lucas C. Parra, of City College New-York, who completes the gallery of researchers with whom I have had the chance to collaborate directly during the course of this Ph.D. Working literally day and night for a week with him and Alain de Cheveigné on three time lags (New-York, Paris, Tokyo) during the preparation of our NIPS paper was one of the most exciting and fun parts of my Ph.D. I have learnt a lot from his way of approaching and solving problems, and really enjoyed the many memorable Skype sessions we had.

My first encounter with audio signal processing and machine learning took place during an internship at NTT CS Labs near Kyoto in Winter 2004, under the supervision of Dr. Erik McDermott. I would like to thank him for introducing me to the field and teaching me most of the basics, and for all the great discussions we have had since then (many of them in the nice environment of an izakaya). I now consider Erik as my "older brother" in Japan, and knowing that I would be meeting him is a significant part of the motivation to attend a conference. I would also like to thank the members and former members of NTT's Signal Processing Research Group and Media Recognition Research Group with whom I have had

the pleasure to interact during the internship and since then at many occasions: Dr. Shinji Watanabe, Dr. Tomohiro Nakatani, Dr. Atsushi Nakamura, Keisuke Kinoshita, Takanobu Ôba, Takuya Yoshioka, Dr. Kunio Kashino, Dr. Shôko Araki, Dr. Hiroshi Sawada, Dr. Shôji Makino, Dr. Yasuhiro Minami, Dr. Takaaki Hori, Dr. Kentarô Ishizuka, Dr. Masakiyo Fujimoto, Dr. Michael Schuster (now with Google) and Dr. Parham Zolfaghari (now with ClientKnowledge).

I would like to express my gratitude to Dr. Frank K. Soong for inviting me to Microsoft Research Asia in Beijing for an internship in 2006, and for managing to find some time in his very busy schedule to supervise my work there. I would also like to thank my fellow interns in the Speech Group for all the pleasant discussions we had and the nice restaurants we went to. I am looking forward to meeting them again in future conferences (though my Chinese skills definitely need a recovery period). I am also indebted to Dr. Philippe de Reffye, of INRIA, who supervised my internship at the Sino-French joint laboratory LIAMA in Beijing in 2001, for revealing to me, among many other things, the real good reason to do a Ph.D.: enjoy student life for a few more years.

I am very grateful to Professor Hiroshi Matano, of the University of Tokyo, for taking care of me during my first year in Japan and for introducing me to Japanese culture, art and language. It is also through him that I got the chance to meet Professor Henri Berestycki, of EHESS, whom I would like to thank for giving me critical pieces of advice on my orientation. I am also indebted to Professor Yôichi Tôkura, of the National Institute of Informatics, for introducing me to Professor Sagayama.

I would like to thank the members of the Educational Affairs Section of the Information Physics department at the University of Tokyo, for their constant help to go through all the administrative matters that an international student doubled with a Ph.D. student inevitably has to face. I am also extremely grateful to Mrs. Reiko Kaneko, the secretary of our lab, and Miss Naoko Ozeki, former secretary, for their efficiency and kindness beyond limit. There has been no administrative difficulty they could not solve and no formality that their never fading smile could not change (sometimes literally) into a piece of cake. I am equally indebted to Mr. Takuya Nishimoto and Dr. Shinji Sakô (now with the Nagoya Institute of Technology) for all the technical assistance and advice they provided me with as well as their efforts to make our lab a better work environment.

The Sagayama/Ono lab has been an incredible place to spend these last four years, and one of the main reasons was of course the quality of the students around me, both on human and academic levels. Many of them have become close friends, and I have to admit that spending the night in the lab was actually a real pleasure knowing they would be around. I

would like to thank in particular Yôsuke Izumi and Shô'ichirô Saitô (now with NTT), who both spent a significant amount of time proof-reading emails and texts I wrote in Japanese, Kenta Nishiki, Emiru Tsunoo and Satoru Fukayama, with whom I had the pleasure to share many dinners and many laughs in the tough period of the last year of the Ph.D., and Ryô Yamamoto, Ken'ichi Miyamoto (now with NEC) and Kyôsuke Matsumoto (now with Sony), whose motivation and positivity with regards to research has been very inspirational. I would also like to thank former members of the lab, with whom I did not overlap while in the lab but who became good friends through the Old Boy/Old Girl (well, Old Boy...) network, Kazuhito Inoue (now with Korg), Shôhei Nakagata (now with Fujitsu) and Yutaka Kamamoto (now with NTT).

Although I spent most of the past four years in Tokyo, I have made a few stays in Paris to visit Alain, in the Équipe Audition of the Laboratoire de Psychologie de la Perception of the Université Paris 5. I am very grateful to the director of the lab Kevin O'Reagan and to all the members of the team, especially Daniel Pressnitzer and Christian Lorenzi, for welcoming me each time so warm-heartedly, even though I would only be spending a few weeks there every year. Christian's lunch rallying cry "Allons manger gras!" ("Let's get some fat food!") always felt particularly appealing to my heart of French expatriate in Japan.

Through all the conferences and meetings I have attended, I came to believe that research is mostly about meeting people and sharing (knowledge, code, publications, or beer). I have had the chance to meet many great researchers during the course of this Ph.D., and I would like to thank them for all the comments, pieces of advice, insights, encouraging words, or simply the time, that they shared with me. To name but a few, I thank Dr. Masataka Gotô, Hiromasa Fujihara and Dr. Tomoyasu Nakano of AIST, Dr. Yasunori Ôishi, of Nagoya University and a future colleague at NTT, Dr. Tetsurô Kitahara of Kwansei Gakuin University, Charles Cadieu and Pierre Garrigues of UC Berkeley, Dr. T. J. Hazen of the MIT Lincoln Laboratory, Dr. Jasha Droppo and Dr. Mike Seltzer of Microsoft Research, Professor Shihab Shamma and Professor Jonathan Z. Simon of the University of Maryland, Professor Simon J. Godsill and Professor Ali Taylan Cemgil of Cambridge University, and Professor Richard Stern and Professor Bhiksha Raj of Carnegie Mellon University. I am particularly indebted to Professor Chin-Hui Lee, of Georgia Tech, and to Dr. Paris Smaragdis, of Adobe Systems, for their many insightful comments during the preparation of my defense.

Finally, I would like to thank all the friends that have made this long journey a pleasant one. Moving around makes it easy to make new friends, but hard to keep the old ones.

I am all the more happy to know that I can count on many of them to gather each time I go back to Paris: my high-school mates, Thierry, François and Énée; the lights of my prépa

years, Bastien and Simon; the fabulous “Gaga-team” from ENS, Gaga, Valeria, Vincent, Anke, Ben, Marie, Jboy, Totodudu, Titoune, Tomtom, Kévin, Smice, Groug, Mathilde, Carole and Thierry M.; my globe-trotter medical doctor friends, Frédéric and Bich-Tram; and my Beijing friends, Sebastian, Antoine, Jean-Aimé, Hélène, Léonore and Pauki. I would like to thank Vincent Plagnol in particular for the constant contact and countless discussions we have had over the years both on technical and personal matters, in spite of the physical distance. I guess I should also thank the Internet for its cooperation there.

I am also grateful to all the friends I met while studying in Beijing. They have showed me how fun and rewarding it is to study abroad, to learn foreign languages and discover foreign cultures. I am especially thankful to the wonderful Japanese friends I first met there, Masato, Sayuri, Minako and Oka, among many others, as they are the reason why I decided to visit Japan in the first place, and to those Beijing has made me meet since, Que, Hayato, Ma Jr., to cite but a few.

Living in Tokyo would not be nearly as fun if I hadn’t had the amazing luck to meet the incredible people who have become my friends here. I am really grateful to the members of my volley-ball circle バレパ and my band circle POMP to have enabled me to experience student life *in the Japanese style*. My colloquial Japanese definitely owes them more than a lot. I would also like to thank Alvaro, Arthur, Arnaud, Fred, Nicolas and Rolf, with whom I have enjoyed so much hanging out. I am particularly grateful to my “little brother” in Japan, Paul Masurel, for being such a cool pop guy. Another important part of my non-academic life in Tokyo has been centered around my band The Empty Bed. I thank sincerely its members Gucchi, aka さわ口まさひロ, Tack-G and Hiromasa for all the fun we have had together and for enabling me to get a glimpse of what it feels like to be creative. For all the amazing friends I met through him and all that he taught me, I really consider my encounter with Gucchi as a revolution in my life in Japan. I hope we can pick up where we left off when I had to focus on the writing of this thesis.

I would finally like to thank Noriko, for being patient with a 29-year old student, my parents, Roselyne and Jean-Pierre, and my sisters, Tiphaine and Pénélope, for their love and care and for enduring a son and little brother often far away. I’ll try to make it up to them.

Abstract

A crucial problem for many audio engineering applications is that most, if not all, real world situations they face are adverse ones, with strong non-stationary background noises, concurrent sources, brief interruptions due to glitches or missing packets, etc. Humans however are able to achieve a great robustness in their perception and understanding of the complex acoustical world that surrounds them, relying on statistical regularities in the original sources and the incoming stimuli.

The goal of this thesis is to propose a statistical approach to the analysis of such natural acoustical scenes, based on models of the regularities of the acoustical environment. Such an approach involves solving mainly three subproblems: inference of what is happening in an acoustical scene as the best explanation of the distorted, mixed, and incomplete observations given models of the environment; reconstruction of incomplete observations based on these models; acquisition of these models from the data. We tackle all of these problems following a common strategy which systematically focuses on a general mathematical formulation based on an objective function, so that the various tasks can be effectively solved as well-posed constrained optimization problems.

We first present a statistical model in the time-frequency power domain for the analysis of acoustical scenes involving speech, with applications such as F_0 estimation and source separation, and explain how to extend such models to perform scene analysis with incomplete data, and how to recover the missing data. Noting the importance of avoiding some inherent limitations of the power domain and of actively making use of the phase information, we successively introduce a general consistency criterion in the complex time-frequency domain, and develop an adaptive template matching framework in the time domain. Finally, we investigate the hypothesis of a data-driven adaptation of the peripheral filters to efficiently extract the modulation structure of the signals.

Keywords: acoustical scene analysis, audio signal processing, signal regularities, probabilistic source models, statistical machine learning, optimization methods, model-based analysis, fundamental frequency estimation, source separation, signal decomposition.

概要

音声アプリケーションや通信システムなどの音を媒体としたあらゆるアプリケーションを実現する上で、非定常で強い背景雑音、複数音源の同時観測、あるいは機器の誤作動や欠損パケット等の影響による通信の中断など、我々が身近に直面しうる複雑な現象を扱うことが重要な課題となる。しかし我々人間は周囲の複雑な音響情景を、信号源の知識や所与の観測信号の統計的規則性に基づいて頑健に認識し理解することができる。

本論文の目的は、音環境に内在する規則性のモデルに基づき、かくのごとく複雑な観測信号から音響情景を解析するための統計的アプローチを確立することである。本アプローチでは主として、（1）ある音環境モデルのもとで、何らかの原因で歪み、複数の音が混在し、あるいは部分的に欠損したような観測信号データに対し、どのような音響的事象が生じているかについての最も自然な解釈を与える問題、（2）上記モデルに基づき欠損した情報を復元する問題、（3）上記モデルを観測データから自律的に獲得する問題を扱う。これらすべての問題に対し、数理的定式化を通して目的関数を立てて解決を図ることを基本戦略とし、これにより様々な問題を拘束つき最適化問題として見通しよく解決することができる。

第2章で関連研究をサーベイした後、第3章ではまず音響情景解析を目的とした時間周波数パワー領域における統計的モデルを導入する。亀岡により導入された調波時間構造化クラスタリング（HTC）法を、音声のように F_0 が滑らかに時間変化する信号を扱えるように拡張する。時間周波数領域でモデル化することにより、人間の聴覚機能の一つとして知られる音脈分凝プロセスにおいて関与しているとされる音声の規則性を拘束条件として導入することができ，これをもとにして拘束付き混合正規分布で表されるパラメトリックな音声信号モデルを立てることができる。滑らかな F_0 軌跡を3次元スプラインにより表現することで、従来の HTC 法の最適化アルゴリズムの効率性を保持しつつ、音声信号に対応できるように拡張することが本章での目的である。また、混合正規分布に基づく背景雑音モデルを導入し、雑音と音声が混合する音響情景のモデル化を検討する。第4章では、評価実験により、音声信号のクリーン・雑音中・複数話者の F_0 推定や音声強調・音源分離などの音響情景解析の様々な課題において提案手法が従来法より優れていたことを示す。

第5章では、不完全データから情景解析を行うためにこのモデルを拡張する方法、欠損データを復元する方法を説明する。一般的な意味で、何らかのモデルをデータにフィッティングさせる問題を、観測データが一部欠損したいわゆる不完全データを対象とした場合にも

扱えるような枠組を補助関数法に基づいて導く。ここで、フィッティングの良さを測る尺度が Bregman ダイバージェンスと呼ぶクラスに属する場合には、補助関数法に基づく上記の枠組が EM（Expectation-Maximization）アルゴリズムによって説明できることを明らかにする。評価実験を通し、提案手法により不完全データからの情景解析と欠損データの復元を同時に行えることを確認した。

続いて、パワースペクトル領域処理の限界を超えること、また、位相情報を有効利用することの重要性に着目し、複素時間周波数領域における位相の無矛盾性規準を第６章で導き、時間領域の適応的テンプレートマッチングの枠組を第７章で構築する。第５章までは位相情報を用いずにパワースペクトル領域で処理を行うが、それによるいくつかの課題を説明する。パワースペクトルから信号の再合成が必要な場合は失われた位相情報をパワースペクトルと位相との無矛盾性を考慮しながら推定する必要がある。また、ここまではパワースペクトルの加法性が近似的に成り立つ (音源間の干渉項を無視できる) 場合を仮定したが、この近似が正当でない場合もある。最後に、音のクラスによって位相情報が有効な特徴になり、それを積極的に利用する手法も考えるべきである。全ての点において、複素時間周波数領域あるいは時間領域での処理が自然な解決となりうることに着目し、各領域での手法を第６章と第７章で導入する。

第６章では短時間フーリエ変換（STFT）で得られる複素スペクトログラムが満たすべき条件を導き、パワースペクトルと位相の無矛盾条件を明らかにする。重なり合うフレームのフーリエ変換により構築される STFT スペクトログラムは時間信号の冗長な表現となるため、複素時間周波数領域での任意の複素数の集合が必ずしもある時間信号から得られた STFT スペクトログラムに対応するとは限らない。実信号に対応するスペクトログラムを無矛盾スペクトログラムと呼び、時間周波数領域での数理的な無矛盾性拘束を導出し、その下で任意のパワースペクトルから無矛盾な位相を復元する高速かつ柔軟性の高いアルゴリズムを提案する。

第７章では、観測信号を、少数のテンプレート波形がそれぞれ任意のオンセット時刻において非負のゲインによって生起し、これらが重畳したものと仮定する。この仮定に基づいて立てられる時間領域の観測モデルを用いた適応的テンプレートマッチング手法を提案する。また、これを実現するための効率的な最適化アルゴリズムを導出する。音楽信号や生理的信号を対象とした動作実験により、テンプレートが重なって観測される状況においても、信号の中に繰り返し生起するテンプレートを自律的に取り出し、そのオンセット時刻及びゲインを推定できることを確認した。

最後に、第３章から第６章までは時間周波数領域で処理を行う要請があったために、ウェーブレット変換や短時間フーリエ変換などの古典的なフィルタバンクから得られる時間周波数表現を扱ってきたが、第８章では、観測データから信号の時間周波数分析に一番適したフィ

ルタバンクを自律的に学習する方法を検討する。人間の聴覚において重要な役割を果たしているとされる信号の変調構造に着目し、信号の変調構造を効率よく獲得するために聴覚末梢系を適応させる人間の機能の仮説を立て、これを数理的な枠組みで定式化する。音声データから、変調エネルギーを規準として最適フィルタバンクを学習したところ、妥当なフィルタバンクが得られることを実験的に確認した。

本論文の大きな成果は、広範囲な用途がありながら解決が難しいとされてきたモノラル入力からの音響情景解析の問題の本質を明らかにし、その代表格であるモノラル信号推定・分解・復元・加工の問題を、統一的な観点から効率的に解決した点にある。

キーワード：音響情景解析、音響信号処理、信号の規則性、確率的音源モデル、統計的機械学習、最適化手法、モデルベースの分析手法、基本周波数推定、音源分離、信号分解

Résumé

Un problème essentiel en ingénieurie audio est que les situations réelles d'application sont caractérisées par des conditions extrêmement défavorables, avec des bruits de fond non-stationnaires, des sources multiples, de brèves interruptions dues à des défauts dans le support d'enregistrement ou des paquets de données manquants, etc. Les humains font néanmoins preuve d'une trés grande robustesse dans leur perception et leur compréhension du monde acoustique complexe qui les entoure, en s'appuyant sur les régularités statistiques des sources d'origine et des stimuli qui leur parviennent.

Le but de cette thèse est de proposer une approche statistique de l'analyse de telles scènes acoustiques naturelles basée sur des modèles des régularités de l'environnement acoustique. Nous nous concentrons systématiquement sur une formulation mathématique générale basée sur une fonction objectif, de sorte que les diverses taches considérées peuvent être résolues efficacement comme autant de problèmes d'optimisation sous contrainte bien posés.

Nous présentons tout d'abord un modèle statistique dans le domaine de puissance temps-fréquence pour l'analyse de scènes acoustiques faisant intervenir de la parole, avec des applications comme l'estimation de F_0 et la séparation de sources, et expliquons comment effectuer cette analyse sur des données incomplètes et restaurer les données manquantes. Notant l'importance d'éviter certaines limitations intrinsèques du domaine de puissance et d'utiliser activement l'information liée à la phase, nous introduisons un critère général de cohérence dans le domaine temps-fréquence complexe, et nous développons une méthode d'appariement adaptatif de templates dans le domaine temporel. Enfin, nous examinons l'hypothèse d'une adaptation des filtres périphériques du système auditif basée sur les données afin d'extraire efficacement la structure de modulation des signaux.

Mots-clés : analyse de scènes acoustiques, traitement du signal audio, régularités du signal, modèles probabilistes de sources, apprentissage statistique, méthodes d'optimisation, analyse à base de modèles, estimation de la fréquence fondamentale, séparation de sources, décomposition de signaux.

Contents

Chapter 1

Introduction

1.1 Immersed in a complex acoustical world

Picture yourself in a boat on a river, with wind in your ears and gulls and their cries. With the loud noise of the engine in the background, you are chatting with friends while listening to the music played by the band on board. Somebody calls you, you answer without any difficulty. Although often interrupted by the birds' calls, your conversation with your friends flows smoothly.

The ease with which humans are able to deal with such adverse situations is astonishing. Inside such an extremely complicated acoustical scene, with several people speaking, from different directions, with various types of noises coming from the boat's engine or from the birds moving around the boat, with reverberation due to the reflections of sound on the boat's walls and from time to time on bridges under which the boat passes, humans are still able to follow a conversation or focus on a melody line, sometimes without even perceiving them as interrupted or corrupted by the surrounding noise, constantly adapting to the changing acoustic environment.

Although the object of intensive research for about half a century, building machines coming anywhere close to the skillfulness of the human auditory system has proven utterly difficult. Automatic speech recognition algorithms see their performance drop dramatically as soon as noise or reverberation come into play; Fundamental frequency estimation becomes a challenge when dealing with noisy or polyphonic signals; Underdetermined source separation, with more sources than channels, and in particular monaural source separation have not yet been satisfactorily solved, etc.

The crucial problem for many audio engineering applications is that most, if not all, real world situations they face are adverse ones, with strong non-stationary background noises,

concurrent sources, reverberant conditions, brief interruptions due to glitches or missing packets, etc. How are humans able to achieve such a robustness in their everyday activity, apparently without particular efforts?

1.2 The importance of the regularities in the environment

This question is not limited to auditory perception and can be considered for general sensory stimuli, with visual ones as another important example. Its answer lies in the humans' extraordinary ability to intensively exploit the statistical regularities of their environment, i.e., of the signals that reach their senses. Through evolution and development, these regularities have contributed to shape the humans' perception of the world they live in: in order to survive or to gain a comparative advantage, it is indeed decisive for an organism to build up an internal model, or representation, of the external world, and it has to rely on sensory stimuli to do so. Extraction and learning of the structures of incoming sensory stimuli, adaptation to their particularities are thus at the root of the information processing performed by the brain. In most situations the brain faces, what is actually observed is generally a mixture of distorted and partial versions of the original signals. To analyze such complex situations and infer from these observations the original signals, the brain needs to rely on models of what these signals are likely to be, it needs to have an idea of what phenomena are possible and to have some knowledge on their dynamics to estimate how they are likely to move or evolve. In other words, the brain needs to be able to answer apparently simple yet fundamental questions such as "what is an object?", in visual perception, or "what is a sound?", in auditory perception.

It then becomes possible to reduce the problem of analyzing a complex perceptual situation from one with infinitely many equally possible solutions to an extremely less ambiguous one where the meaningful solutions are required to respect constraints dictated by these world models. One can for example think here of the problem of separating two audio signals mixed together in a single channel: without any constraint on what the single signals can be, any separation in a given signal and its complement in the mixture would be equally good, but obviously not informative at all. If the signals are actually speech by two persons speaking simultaneously in a single microphone, humans are able to separate them easily.

The objective of this thesis is to similarly learn and exploit such regularities and constraints to develop engineering methods for the analysis of complex acoustical scenes such as the one

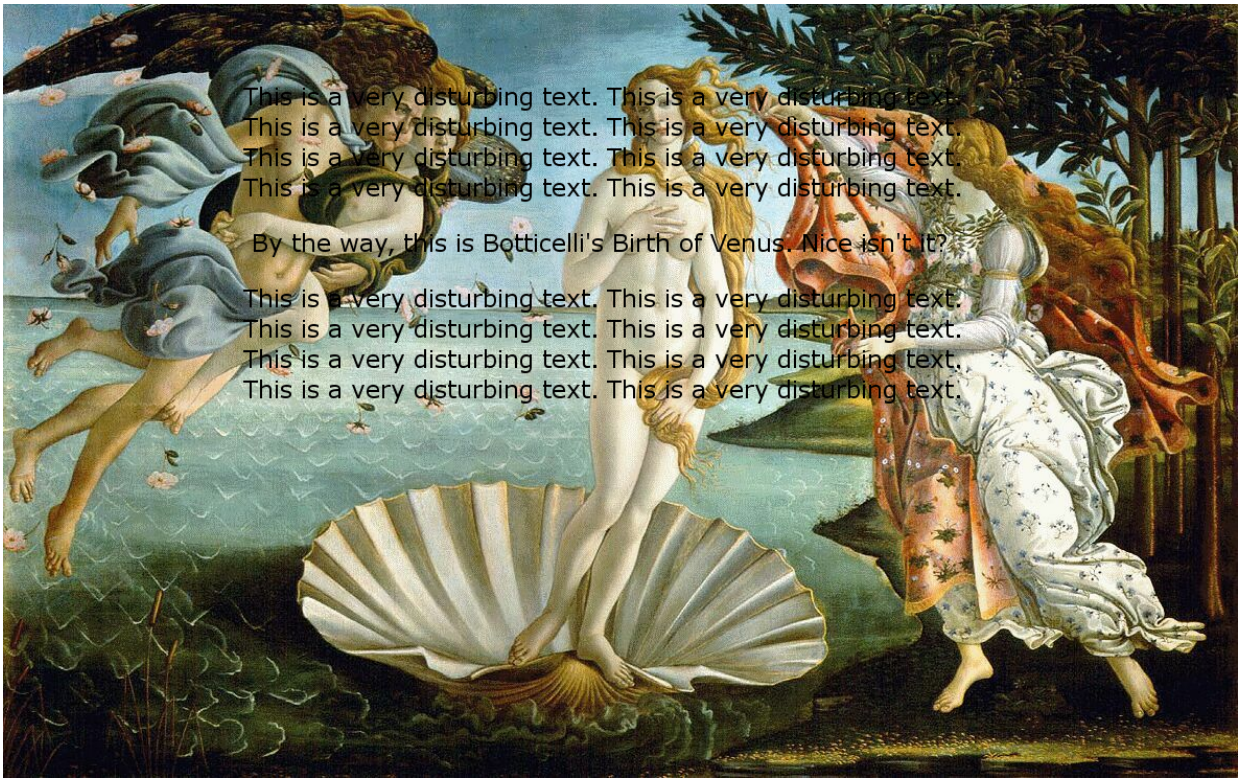


Figure 1.1: *Altered reproduction of Botticelli's Birth of Venus.*

we mentioned above. To give a more precise idea of what we mean by exploiting regularities in scene analysis, let us start with an illustrative example in vision.

1.3 Analyzing natural scenes

1.3.1 Visual scene analysis and image inpainting: an illustrative example

A reproduction of Botticelli's painting "The Birth of Venus" is shown in Fig. 1.1. Parts of the painting are altered by some text. Yet, we are able to make abstraction of that text to analyze the scene in the background: for example, even though parts of Venus's face and body are occluded by letters, we still consider them as a whole and can imagine what the painting would look like without the text. We shall stress here the fact that we do not claim that incomplete sensory stimuli themselves are effectively restored and perceived as complete, but that the underlying representation of the "world" we draw from the observation of the painting is itself complete. Arguably, we can think that it is based on this complete representation that one can imagine a possible completion for the occluded parts of the painting, but the question as to whether the completion occurs at an early or late stage of perceptual organization is left open.

Filling in such altered or missing information is actually a common task in the restoration of damaged paintings, which is called inpainting. To perform such a task, image restoration experts can rely on a wide set of cues: these cues can range from very local ones, such as the continuity of boundaries and colors or the direction of neighboring brush strokes, to higher-level information such as similar patterns or textures occurring elsewhere in the painting, the general style of the painting or even the style of the painter at the time of the composition. More simply, a prior idea of the shape and proportions of a human face or a human body for example could be used to recover parts that are entirely occluded. Altogether, we can consider a whole hierarchy of cues, from models of the signal or stimulus which apply very generally, for example here the continuity of boundaries, to more specific models of the underlying scene, such as the knowledge on the shape of a human face or even on the intentions of the painter. Note the dual relation between perception and restoration: arguably, the same cues which allow us to perceive altered parts of the painting as complete (but occluded) would allow a restoration expert to convincingly reconstruct them.

Such ideas have recently been used in digital image processing to develop computational implementations of image inpainting, for example exploiting local regularities through diffusion, and more global regularities through texture and structure propagation or exemplar-based region filling, where missing pixels are restored using pixels with a similar neighborhood in the intact parts of the image [26, 51].

We would like to develop engineering techniques to perform a similar task on audio signals, i.e., to take advantage of the regularities of natural sounds to analyze an acoustical scene even in adverse situations, with background noises, concurrent sources, or with some parts missing or masked.

1.3.2 Acoustical scene analysis and audio inpainting: towards an audio counterpart

Although it may seem less obvious than for visual scenes, for which we are more accustomed to using such a vocabulary, analyzing auditory scenes is a natural task that we perform constantly. Partly inspired by earlier work on vision, Bregman published in 1990 a comprehensive treatise on the processes underlying auditory organization in humans [34], gathering a large corpus of psychoacoustical experiments in a coherent framework. He showed that a central process is that of "auditory stream segregation", which determines what must be included or excluded from our perceptual descriptions of distinct audio events. The scene is first decomposed into a collection of distinct sensory elements, which are then grouped

together according to some grouping principles accounting for the likeliness that they belong to the same audio event, such as common onset or common amplitude variation. These grouping principles correspond to the "world models" that we mentioned above: they constitute some of the cues that the auditory system uses to determine what to consider as a single source, or as a sound unit, inside a complex scene, in much the same way as we are able to visually distinguish individuals in a crowd or to isolate the various elements of the Birth of Venus. Many engineering applications can be expected from the implementation of such abilities in machines, such as source separation, denoising, dereverberation, audio coding, etc. However, no definitive solution to this problem has been proposed yet.

As we already noted above, most situations that humans face involve not only distorted or mixed but also incomplete stimuli, and yet they perceive the different elements of the scene as complete. The fact that we are able to isolate the various elements of the Birth of Venus even in the presence of an occluding text can be linked to an interesting ability of the human auditory system called auditory induction, or continuity illusion, in which speech with short gaps of silence becomes more understandable when silence is replaced by a louder broadband noise, and is then perceived as complete. In an engineering perspective, the louder broadband noise acting as a masker and the occluding text can be considered playing here similar roles: giving hints on the parts of the scene where underlying data which needs to be treated as a whole may have been occluded or damaged, and leaving a degree of freedom to the processes that determine that underlying entity on what to put in the occluded parts to be coherent with their world models. In the same way as for images, we can thus consider the inpainting of acoustical scenes, or audio inpainting, where missing or occluded parts of a scene are inferred, based on both local and global regularities, for example from the continuity of the power envelope in the time direction to the use of information given by typical spectro-temporal patterns recurring throughout the scene or even learnt *a priori*. We shall stress here the fact that the determination of the underlying acoustic events in spite of the incomplete data on one side, and whether the occluded parts are actually reconstructed or not on the other side, are different issues. As noted earlier on vision, it is still a matter of debate whether incomplete stimuli are actually reconstructed at low levels in human perception (the answer being probably negative). Similarly, actual reconstruction may be unnecessary in certain engineering applications, although it may be obtained as a by-product. In any case, being able to deal with occluded or missing data as well is an important point in the improvement of robustness of many signal processing algorithms and the design of appealing applications such as missing-data speech recognition, audio signal

interpolation, or bandwidth expansion.

We have presented the problem we would like to solve – the analysis of natural acoustical scenes, and the general strategy that we plan to adopt – the use of world models from which to infer and of relevant constraints under which perform optimization. We shall now describe in more details the ideas we will investigate and the different steps we will follow.

1.4 A statistical approach to the analysis of acoustical scenes

The goal of this thesis is to propose a statistical approach to the analysis of acoustical scenes, with applications in many aspects of audio signal processing such as fundamental frequency estimation, source separation, denoising, missing-data reconstruction, audio signal modification, etc. The general idea will be to introduce computational world models on which to rely, so that the considered tasks can be formulated as the optimization of an objective function under relevant constraints. All along this work, our primary focus will be on engineering applications, and, following the strong parallel we presented with perception, we will borrow from knowledge in human perception and psychoacoustics when it may help in the design of such applications.

1.4.1 Models and constraints, inference and optimization

We shall now explain concretely how we plan to apply the philosophy we presented above to the analysis of natural acoustical scenes. We explained the importance of relying on world models to make sense of distorted and incomplete stimuli. Several arguments converge to motivate a statistical approach to the problem, the most obvious of which being that we are focusing on the exploitation of *statistical* regularities of the environment. First, we can see here the analogy with Bayesian inference: the world models can be considered as a sort of prior information, and observations as new evidence. Analyzing the underlying world can be understood as looking for the most likely explanation of the observations given the models of the world we have. Moreover, we need to be able to deal with incomplete stimuli, both to perform inference in spite of the missing parts and to reconstruct, or interpolate, these missing parts, and a statistical framework with hidden variables is a natural and convenient way to formulate such a problem. Furthermore, we also need to tackle the question of the mechanisms by which the world models are acquired. Indeed, understanding how the brain, or here a machine, can rely on such models to perform analysis is by itself a difficult and important question. But the natural question which obviously comes next is to understand

how these models can be learnt, especially as their acquisition needs to be done on the same distorted and incomplete stimuli. A statistical approach seems again natural in such a situation, as statistical learning theory is a well-developed field. Finally, we need to be able to introduce constraints on the possible solutions for the tasks we consider, to ensure that the solutions we obtain are meaningful and that the problems are non-ambiguous. Mathematical formulation enables us to do so and to use a wide range of powerful optimization techniques.

Altogether, a statistical framework provides an elegant and general formulation to the problem, and powerful tools to solve the various tasks we are considering: inference, interpolation, and learning. The key steps to implement this framework will be the design of appropriate models and constraints, the formulation of the inference, interpolation and learning problems, and the derivation of effective optimization methods to solve them.

1.4.2 Global structure model in the time-frequency domain

We begin our work by introducing a statistical model for speech signals. Speech is one of the signals that humans are most often confronted to, certainly one of the richest means of communication, and, as such, lies at the core of most audio signal processing applications. We develop this model in the time-frequency power domain: working in the time-frequency domain enables one to effectively implement grouping principles such as the ones introduced by Bregman, by decoupling (to some extent) the roles of time and frequency, and is presumably close to the domain where human auditory organization is performed, cochlear filtering resulting in a time-frequency decomposition of the signal as well; moreover, we work in the power domain as phase may be seen as a difficult factor to handle, and integrating it out by considering only magnitude (or, equivalently, power) may lead to a more robust analysis, while reducing the problem to a non-negative space may also give better perspectives in the optimization process. Discarding phase is, however, not without raising problems of its own, and modeling in the magnitude or power domain implies inherent limitations and approximations, issues which we shall investigate later in this thesis.

The model we develop for speech, called Harmonic-Temporal Clustering (HTC), is based on a harmonically-constrained Gaussian mixture model, and captures effectively the constraints a signal is likely to respect to be considered by the human auditory organization process as a harmonic acoustic stream with continuously varying pitch, such as the voiced parts of a speech utterance. It is indeed designed to inherently follow Bregman's grouping principles. We also introduce a broadband noise model in order to deal with noisy environments, and explain how to efficiently estimate the parameters of both models. We then show how these

models can be applied to perform several acoustical scene analysis tasks involving speech signals in both concurrent (several speakers) and noisy environments, such as fundamental frequency (F_0) estimation, source separation or denoising. Finally, we show how to extend acoustical scene analysis based on such statistical models to incomplete stimuli, making analysis robust to gaps and enabling us to reconstruct the stimuli as well.

1.4.3 Overcoming some limitations of magnitude or power domain modeling

As noted above, modeling in the magnitude domain has inherent limitations. Indeed, by discarding the phase, we throw away some information and lose the correspondence between time domain and complex time-frequency domain. This raises several issues that need to be dealt with.

First, if resynthesis of a time-domain signal is necessary, the phase information needs to be estimated from the available information, i.e., the magnitude spectrogram. Using a wrong phase, which is not "coherent" with the magnitude spectrogram from which one wants to resynthesize a signal, leads to a reconstructed signal with perceptual artifacts, and whose magnitude spectrogram is actually different from the magnitude spectrogram used in the reconstruction.

Second, additivity of signals is not true anymore: even though the waveform of a mixture is equal to the sum of the waveforms of each component, assuming additivity in the power domain (and even more so in the magnitude domain) is at best an approximation, and at worst totally wrong, as the cross-terms are in general not equal to zero: although it is true that they are in expectation, they may very well be non-zero almost everywhere. The expectation argument fits well in many statistical frameworks, but a better argument to justify to some extent working in the power or the magnitude domain is that of sparseness: many acoustic signals, such as speech, are actually sparse in the time-frequency domain, i.e., their energy is concentrated in very small regions and zero elsewhere. Thus, in a mixture of two sparse signals, regions with non-zero energy for each signal will most likely not overlap, in which case the cross-terms in the power are indeed truly equal to zero. The additivity assumption is then, however, still only as true as the sparseness assumption is.

Third, although modeling the phase is often considered as an intricate problem, and phase is sometimes argued as being a feature which the human ear is relatively blind to, phase still may contain relevant information which could be exploited. In electronic music, and to some extent for some instruments such as piano or percussive instruments, the waveform

of the same note or sound played several times is perfectly or nearly perfectly reproducible from one occurrence to the other. Such information is partially lost when working in the magnitude domain, and exploiting it is another motivation to work in the complex time-frequency domain or the time domain.

We present two frameworks to either overcome or avoid these issues: one is based on a careful study of the structure of complex spectrograms, while the other attempts to model signals directly in the time domain, at the waveform level.

Consistency constraints in the complex time-frequency domain

Under some conditions on the analysis and synthesis windows, there is a perfect equivalence between a time-domain signal and its complex spectrogram constructed using the short-time Fourier transform (STFT). Moreover, the STFT being a linear transform, the additivity assumption still holds true in the complex time-frequency domain. However, as the STFT representation is obtained from overlapping frames of a waveform, it is redundant and has a particular structure. Thus, starting from a set of complex numbers in the complex time-frequency domain, it is not guaranteed whether there exists a signal in the time domain whose STFT is equal to that set of complex numbers. This fact has three consequences: if we were to work in the complex time-frequency domain, for example performing source separation based on some hypothetical model of the complex spectrogram of a sound, we would need to ensure that the separated complex spectrograms are all proper spectrograms (we shall call them "consistent spectrograms"), i.e., that they all are the STFT of some time-domain signal; if we were to work in the time-frequency magnitude or power domain, for example designing a Wiener filter or a binary mask, we would also need to ensure that the obtained magnitude spectrograms are such that there exist consistent complex spectrograms of which they are the magnitude parts; finally, if we now were to reconstruct a time-domain signal from a magnitude spectrogram, we would need to ensure that the phase information we use along with that magnitude spectrogram makes the resulting set of complex numbers a consistent spectrogram as well.

In any case, we need to find a way to ensure that sets of complex numbers indeed respect the structure of a proper complex spectrogram. We do so by defining a consistency constraint in the complex time-frequency domain, and deriving from it a cost function. We show how this cost function can be used to estimate the phase which best corresponds to a given magnitude spectrogram, and explain how it could be used inside many other signal processing algorithms as a consistency prior on complex spectrograms.

Time-domain modeling

The most naive and easy way to avoid non-additivity or resynthesis issues is obviously to stay in the time domain. Although we may then lose some advantages such as the decoupling between time and frequency evolutions, we do not need to make any further analysis assumptions which may introduce a bias (what kind of filter use to perform the time-frequency decomposition, for example), additivity perfectly holds, resynthesis is unnecessary, and the raw waveform is the original stimuli on which the auditory system relies. Even though the argument of a supposed variability of phase across repetitions of a sound and that of a possible phase blindness of the ear are often used to motivate working in the time-frequency magnitude domain, this variability is still an open question. On the contrary, as we mentioned above, for percussive instruments or the piano for example, and obviously for electronic music where sound samples are exactly repeated, the waveforms of several occurrences of a sound repeat with a good reproducibility, and this information could be used to one's advantage. Developing an algorithm directly in the time domain is thus both challenging and promising.

We consider here very simple assumptions to design our model, called shift-invariant semi-non-negative matrix factorization (shift-invariant semi-NMF): we assume that the observed waveform is the superposition of a limited number of elementary waveforms, added with variables latencies and variable but positive amplitudes. The model is more general than the HTC model we developed in the time-frequency power domain in that it does not assume harmonicity and can learn any type of sound or more generally can be applied to data other than audio. We further assume that the amplitudes are sparse, to ensure that the elementary waveforms capture meaningful information reappearing at various time instants. We show that this model can be used to effectively recover recurring templates and their activation times from a mixed waveform, for audio signals (separation of overlapping drum sounds) as well as for extracellular recordings (spike sorting with overlapping spikes).

1.4.4 Data-driven modeling

We have so far put emphasis on how the brain relies on world models to analyze its environment, and how we could design algorithms relying on statistical models to perform similar tasks, but we have not yet addressed the problem of the acquisition of such models. The model of speech signals we introduced above has been tailored, based mainly on prior knowledge on the particularities of speech and on the human auditory organization process. It is interesting, for two reasons, to wonder how one can extract regularities directly from the

stimuli. First, the brain needed to build up the models it is using from distorted, mixed, and incomplete stimuli, in an unsupervised way; modeling this process could both lead to new insights on the way the brain may work and to new theoretical results in machine learning theory. Second, although using tailored models and constraints enables the use of prior knowledge, the results we can expect are limited by the quality and the appropriateness of that prior knowledge. It may be more rewarding to try to learn the most appropriate model directly from the data. Moreover, trying to learn the models which are best fitted to a certain type of signals may indirectly give us information on the structure of that signal. We investigate this problem in two ways.

First, the shift-invariant semi-NMF algorithm we introduced above for time-domain modeling actually achieves a type of data-driven learning, as it extracts, in an unsupervised way, relevant constituents recurring in a waveform. In the same way as what has been done with sparse coding [188], we could also expect that training these constituents on a large corpus of signals will lead to filterbanks which are most suited to obtain a compact and meaningful representation (or in some sense a time-frequency analysis) of that particular type of signals, and in the case of speech and environmental sounds possibly to equivalents of cochlear filters.

Second, we focus on the extraction of the modulation structure of audio signals, and in particular speech, trying to learn time-frequency analysis filters which are "most-suited" to modulation analysis, based on a modulation energy criterion. We noted above that one of the advantages of working in the time domain was to be independent of a choice of particular analysis parameters. Finding "natural" parameters which suit a particular signal for time-frequency analysis is indeed an important issue. Modulation seems to play a central role for auditory perception, and if we were to consider the possibility of a tuning of the initial acoustic processing by exposure to the regularities of natural signals such as speech, it would make sense to assume that during the course of development and/or evolution, the human ear and brain adapted for modulation analysis through a data-driven learning process. We design a mathematical framework to investigate this hypothesis, and show that filterbanks optimized on speech data are close to classical filterbanks.

1.5 Outline of the thesis

A map of the thesis is shown in Fig. 1.2. We first investigate what it means and what it takes to learn and exploit the structures of sound in Chapter 2. We then introduce HTC in Chapter 3 as a global structured model in the time-frequency power domain, and show in Chapter 4 how to use this model for the analysis of acoustical scenes involving speech

signals, with applications such as F_0 estimation, source separation and speech enhancement. In Chapter 5, we explain how to use such global structure models to perform scene analysis with incomplete data, and how to recover the missing data. In Chapter 6, we study the structure of phase and more generally of complex spectrograms, introduce a cost function determining the consistency of complex spectrograms, and present as one of its application a fast and flexible solution to the inherent problem of the determination of phase when working in the magnitude domain. In Chapter 7, we develop another approach to overcoming some of the limitations of the magnitude domain, such as non-additivity, difficulty of resynthesis, and discarding of the phase information, by working directly in the time domain, and introduce the shift-invariant semi-NMF algorithm, which automatically decomposes a waveform as a combination of elementary patterns. Finally, in Chapter 8, we investigate the hypothesis of a data-driven adaptation of the peripheral filters to efficiently extract the modulation structure of the signals.

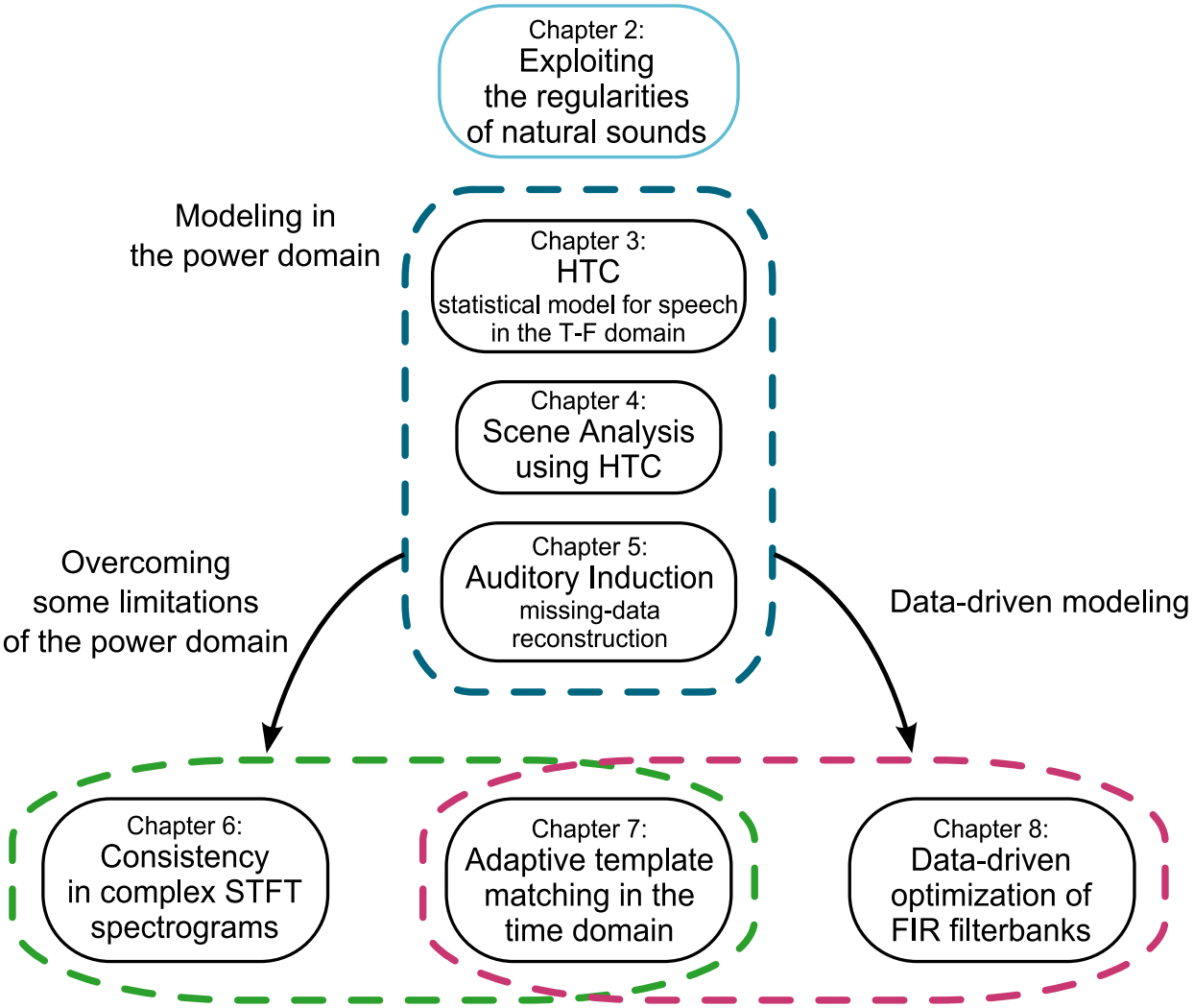


Figure 1.2: *Map of the thesis.*

Chapter 2

Exploiting the regularities of natural sounds

2.1 Introduction

What makes a sound? Humans are constantly immersed in a rich acoustic environment, where sound waves coming from many sources reach their ears. Yet, despite the tremendous complexity of the resulting acoustic signal at the ears, increased by the filtering performed by the surrounding physical environment (reverberation), humans are able to analyze that signal, focus on parts of it, possibly recognize their origin, their type, or understand their meaning. Obviously, for such an analysis to be possible, the constituents of the acoustic signal must be characterized by structures or regularities on which to rely. The regularities of natural sounds are related to their production process. Sound is a vibration, transmitted through media such as gases, liquids or solids. In the air, it is transmitted as a longitudinal pressure wave, characterized by alternating pressure deviations from the equilibrium pressure, forming local regions of compression and rarefaction of air molecules. This wave is generated by a physical process, whose characteristics determine the structure of the emitted sound wave. For example, the strings of a violin vibrate under the action of the bow and resonate through the sound box. Similarly, some insects such as the cricket produce sound by rubbing together parts of their body, a process known as stridulation. Although their complexity is amazingly large, the structure of natural sounds is thus both determined and limited by the physical constraints of the process which generates them. There may be for example limits on the speed variation of the muscles or on the size of the organs of an animal which directly influence the range of possible sounds. At a higher level and at larger time scales, for acoustics waves such as speech, linguistic constraints also occur, while for insects or birds for example, different types of song may exist.

By structure, we shall understand here a restriction of the possible, i.e., a reduction of the dimensionality of the considered problem. For example, knowing how the articulators' dynamics limit the variation rate of the speech spectrum can spare us from modeling the faster variations. Similarly, if the prosody or the grammar of a language does not allow certain sequences, and if we know that the target speech signal is spoken in that particular language, then those sequences need not be considered in the recognition process. The knowledge involved does not need to be a prior, hard-coded one, and could be learnt as well. For example, parametric time series models such as auto-regressive modeling or sinusoidal modeling can be considered as ways to reduce the range of the possible. We will also see that data-driven algorithms such as non-negative matrix factorization (NMF) or sparse coding actually provide ways to train a convenient basis to express a corpus of signals, and thus reduce the dimensionality of that corpus to the signals which can be obtained as a combination of the trained basis. This "dimensionality reduction" or simplification of the problem is crucial for pattern recognition, interpolation, coding, etc., as, without prior assumptions, the dimensionality of possible signals is too large to allow efficient learning. Too many dimensions indeed lead to worse generalization, an issue known as the *curse of dimensionality* [24].

Altogether, natural sounds have a very rich temporal structure, which spans a wide range of time scales and provides a large variety of cues to exploit. Algorithms to extract or learn this structure are key elements in most audio signal processing techniques. This is particularly obvious in the case of speech processing: the production process of speech from linguistic level to acoustic level and the structure of speech signals from the time scale of formants to that of words and sentences have been particularly intensively studied, leading to the development of an extremely large corpus of methods with applications ranging from speech coding to speech enhancement or speech restoration. We shall thus start this chapter by focusing on speech as a typical example to illustrate the development of techniques exploiting the inherent structure of an acoustic signal.

We shall then focus on the analysis of natural scenes, with algorithms which are characterized by a more "perception-oriented" approach, inspired by the remarkable abilities of the human auditory system. As shown by Bregman in a seminal book [34], the human auditory system performs an active analysis of natural scenes by first decomposing the input acoustic signal into spectrogram-like elements, which are then grouped together to form what is considered a sound, based on a set of grouping principles. Many researchers have attempted to develop algorithms which would implement, more or less directly, computational equivalents

to such mechanisms. These methods can be roughly separated into two categories. The first is the so-called computational auditory scene analysis (CASA) framework, which stays closer to the perceptual mechanisms observed in the human auditory system and tries to derive from them effective methods to separate an input acoustic signal into its constituent "sounds". The other category is only indirectly related to the human auditory system, and gathers techniques which attempt to build their own structures, categories and principles, from the statistical structure of environmental stimuli in a data-driven way, as the human auditory system might have itself evolved under the process of evolution.

We will first focus on speech processing, briefly describing speech production models and then presenting a range of techniques which exploit the specificities of speech to perform speech coding, enhancement and restoration. We will then briefly review the principles underlying the analysis of natural scenes by humans, and give an overview of both CASA and data-driven methods which try to develop a computational counterpart to auditory perception.

2.2 What regularities to exploit, and how to exploit them: speech as a typical example

2.2.1 Speech production process

The global process of human speech communication, from the formation of a linguistic message in the speaker's brain to the arrival of the message in the listener's brain, is called the speech chain [59,60,83]. Omitting the final auditory and perceptual level, which consists in the way the acoustic speech signal is perceived by the listener (and similarly by the speaker in a feedback loop, allowing him to continuously control the speech production by his vocal organs), the production part of the speech chain consists in three levels. The highest one is the linguistic level, at which the speaker conceptualizes an idea to be conveyed to the listener, formulates it into a meaningful linguistic structure using words, organized into sentences according to syntactic constraints, and which are successively composed of syllables, phonemes, all the way down to phonological features. At the physiological level, the brain then produces motor nerve commands specified by these features which determine how to move the articulatory muscles (lips, tongue, larynx, jaw, velum, etc.) in order to produce the intended speech sounds, with a constant corrective feedback. Finally, at the acoustic level, the speech wave is produced and propagated. This process can be further decomposed into three subprocesses, namely source generation, articulation, and radiation [78]. Air is

pushed out from the lungs under the action of the diaphragm, and passes through the vocal cords, developing a phonation type. This creates a time-varying sound source which is then filtered by the vocal tract, whose shape is continuously adjusted to produce various linguistic sounds through the action of the moving articulators. Finally, the sound wave is then radiated from the lips and/or nostrils, and can be picked up by microphones that respond to changes in air pressure, sampled and digitally processed by computers.

As a result from this multi-level production process, the speech signal is characterized by a very rich temporal structure, spanning a wide range of time scales. Formants which result from the resonances in the vocal tract typically appear at time scales shorter than the milli-second, while pitch information is located at the order of the milli-second, phonemes at the order of tens of milli-seconds, syllables at the order of hundreds of milli-seconds, up to words and sentences whose time scale is around a second or a few seconds.

2.2.2 Speech modeling and applications

The modeling of a time series can be envisioned in a wide range of fashion, from tailor-made models, in which we put as much knowledge of the precise signal we are studying as possible, to totally data-driven approaches, in which merely the tools for analysis are given and trained to lead to a compact and hopefully meaningful representation of the signal. As speech is such a well-studied signal, most attempts at its modeling have tried to use knowledge on its production process, but recently more data-driven approaches have appeared, trying to go one step further in the fitting to the actual data and automatically extract the structure without having to hard-code pre-decided knowledge whose misfit with the data could lower performance. Being able to obtain a compact or meaningful representation of a signal is a key step in the development of techniques to efficiently transmit it, modify it, or restore some of its parts if they have been lost.

AR modeling

The speech production mechanism at the acoustic level is often modeled as a source-filter model, which completely separates the source from the articulation. The source can be modeled as a periodic impulse train for voiced sounds, and as white noise for unvoiced sounds, possibly allowing mixed excitation for voiced fricatives for example. The filtering by the vocal tract is generally modeled as a concatenation of lossless tubes with time-varying cross-sectional areas, approximating the time-varying vocal tract area function, and can thus be mathematically represented as an all-pole model, leading to the auto-regressive (AR) modeling of speech, which is intensively exploited in speech coding, e.g., in linear predictive

coding (LPC) [14, 75, 103, 131, 158], speech enhancement [39, 124] and restoration of missing speech data, as we shall detail later on.

AR modeling is a very important and widely-used technique in many areas of time series processing, and particularly in speech and audio processing. It consists in representing the current value of a signal as a weighted sum of P previous values and a white noise term:

$$s(n) = \sum_{i=1}^{P} a_i s(n-i) + \epsilon(n). \tag{2.1}$$

It can represent fairly well many stationary linear processes, and even though auto-regressive moving average (ARMA) modeling, which allows for both poles and zeros, is sometimes more appropriate, it is often preferable in practice to use a higher-order AR model instead to benefit from its analytic flexibility. There indeed exists several efficient methods to estimate the LPC coefficients a_i or related parameters. The LPC coefficients can for example be obtained through solving the so-called Yule-Walker equations using the covariance and the auto-correlation methods, or the maximum-likelihood method. As they suffer from the fact that quantization errors may lead to unstable synthesis filters, equivalent representations such as the PARCOR coefficients or the Line Spectrum Pair (LSP) coefficients, for which stability is guaranteed, have been proposed. We refer the reader to [83] for a thorough review of LPC analysis.

The interest of AR modeling for speech coding is that, if the prediction is good, the variance of the error ϵ will be much smaller than that of the original signal s, and, assuming the quantization procedure is adapted to the variance of the signal, quantizing ϵ and s with the same number of bits will lead to a much smaller absolute quantization error for ϵ than for s. Many speech coders have been developed which rely on AR modeling, such as differential pulse code modulation (DPCM), code-excited linear prediction (CELP) [177], or mixed excitation linear prediction (MELP) [136]. A review can be found in [75].

AR modeling assumes that speech is quasi-stationary, and modeling is often performed on a frame-by-frame basis. However, as the characteristics of the vocal tract change continuously, the parameters of the AR model should also evolve continuously. To take this into account, methods based on time-varying AR processes have been developed. These models can be used for both speech modeling and enhancement [200]. The estimation of the parameters is much more arduous, and often relies on involved techniques such as particle filtering [84]. Considering the bad interpolation characteristics of the LPC coefficients and the fact that the stability of the synthesis filter is not guaranteed, methods based on time-varying PARCOR

coefficients have also been developed [81, 85]. Another advantage of such methods is that the variation of the parameters directly reflects the variation of the shape of the vocal tract, as the PARCOR coefficients can indeed be interpreted as parameters of an acoustical tube model approximating its characteristics.

Missing-data interpolation

Deriving a good model of a signal enables to enhance its underlying structure, and by uncovering the connections between various parts of the signal, possibly at different time scales, it enables in particular the prediction and reconstruction of missing parts. In modern communication networks, digital signals are usually transmitted by packets, i.e., in uniform frames, which may be lost or delayed during the transmission, making it necessary to recover the information of the missing frames using the available information so as to minimize the perceptibility of the erasure. Various models have been used to derive missing-data interpolation techniques, some of which are not limited to speech signals.

AR modeling is of course one of them, as by nature it predicts the value of signal samples as a linear combination of previous samples. Janssen, Veldhuis and Vries [104] developed an auto-regressive interpolator which converges to a local maximum of the likelihood by alternately maximizing w.r.t. the missing data and the model parameters. This model has been extended by Vaseghi and Rayner [196] to consider not only immediately preceding samples but also samples which are a pitch period apart. The idea here is to use the quasi-periodicity of voiced parts of the speech signal to provide a better estimate for the missing samples, the down side being that an estimation of the F_0 is necessary. Another modification to the auto-regressive interpolator has been proposed by Rayner and Godsill [164], to deal with the fact that the original algorithm attempts to minimize the energy of the excitation, resulting in an excitation in the gap with lower energy than surrounding excitation and leading to over-smoothed interpolants whose amplitude decreases at the center of the gap. Their algorithm relies on the decorrelation of the missing data samples and on sampling in the transformed domain such that the resulting excitation has a constant energy. Extending the sampling-based interpolation idea, Ó Ruanaidh and Fitzgerald [151] investigated the use of Gibbs sampling to generate typical interpolates from the marginal posterior distribution of the missing samples conditional to the observed samples. Methods based on time-varying AR processes have also been developed [160], expecting a better modeling of the continuously evolving characteristics of the vocal tract. More details and references can be found in [86, 199]. Going one step further in this direction, one could imagine using time-varying

PARCOR models for missing-data interpolation as well, or, considering the necessity for good interpolation characteristics in this situation, developing a framework based on a time-varying LSP model, although this last possibility does not seem to have been explored yet.

Another important model for speech and audio signals, the so-called sinusoidal modeling introduced by McAulay and Quatieri [135], was used by Maher [130] to develop an algorithm for missing-data interpolation. Originally applied to Analysis-by-Synthesis systems, leading to high-quality synthesized speech, the range of application of this model has widened to text-to-speech synthesis, speech modification, or coding. It represents the signal as a superposition of K complex sinusoids which are assumed to have constant frequency and amplitude:

$$s(t) \triangleq \sum_{k=1}^{K} A_k e^{j\mu_k t}, \quad t \in (-\infty, \infty), \tag{2.2}$$

where μ_k and A_k represent respectively the frequency and complex amplitude of the k-th sinusoidal component, the arguments $\arg(A_k)$ representing the phases at time $t = 0$ (initial phase). For each short-time analysis frame, the parameters μ_k and A_k need to be estimated, finding an effective method to do so being a rather intricate problem. We refer to [106] for a review of existing estimation methods. Maher [130] uses this model for the interpolation of missing samples in audio signal by performing the interpolation directly on the parameters of the sinusoidal model.

Data-driven approaches

A recent and promising trend in signal processing is that of data-driven modeling. Structure is extracted directly from the data in an unsupervised way. We will come back in more details to this type of algorithms in the next section, but let us first give a few examples of such processing applied to speech signals, especially in the context of automatic speech recognition (ASR).

Although conventional ASR systems are based on trained stochastic systems (Gaussian mixture models for acoustic modeling, N-gram models for language modeling), Hermansky [94] argues that the feature extraction module should be trained as well: first, using incorrect prior knowledge is worse than using none, so the required knowledge should be acquired directly from the data; moreover, training of the feature extraction module would introduce speech-specific but task-independent knowledge in it, facilitating the work of the subsequent pattern classification module. Several techniques have been introduced which rely on a data-driven modeling at the feature level, such as the data-driven design of RASTA filters [195], the design of optimized spectral basis functions by linear discriminant analysis

and principal component analysis [96, 193], or the training of the feature extraction module based on discriminative training [28, 29].

Going further in this direction, Lee [117] recently proposed a data-driven "knowledge-rich" modeling as a new paradigm to overcome the weaknesses of the conventional "knowledge-ignorant" one. Similarly, Deng [60] claims that, although some tools motivated by knowledge on speech production and speech signal structure are used, such as Mel Frequency Cepstrum Coefficients (MFCCs) or N-gram language models for example, ASR has been treated so far as a "generic, loosely constrained pattern recognition problem" and that we should "put speech science back into speech recognition". He attempts to do so by modeling the dynamics of speech at various levels of the speech chain mentioned earlier in the general framework of Dynamic Bayesian Networks. More references on this approach and on production-oriented models for speech recognition can be found in [60] and in the review papers by McDermott and Nakamura [137] and King et al. [113].

2.3 Analyzing natural scenes

2.3.1 Human auditory functions

The human auditory system is extremely effective at detecting and segregating individual components in sound mixtures. Although humans are almost constantly immersed in a complex acoustical environment composed of signals coming from several sources of a great variety, they are able to distinguish the individual sound sources, and possibly recognize their characteristics or understand their meaning.

This remarkable ability was already noticed in 1863 by Helmholtz [93], who noted that although the auditory scene constituted by a ballroom was "a tumbled entanglement [...] complicated beyond conception", with "a number of musical instruments in action, speaking men and women, rustling garments, gliding feet, clinking glasses, and so on", yet, "the ear is able to distinguish all the separate constituent parts of this confused whole". Cherry [40] later stressed a particular and representative aspect of this phenomenon, the faculty that humans have to be able "to listen to, and follow, one speaker in the presence of others" or of a variety of noises. He noted that no machine had yet been built to deal with this problem, which he called the "cocktail party problem".

It is only in 1990 with Bregman's landmark book [34] that a coherent framework was presented to explain the processes underlying the perceptual separation of sound sources, based on an important corpus of psychophysical experiments. Inspired by earlier works on vision and by the principles proposed by the Gestalt psychologists based on their work on visual

perception (a review can be found in [153]), Bregman described the auditory organization process in humans as a means to solve a problem which he refers to as "auditory scene analysis" (ASA). A central process here is that of "auditory stream segregation", which determines what is included or excluded from our perceptual descriptions of distinct auditory events. This process consists of two stages. The acoustical scene is first decomposed into a collection of distinct sensory elements (segmentation process), and these elements are then combined into a stream according to their likeliness to have arisen from the same sound source (grouping process). Grouping can occur both for simultaneous auditory components or sequentially in time. Bregman also distinguishes between primitive grouping and schema-based grouping, the former being a bottom-up process relying on the intrinsic structure of sounds, while the latter is a top-down process relying on prior knowledge, in which auditory components belonging to the same learnt pattern are more likely to be grouped. Bregman grounds the grouping principles for primitive grouping on five elementary laws of Gestalt psychology, namely proximity, similarity, continuity, closure (completion of fragmentary features), and common fate. More precisely, if we consider the acoustic signal as a time-frequency scene, the cues used for grouping are harmonicity, proximity in frequency and time, common onset and offset, coherent amplitude and frequency modulation, continuity in the time and frequency directions, and common spatial location, although this last cue seems to play a secondary role as source separation can still be performed monaurally.

Bregman's work has been the catalyst of a large amount of work on the computational implementation of ASA, or the design of machine systems which would achieve human performance in that task, a field which is known as computational auditory scene analysis (CASA). We shall give a brief review of the work in this area in the next section.

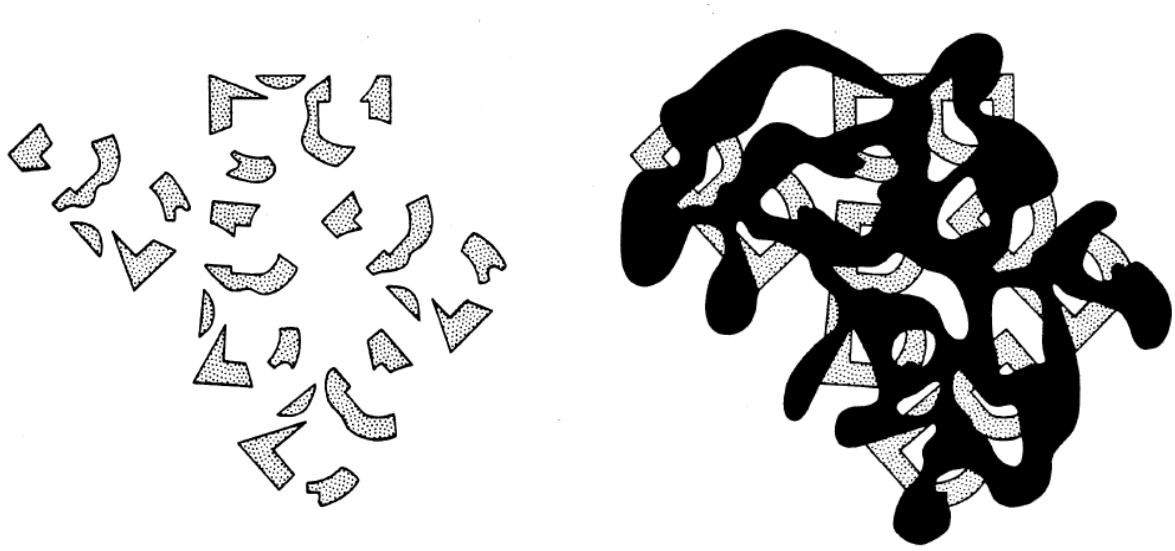
A remarkable faculty of the human auditory system, which seems to have been often overlooked in CASA research (with the notable exception of Ellis [72, 73]), is that of auditory induction, or continuity illusion [111, 207, 208], in which deleted phonemes masked by noise bursts are perceptually reconstructed and genuinely "heard", the listener having difficulties localizing the noise burst within the speech. As summarized by Repp [165], two hypotheses can be considered to explain the origin of the auditory information leading to this restoration. One is that of segregation, supported by Warren [208], in which, guided by top-down expectations, primitive auditory processes attempt to reconstitute the speech signal and separate it from the noise bursts, taking necessary sensory evidence from the noise bursts and leaving a residue perceived as extraneous sound. The other is that of top-down completion: the restoration is an auditory illusion arising from context-induced phonological completion,

or, as described by Bregman [34], "schema-governed stream segregation"; as a schema-driven process, Bregman argues that it does not remove the information it uses "from the array of information that other description-building processes can use", and thus shall not give rise to a residual, unless the segregation is directly induced by the surrounding acoustic context. Experiments conducted by Repp [165] tend to support the latter top-down hypothesis against the former segregation one. In any case, the auditory induction phenomenon is a striking illustration of the law of closure of Gestalt psychology, according to which the human perception system has a tendency to close "strong" perceptual forms which are incomplete, such as a circle partially occluded by an irregular form. More generally, it can be considered as an expression of Mach's "economy of thought", in that it is more rewarding in terms of simplicity of explanation to assume that some parts are occluded but actually present and need to be reconstructed (either at the primitive grouping stage or at higher levels) than to assume that the stimuli is actually composed of several disconnected parts. The fact that auditory induction does not occur if the phonemes are replaced by silence can be linked to the point made by Bregman that our perceptual system needs to be shown that some evidence is missing, as illustrated in Fig. 2.1. Humans can indeed see figures with actual gaps in them, as with no special hint or trigger mechanism, they have no reason to believe that the "missing" parts are not missing but merely hidden.

We shall come back to the auditory induction phenomenon in Chapter 5, where we explain how global structure models in the time-frequency domain can be used to analyze acoustical scenes in the presence of gaps and simultaneously reconstruct the missing data. Following the above discussion on the importance of the presence of information on occlusion, localization of the gaps will be considered known.

2.3.2 CASA and top-down model-based inference

As described above, the human auditory system is extremely effective at separating a sound source from concurrent sources. Inspired by the seminal work of Marr on computational vision [134] and later by Bregman's comprehensive treatise on auditory scene analysis [34], many researchers have attempted to develop computational systems implementing the human ability to actively analyze acoustical scenes and separate sources in a complex acoustical environment, a corpus of work referred to as computational auditory scene analysis. The focus of most research in this area is mainly on the design of artificial systems, but effective algorithms could serve as guides for the investigation of natural processes and a better understanding of the auditory perception process.



(a) Without information for occlusion, fragments do not organize themselves strongly.

(b) When information for occlusion is given, fragments are grouped together.

Figure 2.1: *Illustration of the importance of the presence of information on occlusion for the closure mechanism to happen. (From Bregman [34], with permission)*

The typical structure of most CASA systems follows Bregman's model of auditory organization. Some sort of time-frequency analysis is first performed on the input signal to derive acoustic features. These features may be used to derive an intermediate representation, or directly passed to the grouping processes. Grouping is then performed under a set of grouping principles to gather components which are likely to have arisen from the same source. After identification of the individual sources, the auditory representation can be either inverted to obtain a time-domain waveform for each of the sources, or directly used, for example in an ASR task.

Weintraub [209] was the first to try to use a model of the auditory system to improve the accuracy of speech recognition in the presence of concurrent speech. His system featured the use of time-frequency masks for separation, introduced by Lyon [128], and a computational model for F_0 estimation based on Licklider's pitch perception theory [123], first implemented by Lyon [129] and which was later named correlogram by Slaney and Lyon [181]. His system prepared the ground for subsequent developments in CASA systems, such as the ones by Cooke [48], Brown [35], Ellis [73], Abe and Ando [2–6], Nakatani, Goto and Okuno [144], Wang and Brown [205], Hu and Wang [99,100], among others. More references can be found in [36, 41, 42, 49, 73, 206].

One of the main features present in most CASA systems is the fundamental frequency, F_0, as it is one of the most important cues used by humans to improve the identification of

speech masked by a concurrent speech. An amazingly large number of algorithms have been developed for F_0 estimation, which are reviewed by Hess [98] for single F_0 estimation methods and by de Cheveigné [43] for multiple F_0 estimation. Most methods for F_0 estimation, and more generally for feature extraction and grouping in CASA systems, rely on a two-stage process, where potential candidates for the F_0 are obtained for each frame by a processing in the frequency direction, and a post-processing in the time direction is then performed to eliminate estimation errors and obtain smooth F_0 contours. Separating these optimizations is likely to be suboptimal, and a simultaneous optimization in both time and frequency directions should thus lead to a better accuracy. We will present in Chapter 3 and Chapter 4 a model for scene analysis which performs F_0 estimation through such a joint estimation.

The actual separation based on the results of the segregation and grouping processes is performed in many CASA systems through the use of a time-frequency mask. Introduced by Lyon [128] in a binaural separation system and then by Weintraub [209] in a monaural system, it has been since widely adopted by subsequent CASA systems. Time-frequency masks apply a weight to each bin of a time-frequency representation of the acoustic input (cochleogram, short-time Fourier transform, wavelet transform, etc.) to emphasize regions which are dominated by the target source and suppress regions which are dominated by other sources. The weight values can be either binary or continuous. Continuous values can be interpreted as the energy ratio between the target and the mixture, as in a Wiener filter, or as the probability that the corresponding time-frequency bin belongs to the target source, an interpretation which is a fundamental element of the model we develop in Chapter 3. The use of binary masks can be justified perceptually by invoking the masking phenomenon in human hearing, or computationally by noting that the speech signal is sparsely distributed in high-resolution time-frequency representations and two speech signals thus rarely overlap in that domain, although this last argument becomes invalid for broadband intrusions.

By applying the estimated time-frequency mask to the time-frequency representation of the input signal, a representation of the separated sources can be obtained. This representation is often inverted to resynthesize the separated signal in the time-domain [209], either to make it available directly for listening or evaluation purposes, or to feed it into other systems such as ASR, speaker recognition, or music information retrieval (MIR) systems for example. The inversion of other representations has also been investigated, such as that of the correlogram to segregate sounds according to their F_0 [182]. Some work on separation has also recently been done in the modulation domain [173, 174]. The resynthesis is however not compulsory, and in some situations it may be more rewarding to have a tighter

integration between the segregation and the intended application. This is the case for ASR, for which frameworks have been developed to make the best use of the information given by the time-frequency masks. A missing-feature approach to ASR has been introduced by Cooke et al. [50], in which knowledge of the time-frequency regions which are reliable or not is actively used by either marginalizing out the unreliable features or reconstructing them to fall back on a complete feature vector. However, in this framework the time-frequency mask estimation and the ASR steps are decoupled, making it impossible to model the relationship between segregation and recognition and leading to a suboptimal modeling. Barker et al. [18] proposed to accommodate data-driven and schema-driven processing in a common statistical model called "speech fragment decoding", in which multiple bottom-up organizations of the acoustic mixture are evaluated. Reviews of missing-data approaches can be found in Raj and Stern [159] and Barker [17].

More generally, emphasis has recently been put on model-based approaches to scene analysis, as reviewed by Ellis [74]. Their main motivation, close to that of this thesis, is to make use of knowledge on perception and signal characteristics to design relevant mathematical models and formulate the tasks as statistical inference problems. A representative example of such approaches is that proposed by Kameoka [106], who introduces a complete statistical framework for multipitch analysis. Vincent [201] proposes a model-based approach for source separation of musical signals based on the design of specific instrument models. Durrieu, Richard and David [69] develop a model-based source separation method for singer melody extraction.

2.3.3 Data-driven approaches

The methods we have presented above make a strong use of knowledge on the human auditory system or on the signal to segregate. This is their strength, but it can also be a weakness, as wrong prior knowledge can turn out to be worse than no prior knowledge at all. Moreover, in some cases little is known on a particular signal, and its statistical regularities must be learnt directly from it. The learnt regularities could then be used to derive models to analyze the data.

In addition, as noted by many researchers, from Helmholtz, Mach and Pearson to Attneave and Barlow (references can be found in Barlow's excellent review paper [21]), the exploitation of regularities in the environment, as a key element for the survival of an organism, is likely to have been a major factor in the development and evolution of perception. One could thus imagine designing algorithms which automatically adapt to exploit the regularities of

the incoming stimuli as well. Such a developmental approach for computational auditory perception has been proposed by Smaragdis [183], who showed that the various stages of the computational counterpart of an auditory system could be constructed by letting them evolve based on the input data. Considering further the interaction between action and perception, Philipona et al. [154] showed that a sensory-motor approach [150] could be used to explain how an organism can get to understand the structure of its surrounding world.

For engineering purposes, obtaining compact or low-dimensional representations offers many advantages such as computational efficiency, denoising, greater interpretability, or easier visualization, for example. We have already presented in Section 1 some examples of algorithms which attempt to extract structure directly from the data in relation with ASR. In this section, we give an overview of general methods for the decomposition and the dimensionality reduction of signals, and focus on two recent computational approaches for the design of compact and meaningful representations, sparse coding and non-negative matrix factorization.

General methods for signal decomposition and dimension reduction

The structure of the data one wants to analyze is generally unknown, and a wide range of methods have been developed to try to estimate this structure in an unsupervised way.

The most basic and nonetheless widely used technique is certainly principal component analysis (PCA), which performs a linear rotation of the data by recursively selecting the direction with the largest variance in the subspace orthogonal to the already estimated directions. It is usually computed by singular value decomposition of the data matrix. No particular assumption is made on the distribution of the data, and PCA is really useful only under the assumption that directions with large variances are the interesting ones and the ones with small variance merely noise. The directions with smallest variances can then be skipped, leading to a representation of the data with smaller dimension, most of the variance of the data being preserved. While PCA performs a decorrelation of the data elements, independent component analysis (ICA) attempts to extract independent sources in the signal, and has recently been widely used in blind source separation (BSS) methods based on the assumption that the observed signals are the result of the mixing (with unknown characteristics) of a small number of independent sources (see for example [102] and [132]).

Other methods have been developed based on PCA by introducing additional steps. In slow feature analysis (SFA), Wiskott and Sejnowski [210] transform the input signal through non-linear expansions and apply PCA to the derivatives of the obtained signals to extract

slowly varying features; due to the non-linearity, several consecutive blocks of SFA can be used to lead to higher-level slowly changing features. In close relation to SFA, Valpola and Särelä propose a framework called denoising source separation (DSS) [171,194]. A first PCA step is used to whiten (decorrelate and normalize) the data, and is followed by a "denoising" step which enhances some directions according to prior knowledge on the signal or the particular goal of the task. A final PCA is then performed to estimate the directions which have been enhanced by the denoising step. A different type of data expansion is used in singular-spectrum analysis (SSA) [197,198]. A time series (s_n) is embedded in an M-dimensional space by considering as state vectors the sets of consecutive values $X_n = (s_n, s_{n+1}, \ldots, s_{n+M-1})$, and PCA is applied to extract so-called "empirical orthogonal functions" which are used as a basis for a decomposition of the data. This technique can also be used to deal with time series which are unevenly sampled or contain missing data [115]. Finally, singular value decomposition is also used in latent semantic mapping (LSM) [23] for language modeling or text-to-speech unit selection, modeling meaningful global relationships which are implicit in a large data corpus. While PCA finds directions which are useful for representing data, there is no reason to assume that the obtained directions are also relevant for discriminating between data in different classes. Linear discriminant analysis (LDA) has been developed for such a purpose, and does not look for the maximum variance directions, but for a linear transform of the feature space which would maximize the linear separability of classes by maximizing the ratio of between-class scatter to within-class scatter [68]. This technique has been used for example to derive data-driven spectral basis for speech analysis [193].

We will conclude this short review by noting that techniques such as PCA assume that data lie on a linear subspace, and thus cannot be used when it is not the case, even if the data still inherently have only few degrees of freedom. Manifold learning techniques have been developed to deal with such a situation, such as Tenenbaum's Isomap [191] or Roweis and Saul's locally linear embedding (LLE) [167]. Both techniques rely on the construction of a neighborhood graph. Isomap uses it to extract geodesic distances and perform multidimensional scaling on them, while LLE defines barycentric coordinates using a fixed number of neighbors and uses these coordinates to extract a low-dimensional global coordinate system.

Sparse Coding

In 1961, Barlow [19] suggested that a potential mechanism governing the development of the sensory and perceptual systems might have been efficiency of coding, or redundancy reduction. Although he later slightly changed his viewpoint on the subject [20] to insist more

on redundancy exploitation, the idea of a sparse or efficient coding to represent natural data in a simple way has been investigated by many authors since he made this proposal. Work by Olshausen and Field [145], Bell and Sejnowski [22] and Hyvärinen and Hoyer [101] (a review can be found in [146]) showed using natural images that optimizing receptive fields of an entire population of neurons to produce sparse representations led to the emergence of a set of receptive fields resembling those of simple cells. A similar account has been made for sound by Smith and Lewicki [188], who showed that optimizing a population of spike codes to efficiently encode natural sounds leads to features which show a great similarity with time-domain cochlear filter estimates.

Sparse coding assumes that a signal can be expressed, based on an overcomplete (redundant) dictionary of functions, using only a few number of words from the dictionary. A difficult problem in sparse coding is that of encoding. Although the signal is assumed to be a linear combination of elements of the code, inferring the optimal representation for a given signal is highly non-linear. Determining exactly the sparsest representation can actually be shown to be an NP-hard problem [58]. Greedy algorithms have thus been developed to deal with this problem, such as matching pursuit or orthogonal matching pursuit [133]. The matching pursuit algorithm iteratively selects a waveform from the overcomplete dictionary which best explains part of the signal, and subtracts it from the signal.

The use of sparse coding in audio has recently been intensively investigated, with, for example, applications to polyphonic music analysis by Abdallah and Plumbley [1], object representation of music signals by Leveau [120,121], audio coding by Daudet [55] and Ravelli, Richard and Daudet [163], denoising of music signals by Févotte, Torrésani, Daudet and Godsill [80], monaural speaker separation by Shashanka, Raj, and Smaragdis [180], or source separation based on differential filtering by the head-related transfer function (HRTF) by Asari, Pearlmutter, and Zador [13]. Shift-invariant versions of sparse coding have also been applied to image and audio processing by Mørup, Schmidt, and Hansen [142] and Plumbley, Abdallah, Blumensath and Davies [155].

Non-negative matrix factorization

Since its introduction by Lee and Seung [118] in 1999, non-negative matrix factorization (NMF) has been widely adopted as a very effective technique for linear decomposition and dimensionality reduction of non-negative data. It is considered likely to lead to meaningful representations of the data as it does not allow for cancellations. Mathematically, NMF attempts to decompose a non-negative matrix $V \in \mathbb{R}^{\geq 0, M \times N}$ as the product of two

usually lower-rank non-negative matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$,

$$V \approx WH.$$

Effective multiplicative updates guaranteeing non-negativity and local convergence have been derived by Lee and Seung [119] for two distance functions measuring the goodness of the approximation of V by the product WH, the L^2-norm and the $\mathcal{I}$-divergence. Dhillon and Sra [61,190] discussed NMF algorithms based on Bregman divergences, of which the L^2-norm and the $\mathcal{I}$-divergence are particular cases, and considered the use of penalty functions. NMF can indeed be considered in a Bayesian framework, as already noted by Lee and Seung [118] and explicitly stated in Sajda, Du, and Parra [170], and penalty functions can thus be viewed as prior distributions. We refer to Cemgil [37] for a complete description of NMF in a statistical framework. A related model called semi-NMF has also been introduced by Ding, Li, and Jordan [63, 122], in which only the amplitudes are constrained to be non-negative while the templates can take negative values.

Originally mainly applied to images, NMF was first applied to audio signals for multi-pitch analysis by Smaragdis and Brown [186], Saul, Sha, and Lee [172], and Sha and Saul [178], by considering the magnitude or power spectrogram as a non-negative matrix to decompose linearly. However, additivity in the magnitude or power domain does not hold, but linear decompositions in these domains are usually motivated by invoking the sparsity of the modeled signals, arguing that additivity is then approximately true. In most approaches which consider the power spectrum probabilistically, i.e., as the variance of a random variable, the fact that additivity is true in expectation if the sources are uncorrelated is usually implicitly used. Févotte [79] argues that, in the case of NMF, the right distance to use is the Itakura-Saito distance, as applying NMF based on this distance to the power spectrogram corresponds to assuming additivity of waveforms if the sources are independently Gaussian distributed. One may however argue, as noted by Kameoka [109], that such independence or decorrelation assumptions are only true in the time-frequency domain for analysis windows with infinite length, and using finite-length windows thus results in some sort of approximation. Kameoka recently proposed a promising NMF framework where addition is effectively done in the complex domain [109], thus not relying on any assumption or approximation concerning additivity.

Since the first applications to audio, the range of methods based on NMF has greatly broadened. Smaragdis first extended NMF to take shifts in the time direction into account, developing a framework called non-negative matrix factor deconvolution (NMFD) [184,185].

NMFD was later extended by Schmidt, Mørup and colleagues to also consider shifts in the frequency direction, and subsequently to include a sparseness constraint. Their methods, called non-negative matrix factor 2-D deconvolution (NMF2D) and sparse NMF2D (SNMF2D) [140, 141,175,176], can be applied to audio data under the assumption that log-frequency spectral patterns are approximately pitch-invariant. On the same idea of sparse and shift-invariant feature extraction of non-negative data, Smaragdis [187] recently proposed a framework called probabilistic latent component analysis (PLSA) which has the advantage to be fully probabilistic. The use of priors with NMF for audio modeling is investigated by Virtanen [204], who also used NMF for monaural source separation [203]. Vincent et al. [202] considered the introduction of harmonicity and inharmonicity constrains in NMF models for polyphonic pitch transcription. Finally, we note that, in the same way as what had been done with ICA and sparse coding, unsupervised learning of auditory filterbanks using NMF has also been investigated [27].

We will present in Chapter 5 a general framework, including an extension of SNMF2D, to perform the analysis of acoustical scenes on incomplete data. In Chapter 7, we will develop a shift-invariant semi-NMF algorithm, which can be used to directly model a signal in the time domain as a combination of a limited number of recurring elementary patterns learnt from the data. Finally, in Chapter 8 we will present a data-driven learning of filterbanks based on a modulation energy criterion.

2.4 Summary of Chapter 2

In this chapter, we have presented an overview of theories and methods on the structure of natural sounds and the way to exploit it. We first focused on speech as an intensively-studied example of natural sound, explaining its production mechanism and reviewing an array of methods which have been developed to exploit its statistical regularities. We then briefly presented human auditory functions involved in the analysis of natural scenes and reviewed algorithms which draw on insights on human auditory organization to computationally implement this analysis. Finally, we reviewed concepts and methods on the exploitation of regularities in natural signals, particularly visual and audio. This corpus of works constitutes the conceptual basis for the research we present in this thesis.

Chapter 3

HTC, a statistical model for speech signals in the time-frequency domain

3.1 Introduction

In this chapter, we introduce a statistical model for speech signals designed to capture effectively the constraints a signal is likely to respect to be considered by the human auditory organization process as a harmonic acoustic stream with continuously varying pitch, such as the voiced parts of a speech utterance. The model is part of a more general framework called Harmonic-Temporal Clustering (HTC), first introduced for music signals by Kameoka [106], which represents the wavelet power spectrum of an acoustical scene as a combination of parametric models inherently respecting Bregman's grouping principles such as harmonicity, common onset and offset, coherent amplitude and frequency modulation, continuity of the components in the time and frequency directions, etc.

Many methods utilizing Bregman's grouping cues for the computational implementation of acoustical scene analysis have been proposed, as we saw in Section 2.3, in most of which the grouping process is usually implemented in two steps. Instantaneous features are first extracted at each discrete time point, which corresponds to the grouping process in the frequency direction, and a post-processing is then performed on these features to reduce errors and/or obtain continuous tracks, through hidden Markov model (HMM), multiple agents, or some dynamical system such as Kalman filtering, corresponding to the grouping process in the time direction. Considering that, from an engineering point of view, it should be more efficient to perform the analysis in both time and frequency directions simultaneously, we consider here with HTC, in contrast to the conventional strategy, a unified estimation

framework for the two dimensional structure of time-frequency power spectra.

We explain in this chapter how to extend the HTC framework to model speech signals. In brief, we model the power spectrum $W(x,t)$ as a sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$, where x is log-frequency, t is time and $\mathbf{\Theta}$ is the set of model parameters: $W(x,t) \approx \sum_{k=1}^{K} q_k(x,t;\mathbf{\Theta})$, and apply two constraints to the spectro-temporal model. Along the frequency axis, it consists of a series of harmonics with frequencies multiple of a common F_0. Along the temporal axis, this F_0 follows a contour modeled as a cubic spline, and the amplitude of each partial follows a smooth temporal envelope modeled as a sum of Gaussian functions. The cubic spline F_0 contour was preferred to other options such as the Fujisaki model [82] as it can be applied to a wider range of stimuli in addition to speech, and enables us to obtain analytic update equations during the optimization process. In the original formulation of HTC, each source model represents a sound stream with *constant* F_0. However, by considering speech as a succession of models that each correspond to a phoneme (or more generally to a segment of the speech utterance with steady acoustic characteristics), we can use the HTC method to model the spectrum as a sequence of spectral cluster models with a continuous F_0 contour. Contrary to HMMs which assume discrete phonetic states, we aim here to model smooth transitions in the temporal succession of the spectral structures.

We also introduce a noise model to deal with the non-harmonic power coming from background broadband noise. While the spectrogram of voiced speech is characterized by harmonic parts with strong relative power, noise tends to have a more flat spectrogram. Extracting voiced speech from such noise corresponds to searching for local, harmonically structured "islands" within a "sea" of unstructured noise.

The parametric model we obtain in this way is optimized using a new formulation of the EM algorithm. The spectral clusters are obtained by an unsupervised 2D clustering of the power density, performed simultaneously with the estimation of the F_0 contour of the whole utterance. We first present the general HTC method, then introduce the speech model, and finally the noise model.

3.2 General HTC method

Consider the wavelet power spectrum $W(x,t)$ of a signal recorded from an acoustical scene, defined on a domain of definition $D = \{x,t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$. The problem considered is to approximate the power spectrum as well as possible as the sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$ modeling the power spectrum of K "objects" each with its own F_0 contour $\mu_k(t)$ and its own harmonic-temporal structure. We note that the

formulation described hereafter is not limited to any particular time-frequency representation, and it could work with a standard linear-frequency STFT. However we found better performance with wavelets, possibly because it offers relatively better spectral resolution for the low-frequency harmonics of the voice. We will thus use the wavelet power spectrum in the following.

The source models $q_k(x,t;\mathbf{\Theta})$ are expressed as a Gaussian Mixture Model (GMM) with constraints on the characteristics of the kernel distributions: assuming that there is harmonicity with N partials modeled in the frequency direction, and that the power envelope is described using Y kernel functions in the time direction, we can rewrite each source model in the form

$$q_k(x,t;\mathbf{\Theta}) = \sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\mathbf{\Theta}), \tag{3.1}$$

where $\mathbf{\Theta}$ is the set of all parameters and with kernel densities $S_{kny}(x,t;\mathbf{\Theta})$ which are assumed to have the following shape:

$$S_{kny}(x,t;\mathbf{\Theta}) \triangleq \frac{w_k v_{kn} u_{kny}}{2\pi\sigma_k\phi_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}}, \tag{3.2}$$

where the parameters w_k, v_{kn} and u_{kny} are normalized to unity. A graphical representation of an HTC source model $q_k(x,t;\mathbf{\Theta})$ can be seen in Fig. 3.1.

Our goal is to minimize the difference between $W(x,t)$ and $Q(x,t;\mathbf{\Theta}) = \sum_{k=1}^{K} q_k(x,t;\mathbf{\Theta})$ according to a certain criterion. We use the $\mathcal{I}$-divergence [52] as a classical way to measure the "distance" between two distributions:

$$\mathcal{I}(W|Q(\mathbf{\Theta})) \triangleq \iint_D \left(W(x,t)\log\frac{W(x,t)}{Q(x,t;\mathbf{\Theta})} - \Big(W(x,t) - Q(x,t;\mathbf{\Theta})\Big)\right) dx\,dt, \tag{3.3}$$

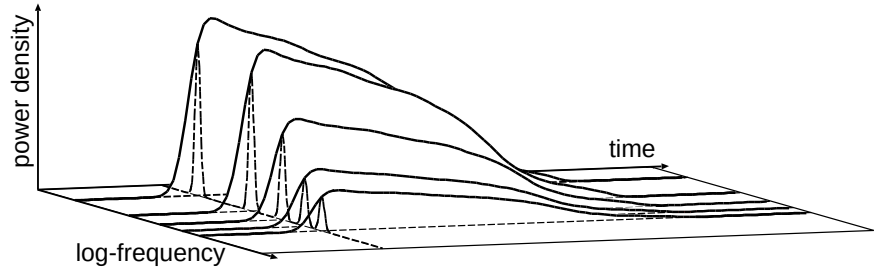
and we are thus looking for $\mathbf{\Theta}_{\text{opt}} = \underset{\mathbf{\Theta}}{\operatorname{argmin}}\, \mathcal{I}(W|Q(\mathbf{\Theta}))$. Keeping only the terms depending on $\mathbf{\Theta}$ and reversing the sign of this expression, one defines the following function to maximize w.r.t. $\mathbf{\Theta}$:

$$\mathcal{J}(W,\mathbf{\Theta}) = \iint_D \Big(W(x,t)\log Q(x,t;\mathbf{\Theta}) - Q(x,t;\mathbf{\Theta})\Big)\, dx\,dt. \tag{3.4}$$

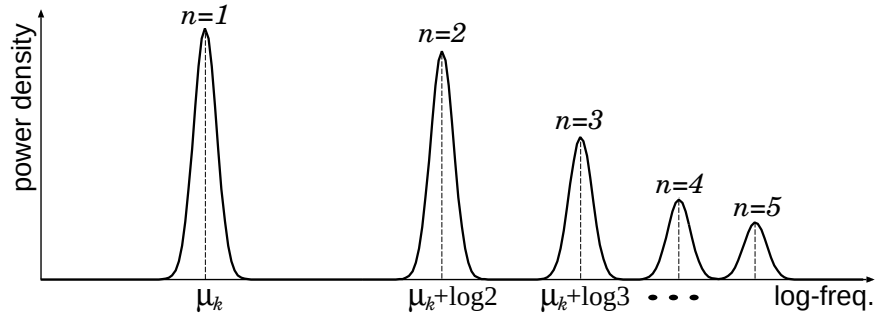
Using this function $\mathcal{J}$, one can derive the likelihood of the parameter $\mathbf{\Theta}$:

$$P(W|\mathbf{\Theta}) \triangleq e^{\mathcal{J}(W,\mathbf{\Theta}) - \iint_D \log\Gamma(1+W(x,t))\,dx\,dt}, \tag{3.5}$$

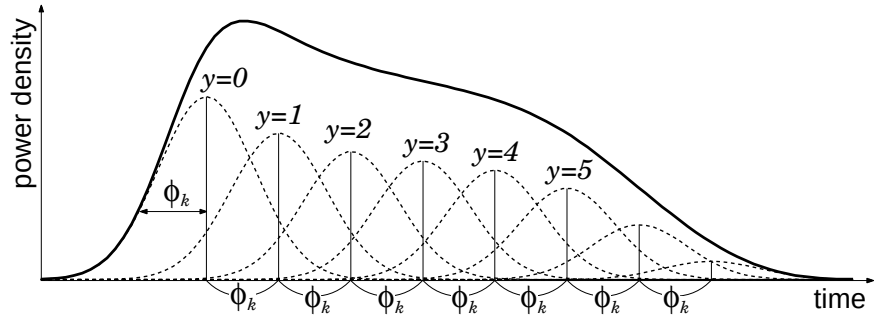
where $\Gamma(\cdot)$ is the Gamma function. If the model and the data are assumed to take integer values, the second part of the exponent ensures that we obtain a probability measure on



(a) Profile of an HTC source model $q_k(x, t; \mathbf{\Theta})$



(b) Cross-section of $q_k(x, t; \mathbf{\Theta})$ at constant time



(c) Power envelope function

Figure 3.1: *Graphical representation of an HTC source model. Fig. (a) shows the time-frequency profile of the model, while Fig. (b) shows a cross-section of the model at constant time and Fig. (c) the evolution in time of the power envelope function. The harmonic structure of the model can be seen in Fig. (b), and the approximation of the power envelope in the time direction as a sum of Gaussian kernels can be seen in Fig. (c).*

W. One can indeed see this probability as the joint probability of all the variables $W(x, t)$ independently following Poisson distributions of parameter $Q(x, t)$. If, as is usually the case, the model and the data are assumed to take continuous real non-negative values, one could formally assume that they have actually been discretized (which happens anyway when handling digital data in computers) and rescaled to justify the above likelihood derivation. Although practical, this formal workaround may however be considered as somewhat theoretically unsatisfactory. We investigate in Appendix A the behavior of the measure defined

by extending the Poisson distribution to the non-negative real numbers using the Gamma function normalization introduced in (3.5), and argue that a more theoretical justification to the likelihood derivation can be obtained in this way. Going back to the likelihood defined in (3.5), this way of presenting the problem enables us to interpret it as a Maximum *A Posteriori* estimation problem and to introduce prior functions on the parameters as follows, using Bayes theorem:

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_{\mathrm{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}|W) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\log P(W|\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\mathcal{J}(W,\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big). \end{aligned} \tag{3.6}$$

Our goal is now equivalent to the maximization with respect to $\boldsymbol{\Theta}$ of $\mathcal{J}(W,\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})$. In the following, we will write simply $\mathcal{J}(\boldsymbol{\Theta})$ for $\mathcal{J}(W,\boldsymbol{\Theta})$. The problem is that there is a sum inside the logarithm in the term

$$\iint_D W(x,t) \log \sum_{k,n,y} S_{kny}(x,t;\boldsymbol{\Theta})\, dx\, dt, \tag{3.7}$$

and we thus cannot obtain an analytical solution. But if we introduce non-negative membership degrees $m_{kny}(x,t)$ summing to 1 for each (x,t), one can write, using the concavity of the logarithm:

$$\begin{aligned}\log \sum_{k,n,y} S_{kny}(x,t;\boldsymbol{\Theta}) &= \log \sum_{k,n,y} m_{kny}(x,t) \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \\ &= \log \left\langle \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \right\rangle_m \\ &\geq \left\langle \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \right\rangle_m \\ &\geq \sum_{k,n,y} m_{kny}(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)}, \end{aligned} \tag{3.8}$$

where $\langle\cdot\rangle_m$ denotes the convex combination with coefficients m. Moreover, the inequality (3.8) becomes an equality for

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}. \tag{3.9}$$

We can thus use an EM-like algorithm to maximize the likelihood by alternately updating $\boldsymbol{\Theta}$

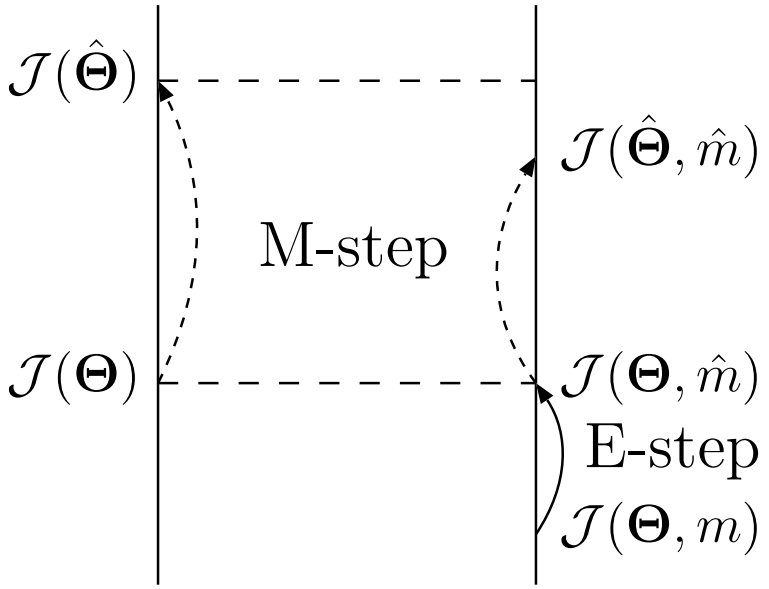


Figure 3.2: *Optimization through the EM algorithm. During the E-step, the auxiliary parameter m is updated to $\hat{m}$ so that $\mathcal{J}(\boldsymbol{\Theta}) = \mathcal{J}(\boldsymbol{\Theta}, \hat{m})$. Then, during the M-step, $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ is optimized w.r.t. $\boldsymbol{\Theta}$, ensuring that $\mathcal{J}(\hat{\boldsymbol{\Theta}}) \geq \mathcal{J}(\hat{\boldsymbol{\Theta}}, \hat{m}) > \mathcal{J}(\boldsymbol{\Theta}, \hat{m}) = \mathcal{J}(\boldsymbol{\Theta})$. The maximization of $\mathcal{J}(\boldsymbol{\Theta})$ can thus be performed through the maximization of the auxiliary function $\mathcal{J}(\boldsymbol{\Theta}, m)$ alternately w.r.t. m and $\boldsymbol{\Theta}$.*

and the membership degrees m, which act as auxiliary parameters, while keeping the other fixed:

$$\textbf{(E-step)} \qquad \hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})},$$

$$\textbf{(M-step)} \qquad \widehat{\boldsymbol{\Theta}} = \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\mathcal{J}(\boldsymbol{\Theta}, \hat{m}) + \log P(\boldsymbol{\Theta}) \Big),$$

with

$$\mathcal{J}(\boldsymbol{\Theta}, m) \triangleq \iint_D \Big(\sum_{k,n,y} \ell_{kny}(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} - Q(x,t;\boldsymbol{\Theta}) \Big)\, dx\, dt, \tag{3.10}$$

where $\ell_{kny}(x,t) = m_{kny}(x,t)W(x,t)$. For all m, we indeed have from (3.8) that

$$\mathcal{J}(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \geq \mathcal{J}(\boldsymbol{\Theta}, m) + \log P(\boldsymbol{\Theta}), \tag{3.11}$$

and $\mathcal{J}(\boldsymbol{\Theta}, m)$ can be used as an auxiliary function to maximize, enabling us to obtain analytical update equations. The optimization process is illustrated in Fig. 3.2.

A slightly different approach was presented in [108], but leads to the same optimization

process as the one we present here. The E-step is straightforward. However, the possibility to obtain analytical update equations during the M-step depends on the actual expression of $\mu_k(t)$, and it could be performed there in the special case of a piece-wise flat F_0 contour $\mu_k(t) = \mu_k$. More generally, if the other HTC parameters (w_k, τ_k, u_{kny}, v_{kn}, ϕ_k, σ_k) do not enter in the expression of $\mu_k(t)$, then the update equations obtained in [108] for these parameters can be used as is, and we only need to obtain new M-step update equations for the F_0 contour parameters. We will explain in the following subsection how we define the F_0 contour model and how we proceed to estimate its parameters, and shall give here, for the sake of completeness, the update equations for the other HTC parameters at step p:

$$w_k^{(p)} = \sum_{n,y} \iint_D \ell_{kny}^{(p)}(x,t)\,dx\,dt, \tag{3.12}$$

$$\tau_k^{(p)} = \frac{1}{w_k^{(p)}} \sum_{n,y} \iint_D (t - y\phi_k^{(p-1)})\ell_{kny}^{(p)}(x,t)\,dx\,dt, \tag{3.13}$$

$$v_{kn}^{(p)} = \frac{1}{w_k^{(p)}} \sum_{y} \iint_D \ell_{kny}^{(p)}(x,t)\,dx\,dt, \tag{3.14}$$

$$u_{kny}^{(p)} = \frac{1}{w_k^{(p)}} \iint_D \ell_{kny}^{(p)}(x,t)\,dx\,dt, \tag{3.15}$$

$$\phi_k^{(p)} = \frac{1}{2w_k^{(p)}} \left(\left(a_k^{(p)2} + 4b_k^{(p)} w_k^{(p)} \right)^{\frac{1}{2}} - a_k^{(p)} \right), \tag{3.16}$$

$$\text{with} \begin{cases} a_k^{(p)} & \triangleq \sum_{n,y} \iint_D y(t - \tau_k^{(p)})\ell_{kny}^{(p)}(x,t)\,dx\,dt, \\ b_k^{(p)} & \triangleq \sum_{n,y} \iint_D (t - \tau_k^{(p)})^2 \ell_{kny}^{(p)}(x,t)\,dx\,dt, \end{cases}$$

$$\sigma_k^{(p)} = \left(\frac{1}{w_k^{(p)}} \sum_{n,y} \iint_D (x - \mu_k^{(p-1)}(t) - \log n)^2 \ell_{kny}^{(p)}(x,t)\,dx\,dt \right)^{\frac{1}{2}}. \tag{3.17}$$

3.3 Speech modeling

In the following, in order to model the spectrum of a speech utterance, we will make several assumptions. First, we assume that the F_0 contour is smooth and defined on the whole interval: we will not make voiced/unvoiced decisions, and F_0 values are assumed continuous. Second, we fix the harmonic structure of each HTC source model so that it corresponds to segments of speech with steady acoustic characteristics, and assume that a speech segment is a succession of such steady segments sharing a common F_0 contour.

3.3.1 Spline F_0 contour

In previous works [108], the HTC method has only been applied to piece-wise flat F_0 contours, which is relevant for certain instruments like the piano for example, but of course not in speech. Looking for a smooth F_0 contour, we chose to use cubic spline functions as a general class of smooth functions, so as to be able to deal in the future with a wide variety of acoustic phenomena (background music, phone ringing, etc). Moreover, their simple algebraical formulation enables us to optimize the parameters through the EM algorithm, as update equations can be obtained analytically. It may happen that the smooth F_0 assumption is invalid, such as at the end of utterances where F_0-halving can occur, but this problem is faced by all algorithms that exploit continuity of F_0, and the assumption is justified empirically in that including it tends to reduce overall error rates. Moreover, failure to track F_0-halving at the end of utterances is perhaps not too serious, as the halving is usually not salient perceptively, other than as a roughness of indeterminate pitch. Furthermore, it is one of a wider class of irregular voicing phenomena (diplophony, creak) for which F_0 is hard to define [92].

The analysis interval is divided into subintervals $[t_i, t_{i+1})$ which are assumed of equal length. Following [156], the parameters of the spline contour model are then the values z_i of the F_0 at each bounding point t_i. The values z_i'' of the second derivative at those points are given by the expression $\mathbf{z}'' = M\mathbf{z}$ for a certain matrix M which can be explicitly computed offline, under the hypothesis that the first-order derivative is 0 at the bounds of the analysis interval. We can assume so if we set the bounds of the interval outside the region where there is speech. One can then classically obtain a simple algebraic formula for the contour $\mu(t;\mathbf{z})$ on the whole interval. For $t \in [t_i, t_{i+1})$:

$$\begin{aligned}\mu(t;\mathbf{z}) \triangleq \frac{1}{t_{i+1}-t_i}\Big(& z_i(t_{i+1}-t) + z_{i+1}(t-t_i) \\ & -\tfrac{1}{6}(t-t_i)(t_{i+1}-t)\big[(t_{i+2}-t)z_i'' + (t-t_{i-1})z_{i+1}''\big]\Big).\end{aligned} \tag{3.18}$$

3.3.2 Optimization of the model

For simplicity, we first describe here the case of a single speaker, and the multiple-speaker case will be presented in 3.3.4.

To design our model, we further make the following assumptions. We make all the HTC source models share the same F_0 contour: $\mu_k(t) = \mu(t), \forall k$, by plugging the analytical expression (3.18) of the spline F_0 contour into (3.2), such that all the source models are

driven by the same F_0 expression. Our intention is to have a succession in time of slightly overlapping source models which correspond if possible to successive phonemes, or at least to segments of the speech utterance with steady acoustic characteristics. As the structure is assumed harmonic, the model takes advantage of the voiced parts of the speech utterance, which it uses as anchors. When used on natural speech, if the unvoiced/silent parts are too long, it may happen that the spline contour becomes unstable, which can deteriorate the accuracy of the F_0 contour extraction immediately before or after a section of unvoiced speech, especially if the neighboring voiced parts are not strongly enough voiced. If they are not, we believe there is no particular incidence on the accuracy of the F_0 contour near unvoiced parts of the speech. The results of the experimental evaluations will show that this assumption is justified.

We also assume that inside a source model the same power envelope is used for all harmonics, as we want to isolate structures in the speech flow with stable acoustic characteristics. The model allows source models to overlap, so a given spectral shape can merge progressively into another, which allows it to fit arbitrary spectro-temporal shapes. The subscript n can thus be excluded in u_{kny}, resulting in an extra summation on n in Eq. (3.15). We note however that the following discussion on the optimization of the model is independent of this assumption, the algorithm being general enough to allow separate power envelope functions.

The optimization process goes as follows: we start by updating the HTC parameters which do not enter in the spline model (namely w_k, τ_k, u_{ky}, v_{kn}, ϕ_k, σ_k) through the analytical update equations given in Eq. (3.12)–(3.17). Once these parameters have been updated, we compute the derivatives of $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ with respect to the spline parameters $\mathbf{z} = (z_0, \ldots, z_J)$. We then update the z_j's one after the other, starting for example from z_0 and using the already updated parameters for the update of the next one. This way of performing the updates is referred to as the coordinate descent method [214], and can be summarized as in Eq. (3.19):

$$\begin{cases} z_0^{(p)} \leftarrow \underset{z_0}{\operatorname{argmax}}\, \mathcal{J}(z_0, z_1^{(p-1)}, \ldots, z_J^{(p-1)}, \boldsymbol{\Theta}_{-\mathbf{z}}, \hat{m}), \\ z_1^{(p)} \leftarrow \underset{z_1}{\operatorname{argmax}}\, \mathcal{J}(z_0^{(p)}, z_1, z_2^{(p-1)}, \ldots, z_J^{(p-1)}, \boldsymbol{\Theta}_{-\mathbf{z}}, \hat{m}), \\ \quad \vdots \\ z_J^{(p)} \leftarrow \underset{z_J}{\operatorname{argmax}}\, \mathcal{J}(z_0^{(p)}, \ldots, z_{n-1}^{(p)}, z_J, \boldsymbol{\Theta}_{-\mathbf{z}}, \hat{m}), \end{cases} \tag{3.19}$$

where $\boldsymbol{\Theta}_{-\mathbf{z}}$ denotes the set of all the parameters except $\mathbf{z}$. The corresponding optimization

procedure, called the Expectation-Constrained Maximization algorithm (ECM) [138], does not ensure the maximization in the M-step but guarantees the increase of the function $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ (we will see in Section 5.5.2 how to derive analytical updates which guarantee to reach the maximum at each iteration). Putting the derivatives with respect to z_j to 0, one then finds update equations analytically at step p:

$$z_j^{(p)} = \frac{\sum\limits_{k,n,y} \iint_D \Big(x - \hat{\mu}_j^{(n)}\big(t; \mathbf{z}^{(j,p)}\big)\Big) \dfrac{\partial \mu}{\partial z_j}(t) \dfrac{\ell_{kny}^{(p-1)}(x,t)}{{\sigma_k^{(p)}}^2}\, dx\, dt}{\sum\limits_{k,n,y} \iint_D \left(\dfrac{\partial \mu}{\partial z_j}(t)\right)^2 \dfrac{\ell_{kny}^{(p-1)}(x,t)}{{\sigma_k^{(p)}}^2}\, dx\, dt} \tag{3.20}$$

where

$$\mathbf{z}^{(j,p)} = \big(z_0^{(p)}, \ldots, z_{j-1}^{(p)}, z_j^{(p-1)}, z_{j+1}^{(p-1)}, \ldots, z_J^{(p-1)}\big)$$

and

$$\hat{\mu}_j^{(n)}\big(t; \mathbf{z}^{(j,p)}\big) = \mu\big(t; \mathbf{z}^{(j,p)}\big) - \frac{\partial \mu}{\partial z_j}(t) z_j^{(p)} + \log n.$$

$\mathbf{z}^{(j,p)}$ does not depend on z_j and $\frac{\partial \mu}{\partial z_j}(t)$ only depends on t and the fixed matrix M.

3.3.3 Prior distribution

As seen in 3.2, the optimization of our model can be naturally extended to a Maximum *A Posteriori* (MAP) estimation by introducing prior distributions $P(\boldsymbol{\Theta})$ on the parameters, which work as penalty functions that try to keep the parameters within a specified range. The parameters which are the best compromise with empirical constraints are then obtained through Eq. (3.6).

By introducing such a prior distribution on v_{kn}, it becomes possible to prevent half-pitch errors, as the resulting source model would usually have a harmonic structure with zero power for all the odd order harmonics, which is abnormal for speech. We apply the Dirichlet distribution, which is explicitly given by:

$$p(\mathbf{v}_k) \triangleq \frac{\Gamma\Big(\sum_n (d_v \bar{v}_n + 1)\Big)}{\prod_n \Gamma(d_v \bar{v}_n + 1)} \prod_n {v_{kn}}^{d_v \bar{v}_n}, \tag{3.21}$$

where $\bar{v}_n$ is the most preferred 'expected' value of v_{kn} such that $\sum_n \bar{v}_n = 1$, d_v the contribution degree of the prior and $\Gamma(\cdot)$ the Gamma function. The maximum value for $p(\mathbf{v}_k)$ is taken when $v_{kn} = \bar{v}_n$ for all n. When d_v is zero, $p(\mathbf{v}_k)$ becomes a uniform distribution. The choice of this particular distribution allows us to give an analytical form of the update equations

of v_{kn}, which become

$$v_{kn}^{(p)} = \frac{1}{d_v + w_k^{(p)}} \left(d_v \bar{v}_n + \sum_y \iint_D \ell_{kny}^{(p)}(x,t) dx dt \right). \tag{3.22}$$

Although the spline model can be used as is, one can also introduce in the same way a prior distribution on the parameters z_j of the spline F_0 contour, in order to avoid an overfitting problem with the spline function. Indeed, spline functions have a tendency to take large variations, which is not natural for the F_0 contour of a speech utterance. Moreover, the F_0 contour might also be hard to obtain on voiced parts with relatively lower power, or poor harmonicity. The neighboring voiced portions with higher power help the estimation over these intervals by providing a good prior distribution.

To build this prior distribution, we assume that the z_j form a Markov chain, such that

$$P(z_0, \ldots, z_J) = P(z_0) \prod_{j=1}^{J} P(z_j | z_{j-1}),$$

and assume furthermore that z_0 follows a uniform distribution and that, conditionally to z_{j-1}, z_j follows a Gaussian distribution of center z_{j-1} and variance σ_s^2 corresponding to the weighting parameter of the prior distribution:

$$P(z_j | z_{j-1}) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(z_j - z_{j-1})^2}{2\sigma_s^2}}.$$

In the derivative with respect to z_j used above to obtain (3.20) add up two new terms

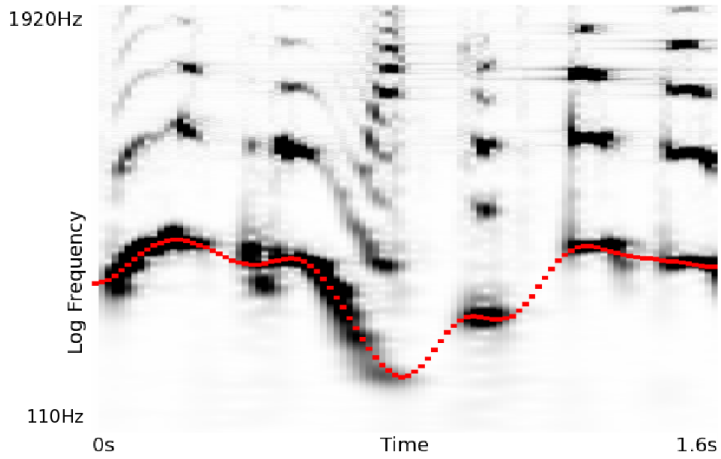
$$\frac{\partial \log P(z_j | z_{j-1})}{\partial z_j} + \frac{\partial \log P(z_{j+1} | z_j)}{\partial z_j},$$

and the update equation (3.20) then becomes

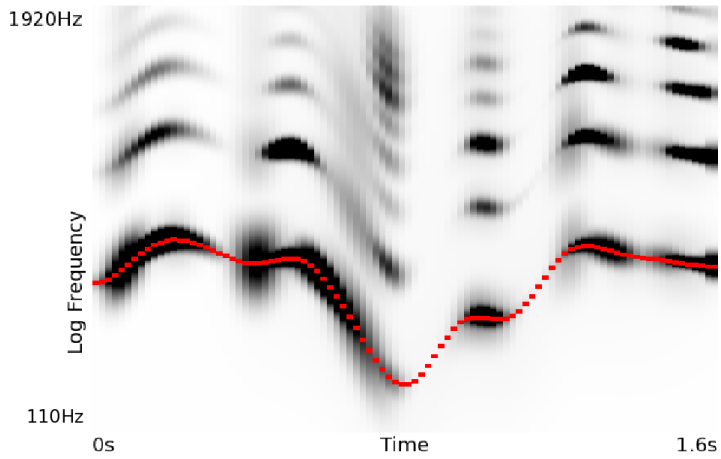
$$z_j^{(p)} = \frac{\dfrac{2}{\sigma_s^2} \cdot \dfrac{z_{j-1}^{(p)} + z_{j+1}^{(p-1)}}{2} + A_j^{(p)}}{\dfrac{2}{\sigma_s^2} + B_j^{(p)}}, \tag{3.23}$$

where $A_j^{(p)}$ and $B_j^{(p)}$ are respectively the numerator and denominator of the right-hand side term of equation (3.20). The update equation for the boundary points is derived similarly.

An example is presented in Fig. 3.3, based on the Japanese sentence "Tsuuyaku denwa kokusai kaigi jimukyoku desu" uttered by a female speaker. Shown are 2D representations of the observed and modeled spectra (after 30 iterations of the estimation algorithm). The F_0 contour estimated through our method is reproduced on both the observed and modeled



(a) Observed spectrogram and estimated F_0 contour



(b) Modeled spectrogram and estimated F_0 contour

Figure 3.3: *Comparison of observed and modeled spectra ("Tsuuyaku denwa kokusai kaigi jimukyoku desu", female speaker). The estimated F_0 contour is reproduced on both the observed and modeled spectrograms to show the precision of the algorithm.*

spectrograms to show the precision of our algorithm. One can see that the model approximates well the spectrum and that F_0 contour is accurately estimated.

3.3.4 Multiple speakers

The multiple-speaker case is a simple extension of the single-speaker one. While for a single F_0 estimation all the source models share the same F_0 contour, for multiple speakers, according to the number of F_0 contours that we want to estimate, we group together source models into subsets such that source models inside a subset share a common F_0 contour. For example, if we use $K = 10$ models in total for two speakers, the models with index $k \in \{1, \ldots, 5\}$ will be attributed to the first speaker, while the others will be attributed to the second one. We thus have pools of source models driven by a single F_0 contour for each

of the pools and corresponding to one of the speakers, and we only need to introduce a set of spline parameters for each of the F_0 contours. These sets can be optimized independently and simultaneously in the exact same way as in the single-speaker case: the E-step is unchanged, and the M-step is performed by first updating the HTC parameters which do not enter in the spline model, and then updating the spline parameters of each F_0 contour through the ECM algorithm. This last update can be done independently as the derivatives of $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ with regards to the parameters of one of the F_0 contours do not include any term depending on the parameters of the other contours.

The method handles overlapping harmonics by, at each iteration of the algorithm, re-estimating simultaneously the contributions of each voice in the spectrum. It relies on the assumption that the spectra of the contributing sources combine additively, and neglects phase-dependent vector summation. This is of course a rough approximation in the view of perfect source separation, but as the speech spectrum is usually relatively sparse, in a mixture of two speech signals, regions of the time-frequency plane with non-zero energy for each signal will most likely not overlap, in which case the cross-terms in the power are indeed truly equal to zero. We shall note that the additivity assumption is then, however, still only as true as the sparseness assumption is.

3.4 Noise modeling

We introduce a noise model to cope with the background noise that can be a disturbance in the process of clustering the harmonic portions of speech. Indeed, it would be more rewarding, in the purpose of decreasing the $\mathcal{I}$-divergence, for the harmonic source models to take very large variances in the log-frequency direction and have the centers of the Gaussian distribution go on portions of the spectrogram with strong broad power, even though there is no harmonic structure corresponding to these portions, especially if the noise power is comparable to or even larger than the speech signal power.

The spectrogram of a quasi-periodic signal such as voiced speech consists of a large number of line spectrum components and has spikes that are strongly marked, while the spectrogram of a white or pink noise is more flat, without significant spiky regions. The idea to design the noise model was thus that detecting the harmonic parts of the spectrogram in a noisy background corresponds to searching for thin and harmonically distributed "islands" which rise out of a "sea" of noise. We thus chose to model the noise using a mixture of Gaussian distributions with large fixed variance and with centers fixed on a grid, the only parameters of the model being the respective weights of the Gaussians in the model and the ratio of noise

power inside the whole spectrogram. This noise-cancelling approach can be considered quite close to spectral subtraction in the sense that the power spectrum is in both cases assumed additive. However, while spectral subtraction generally needs voiced/unvoiced decision to obtain the power spectrum of the background noise, which furthermore has to be assumed stationary, our approach estimates adaptively the noise part of the spectrum even when there is speech, taking advantage of the valleys of the spectral structure. The only assumption we make is that the noise spectrum is smooth in both time and frequency directions whereas speech spectrum is more spiky in the frequency direction. Therefore, we can expect our model's performance to be close to the best performance that spectral subtraction could reach.

Let N_c be the number of columns of the grid (in the time direction) and N_r the number of rows (in the log-frequency direction). The noise model is defined as

$$\mathcal{N}(x,t;\mathbf{\Theta}) = \rho \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} w_{ny}^{\mathcal{N}} \frac{1}{2\pi\sigma_r^{\mathcal{N}}\sigma_c^{\mathcal{N}}} e^{-\frac{(x-\alpha_n)^2}{2(\sigma_r^{\mathcal{N}})^2}} e^{-\frac{(t-\beta_y)^2}{2(\sigma_c^{\mathcal{N}})^2}}, \tag{3.24}$$

where ρ is the proportion of the noise model in the total model, the $w_{ny}^{\mathcal{N}}$ are the weights of the Gaussian functions in the mixture and add up to 1, $\sigma^{\mathcal{N}} = (\sigma_r^{\mathcal{N}}, \sigma_c^{\mathcal{N}})$ the variances of the Gaussian functions, α_n the log-frequency index of the centers of the n-th row and β_y the time index of the centers of the y-th column.

If we note $\mathcal{N}(x,t;\mathbf{\Theta}) = \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} S_{0ny}(x,t;\mathbf{\Theta})$ and introduce $\ell_{0ny}(x,t)$ and $m(0,n,y;x,t)$ as in 3.2, the EM algorithm can be applied in the same way as described above, just differing in the range of the summations on k, n and y. We only need to specify the update equations of the M-step for the noise model parameters:

$$\rho^{(p)} = \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} \iint_D \ell_{0ny}^{(p)}(x,t)\,dx\,dt, \tag{3.25}$$

$$w_{ny}^{\mathcal{N}\,(p)} = \frac{1}{\rho^{(p)}} \iint_D \ell_{0ny}^{(p)}(x,t)\,dx\,dt, \tag{3.26}$$

where the superscript (p) refers to the iteration cycle. The update equations for the other parameters remain the same, the only changes coming from the E-step where the noise model now enters in the summation.

The noise cancelling performed through the introduction of the noise model is illustrated in Fig. 3.4. Fig. 3.4 (a) gives a 3D view of the original clean spectrogram of a part of the sentence "It was important to be perfect since there were no prompts" uttered by a female speaker, and Fig. 3.4 (b) shows the spectrogram of the same part of the utterance to

which white noise at an SNR of -2 dB has been added. The estimation of the spectrogram where the noise has been cancelled, shown in Fig. 3.4 (c), is obtained through the optimized masking functions $\hat{m}_{kny}(x,t)$ as in the following formula:

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\hat{m}_{kny}(x,t)W(x,t). \tag{3.27}$$

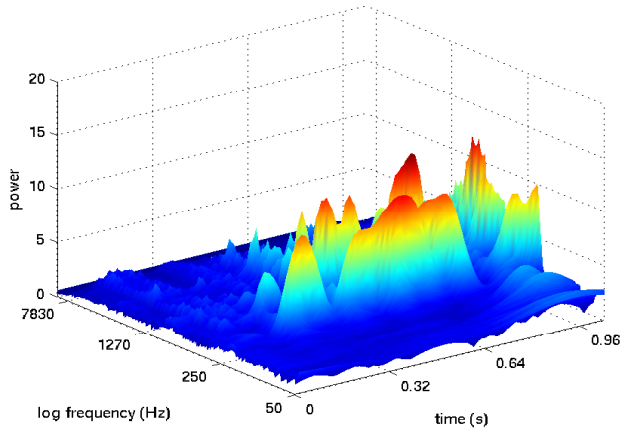
We can notice from the definition of the masking functions that this amounts to performing a Wiener filtering of the signal. In the course of the optimization of the model, the F_0 contour estimation is performed on this "cleaned" part of the spectrogram, which enables the F_0 estimation to perform well even in very noisy environments, as we will show in the next chapter.

3.5 Parametric representation and applications

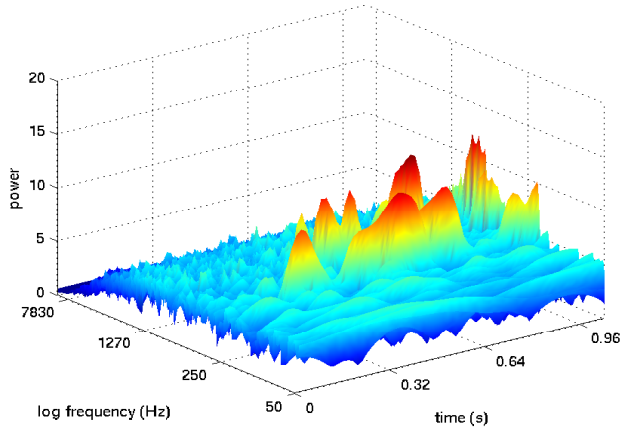
The algorithm we presented not only estimates the F_0 contour, but also more generally gives a parametric representation of the voiced parts of the spectrogram. This can be useful especially in the analysis of co-channel speech by multiple speakers, as one can get a parametric representation of the harmonic parts of the separated spectrograms of each utterance, as shown in Fig. 3.5: Fig. 3.5 (b) represents the modeled spectrogram of the Japanese utterance "oi wo ou" by a male speaker, and Fig. 3.5 (c) the one of the utterance "aoi" by a female speaker, extracted from the mixed signal shown in Fig. 3.5 (a) (the F_0 contour near the boundary is not relevant as there is no sound by the second speaker there). These parametric representations can be used for example to cluster the spectrogram of the mixed sound and separate the speakers, as well as for noise cancelling, as we showed in Fig. 3.4 and shall investigate in more detail in Chapter 4.

3.6 Summary of Chapter 3

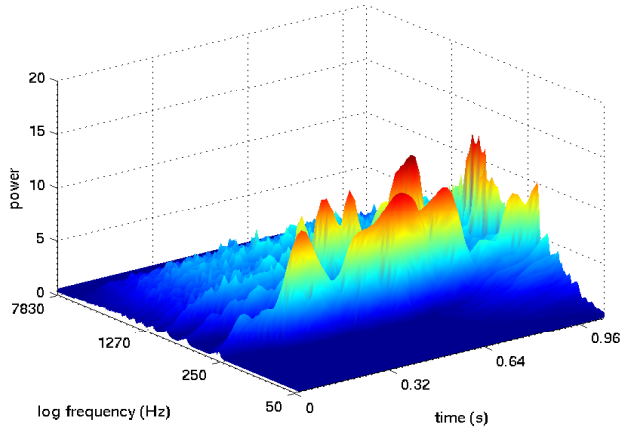
We introduced a new model describing the spectrum as a sequence of spectral cluster models governed by a common F_0 contour function, with smooth transitions in the temporal succession of the spectral structures. The model enables an accurate estimation of the F_0 contour on the whole utterance by taking advantage of its voiced parts in clean as well as noisy environments, and gives a parametric representation of the voiced parts of the spectrogram. We explained how to optimize its parameters efficiently, and shall now perform in the next chapter several experiments to evaluate its performance for F_0 estimation, speech enhancement, and speaker separation.



(a) Original clean spectrogram

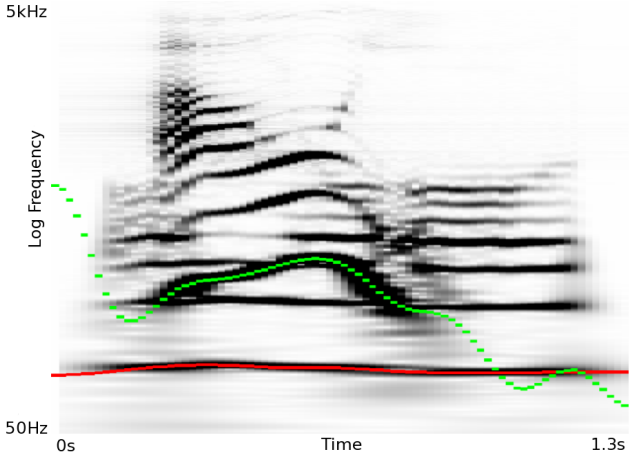


(b) Noisy spectrogram (clean speech mixed with white noise, SNR = −2 dB)

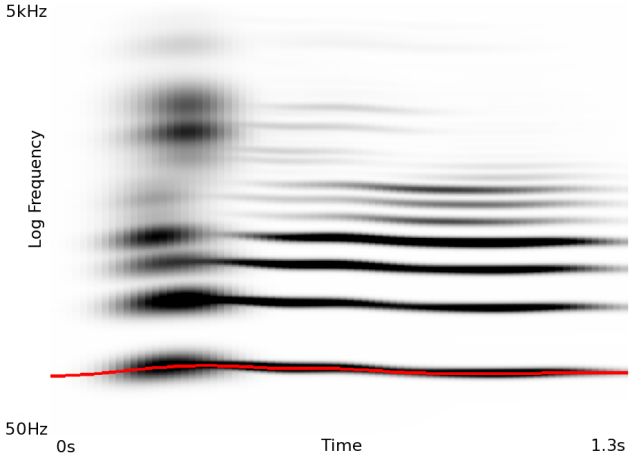


(c) Estimated noise-cancelled part of the spectrogram

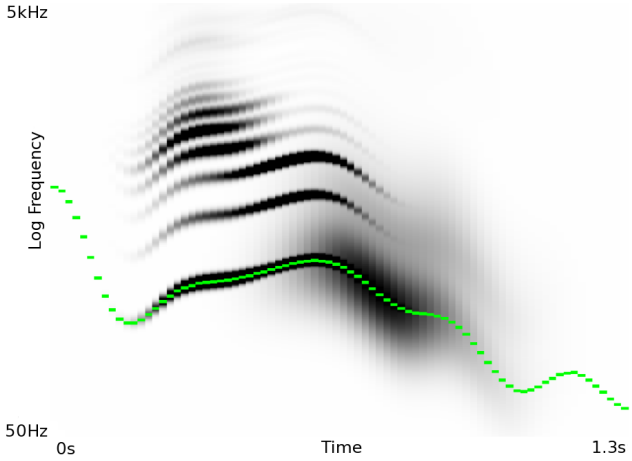
Figure 3.4: *Estimation of the clean part of a noisy spectrogram. Fig. (a) shows a 3D view of the original clean spectrogram of a part of the sentence "It was important to be perfect since there were no prompts" uttered by a female speaker. Fig. (b) shows the spectrogram of the same part of the utterance to which white noise at an SNR of −2 dB has been added. Fig. (c) shows the estimated noise-cancelled part of the spectrogram of Fig. (b).*



(a) Original Spectrogram and estimated F_0 contour



(b) Modeled spectrogram, speaker 1



(c) Modeled spectrogram, speaker 2

Figure 3.5: *Parametric representation of separated spectrograms. Fig. (a) shows the spectrogram of a signal obtained by mixing the two Japanese utterances "oi wo ou" by a male speaker and "aoi" by a female speaker, together with the F_0 contours estimated by our method. Fig. (b) and Fig. (c) show the parametric representations of the spectrograms of the utterances by the male and female speaker respectively, extracted from the mixed signal shown in Fig. (a).*

Chapter 4

Scene analysis through HTC: F_0 estimation, speech enhancement and speaker separation

4.1 Introduction

The design of an effective method for the analysis of complex and varied acoustical scenes is a very important and challenging problem. Many applications, such as automatic speech recognition (ASR) or speaker identification, would for example benefit from the ability of such a system to reduce acoustic interferences which often occur simultaneously with speech in real environments. Being able to locate and extract a portion of an acoustical scene, or on the contrary to cancel it, would also lead to very appealing applications such as instrument separation inside a multipitch track, background music recovery or voice activity detection (VAD). Although there exist general methods for signal separation or enhancement in multisensor frameworks, based for example on independent component analysis or spatial filtering, single-channel solutions are necessary for many applications, such as in telecommunication, analysis of monaural CD recordings, automatic news search or background music determination in television programs, for example. Implementing the single-channel separation problem in computers has proven to be extremely challenging.

As fundamental frequency is one of the most prominent cues in speech communication and in our perception of our auditory environment, determining precisely and robustly the F_0 contour of harmonic signals such as speech is an important step in the development of applications related to the analysis of acoustical scenes. So far, many pitch determination

algorithms (PDAs) have been proposed [98], some of them with very good accuracy [44]. However, they are usually limited to the clean speech of a single speaker and fail in moderate amounts of background noise or the presence of other speakers. Ideally, the performance of a PDA should stay high in as wide a range of background noises as possible (white noise, pink noise, noise bursts, music, other speech...). Furthermore, the possibility to extract simultaneously the F_0 contours of several concurrent voices is also a desirable feature. Several PDAs already exist that deal with the tracking of multiple F_0s [43, 91, 114, 192, 212]. Several of these algorithms rely on an initial frame-by-frame analysis followed by post-processing to reduce errors and obtain a smooth F_0 contour, for example using hidden Markov models (HMM) (see [43] for a review). In contrast, the method we propose, HTC, can be interpreted as simultaneously performing estimation of F_0 and model-based interpolation and looking for a joint optimum, which is expected to be more effective in an engineering perspective. Indeed, as explained in the previous chapter, HTC performs the analysis of an acoustical scene by fitting a parametric model to its power spectrum on the whole time-frequency plane, looking for a global optimum.

This remark is more generally true for source separation algorithms. Most of the recent attempts to tackle the problem of monaural source separation belong to a field called computational auditory scene analysis (CASA), which we briefly reviewed in Section 2.3, and try to solve the separation problem by implementing in computers the remarkable separation ability of the human auditory system. As we noted in Chapter 3, most CASA methods inspired by the auditory organization process described by Bregman implement the grouping process in two steps: extraction of instantaneous features at each discrete time point, followed by a post-processing in the time direction. Although the design of HTC is also inspired from auditory organization as the model is built to respect Bregman's grouping cues, one of its main differences with these methods is that it relies on a global parametric representation of the acoustical scene, and that the grouping process in time and frequency directions can be considered to be performed simultaneously in a joint optimization process.

In this chapter, we show that the HTC framework introduced in Chapter 3 is well-suited to perform the analysis of complex acoustical scenes in a wide range of noisy environments. We first show that it outperforms state-of-the-art methods for the F_0 estimation of single-speaker speech in noisy environments and for multi-pitch estimation of concurrent speech. Then, we show how it can be used to perform single-channel speech signal processing tasks such as speech enhancement, background retrieval, and speaker separation.

4.2 F_0 estimation

As the F_0 contour is a fundamental component of the HTC model, fitting HTC models onto an acoustical scene inherently performs F_0 estimation of the components of the scene. We evaluate here the performance of the HTC framework in various situations.

4.2.1 Single-speaker F_0 estimation in clean environment

We evaluated the accuracy of the F_0 contour estimation of our model on a database of speech recorded together with a laryngograph signal by Bagshaw [15], consisting of one male speaker and one female speaker who each spoke 50 English sentences for a total of 0.12 h of speech, for the purpose of evaluation of F_0-estimation algorithms.

The power spectrum $W(x,t)$ was calculated from an input signal digitized at a 16 kHz sampling rate (the original data of the database was converted from 20 kHz to 16 kHz) using a Gabor wavelet transform with a time resolution of 16 ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The lower bound of the frequency range and the frequency resolution were respectively 50 Hz and 14 cent. The spline contour was initially flat and set to 132 Hz for the male speaker and 296 Hz for the female speaker. These values were tuned in frequency bins after a few preliminary experiments. We shall note that, as we were able to obtain very good performance with the same initial conditions on various data, we believe that the performance is not very sensitive to the priming of the F_0 contour parameters. The length of the interpolation intervals was fixed to 4 frames. For HTC, we used $K = 10$ source models, each of them with $N = 10$ harmonics. This is enough for F_0 estimation. For a better modeling of the spectrogram, one can use 40 or 60 harmonics for example. Temporal envelope functions were modeled using $Y = 3$ Gaussian kernels. The parameters w_k were set to $1/K$, u_{ky} to $1/Y$, τ_k to $T_0 + (k-1)T/K$, ϕ_k to 32 ms and σ_k to 422 cents. For the prior functions, σ_s was fixed to 0.4, d_v to 0.04 and $(\bar{v}_n)_{1\leq n\leq N} = \frac{1}{N}(8,8,4,2,1,\ldots,1)$. The algorithm was run on the utterances from which the initial and final silence parts were manually removed. We note that the minimum time window on which the algorithm can work is one frame, in which case the algorithm works frame-by-frame. This gives acceptable results as described in [107], but performance improves with longer windows, as documented in [108]. It is thus used here on the whole time interval. The computational cost is also discussed in [108].

We used as ground truth the F_0 estimates and the reliability mask derived by de Cheveigné et al. [44] under the following criteria: (1) any estimate for which the F_0 estimate was

Table 4.1: *Gross error rates for several F_0 estimation algorithms on clean single-speaker speech*

Method	Gross error (%)
pda	19.0
fxac	16.8
fxcep	15.8
ac	9.2
cc	6.8
shs	12.8
acf	1.9
nacf	1.7
additive	3.6
TEMPO	3.2
YIN	1.4
HTC (proposed)	3.5

obviously incorrect was excluded and (2) any remaining estimate for which there was evidence of vocal fold vibration was included. Frames outside the reliability mask were not taken into account during our computation of the accuracy, although our algorithm gives values for every point of the analysis interval by construction. As the spline function gives an analytical expression for the F_0 contour, we compare our result with the reference values at a sampling rate of 20 kHz although all the analysis was performed with a time resolution of 16 ms.

Deviations over 20% from the reference were deemed to be gross errors. The results can be seen in Table 4.1, with for comparison the results obtained by de Cheveigné et al. [44] for several other algorithms. Notations stand for the method used, as follows: **ac**: Boersma's autocorrelation method [31, 32], **cc**: cross-correlation [32], **shs**: spectral subharmonic summation [32, 97], **pda**: eSRPD algorithm [15, 216], **fxac**: autocorrelation function (ACF) of the cubed waveform [217], **fxcep**: cepstrum [217], **additive**: probabilistic spectrum-based method [67], **acf**: ACF [44], **nacf**: normalized ACF [44], **TEMPO**: the TEMPO algorithm [112], **YIN**: the YIN algorithm [44]. More details concerning these algorithms can be found in [44]. We can see that our model's accuracy for clean speech is comparable to the best existing single-speaker F_0 extraction algorithms designed for that purpose.

4.2.2 Single F_0 estimation on speech mixed with white and pink noise

We performed F_0 estimation experiments on speech to which a white noise, band-passed

Table 4.2: *Accuracy (%) of the F_0 estimation of single-speaker speech mixed with white and pink noises*

	HTC (YIN,WWB)			
	White noise			Pink noise
	SNR = 0 dB	SNR = −2 dB	SNR = −10 dB	SNR = −2 dB
Female speaker	95.8 (83.5, 56.1)	96.3 (77.8, 48.8)	88.2 (36.7, 09.0)	91.9 (46.1, 44.6)
Male speaker	92.1 (82.5, 69.3)	92.2 (77.2, 59.2)	79.7 (41.5, 19.2)	74.0 (58.1, 37.6)
Total	94.0 (83.0, 62.5)	94.3 (77.5, 53.8)	84.1 (39.0, 13.9)	83.2 (51.9, 41.2)

between 50 Hz and 3300 Hz, was added, with SNRs of 0 dB, −2 dB and −10 dB. These SNRs were selected because 0 dB corresponds to equal power, −2 dB is used in a study [48] performing evaluation on the same database as we use in 4.2.3 and −10 dB is a relatively noisy condition to illustrate the effectiveness of our algorithm in that case. The database mentioned above [15] was again used, and the white noise added was generated independently for each utterance. We also performed experiments with pink noise, band-passed between 50 Hz and 3300 Hz, with an SNR of −2 dB. The spectrum of pink noise is closer to that of speech than white noise. The noise model was initialized with $\rho = 0.1$ and the $w_{ny}^{\mathcal{N}}$ all equal to $1/(N_c N_r)$, while the variances were fixed to $\sigma_r^{\mathcal{N}} = 1120$ cent and $\sigma_c^{\mathcal{N}} = \sigma_r^{\mathcal{N}}/3$, and the centers (β_y, α_n) of the Gaussian distributions of the noise model were fixed on a grid such that the distances between them in the time and log-frequency directions were all equal to $\sigma_c^{\mathcal{N}}$ and $\sigma_r^{\mathcal{N}}$ respectively, to ensure a good overlap between the Gaussian distributions. The determination of the variances was made after a few experiments while keeping in mind that $\sigma_r^{\mathcal{N}}$ should be significantly larger than the typical variance in the log-frequency direction of the Gaussian of the harmonic model but small enough to still be able to model fluctuations in the noise power.

As a comparison, we present results obtained on the same database using YIN [44] and the algorithm of Wu, Wang and Brown [212], specifically designed for F_0 tracking in a noisy environment, and that can also handle the estimation of two simultaneous F_0s. Their code is made available on the Internet by their respective authors. The algorithm of Wu, Wang and Brown will be referred to as the WWB algorithm. According to its authors, the parameters of this algorithm could be tuned on a new database to obtain the best performances, but they mention [212] that it is supposed to work fine in the version made available (trained on a corpus [48] that we will use later).

We obtained good results, presented in Table 4.2, showing the robustness of our method

Table 4.3: *Categorization of interference signals*

	Interference signals
Category 1	White noise, noise bursts
Category 2	1 kHz tone, "cocktail party" noise, rock music, siren, trill telephone
Category 3	Female utterance 1, male utterance, female utterance 2

on noisy speech, when noise is not harmonic. Note that the level of the noise added is greater than that of the original signal, and that at -10 dB it is even difficult for human ears to follow the pitch of the original signal. The harmonic structure of our model is effective for detecting speech in the presence of background noise. YIN and the WWB algorithm were both outperformed, although we should note again that their code was used as is, whereas ours was developed with the task in mind. Thus, this comparison may not do them full justice.

4.2.3 Validation on a corpus of speech mixed with a wide range of interferences

In order to show the wide applicability of our method, we also performed experiments using a corpus of 100 mixtures of voiced speech and interference [48], commonly used in CASA research. In [212], half of the corpus is used for model parameter estimation and the other half for system evaluation. As it is not specified in that paper which part of the corpus was used for which purpose, we decided to use the full corpus as the evaluation set for comparison of the algorithms, which can only be an advantage for the WWB algorithm. The results we present for the WWB algorithm differ from the ones given in [212] as the criterion we use is different. To be able to compare it with our method, which does not perform a voiced/unvoiced decision, we do not take into account errors on the estimation of the number of F_0s, but only look at the accuracy of the output of the pitch determination algorithm. Moreover, we focus on the F_0 estimation of the main voiced speech, as we want here to show that our algorithm robustly estimates the F_0 in a wide range of noisy environments. The ten interferences are grouped into three categories: 1) those with no pitch, 2) those with some pitch qualities, and 3) other speech, as shown in Table 4.3. The reference F_0 contours for the ten voiced utterances were built using YIN on the clean speech and manually corrected.

The experiments were performed in the same conditions as described in 4.2.2 for HTC, and the results are presented in Table 4.4. One can see that our algorithm again outperforms YIN and the WWB algorithm in all the interference categories.

Table 4.4: *Accuracy (%) of the F_0 estimation of single-speaker voiced speech with several kinds of interferences*

	HTC	WWB	YIN
Category 1	99.7	90.8	93.1
Category 2	98.6	96.1	75.7
Category 3	99.5	97.8	87.1

Table 4.5: *F_0 estimation of concurrent speech by multiple speakers, gross error for a difference with the reference higher than* 20 % *and* 10 %

Gross error threshold	20 %		10 %	
	HTC	WWB	HTC	WWB
Male-Female	93.3	81.8	86.8	81.5
Male-Male	96.1	83.4	87.9	69.0
Female-Female	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

4.2.4 Multi-pitch estimation

We present here results on the estimation of the F_0 contour of the co-channel speech of two speakers speaking simultaneously with equal average power. We used again the database mentioned above [15], and produced a total of 150 mixed utterances, 50 for each of the "male-male", "female-female" and "male-female" patterns, using each utterance only once and mixing it with another such that two utterances of the same sentence were never mixed together. We used our algorithm in the same experimental conditions as described in 4.2.1 for clean single-speaker speech, but using two spline F_0 contours. The spline contours were initially flat and set to 155 Hz and 296 Hz in the male-female case, 112 Hz and 168 Hz in the male-male case, and 252 Hz and 378 Hz in the female-female case.

The evaluation was done in the following way: only times inside the reliability mask of either of the two references were counted; for each reference point, if either one of the two spline F_0 contours lies within a criterion distance of the reference, we considered the estimation correct. We present scores for two criterion thresholds: 10 % and 20 %. For comparison, tests using the WWB algorithm [212] introduced earlier were also performed, using the code made available by its authors. YIN could not be used as it does not perform multi-pitch estimation. Results summarized in Table 4.5 show that our algorithm outperforms the WWB algorithm on this experiment.

4.3 Spectrogram modification

Speech enhancement, background retrieval and speaker separation can be performed very simply through HTC, based on the classic idea of masking function, often used in CASA oriented methods.

4.3.1 Ratio of HTC models

When the noise to be cancelled is broadband, an estimation of the spectrogram where the interference has been cancelled can be obtained from the original spectrogram by looking, at each point (x, t), at the proportion of the "clean" part inside the whole parametric model. In the same way, in the multiple-speaker case, when several speech models are used simultaneously, the ratio of one speech model inside the whole at each time-frequency bin can be used as a mask. This corresponds to the E-step in the EM algorithm, and can be easily seen to be equivalent to performing a Wiener filtering of the input. In the course of the optimization of the model, it is actually on this "cleaned" part of the spectrogram that the F_0 contour estimation is performed, during the M-step, enabling the F_0 estimation to perform well even in very noisy environments or within multiple speakers, as we showed in the previous section.

4.3.2 Direct use of HTC models

However, when the noise to be cancelled is not a broadband noise, we can expect that the noise model will not be able to cope totally with the background noise. It might thus as well happen that the total power of the noise model is very low in absence of broadband noise, and using the ratio of speech model in the whole would then make less sense, as this ratio would almost always be close to 1, thus leading to a very ineffective mask. Introducing noise models to cope with more types of background noise could be a way to deal with that problem, but it is limited by several problems: background noises could be of infinitely many kinds, and the multiplication of models would lead to a larger computation cost, and might also conflict with the estimation of the speech model. It might thus be simpler and more effective to look directly at the estimated speech model itself to build a masking function. The speech model has been designed to encompass the acoustic characteristics of speech, and has by construction a harmonic structure.

To use the speech model as a mask, we use a "filtered" version of the speech model which

broadens its peaks:

$$\tilde{Q}(x,t) = \frac{1}{1 + (\frac{\epsilon}{Q(x,t)})^p}, \tag{4.1}$$

where Q is the speech model, normalized such that its maximum is 1, and ϵ a small constant, typically between 10^{-3} and 10^{-1}.

The interference can also be retrieved using a masking function obtained through HTC. We simply consider $1 - \tilde{Q}$, where $\tilde{Q}$ is defined as in (4.1), as a mask function to apply to the noisy spectrogram to retrieve the interference part from the mixture.

In all cases, the modified power spectrogram is coupled with the phase of the noisy spectrogram to obtain an estimation of the denoised complex spectrogram. An inverse transform is then used to synthesize the denoised signal back.

4.3.3 STFT and wavelet spectrograms

The HTC models are optimized to fit the wavelet power spectrum of an utterance. The basic idea to synthesize an enhanced or denoised speech, is, as described above, to modify the wavelet power spectrum and use an inverse wavelet transform. So far, a Gabor transform has been used for HTC analysis, and a first option is to perform the analysis-synthesis with this transform. As argued in [106], the HTC framework is naturally designed to fit a power spectrogram obtained with a constant-Q filterbank based on an analyzing wavelet whose Fourier transform is of the form

$$\Psi(\omega) = \Psi^*(\omega) = \begin{cases} \exp\left(- \frac{\left(\log \omega\right)^2}{4\sigma^2} \right) & (\omega > 0) \\ 0 & (\omega \leqq 0) \end{cases}. \tag{4.2}$$

One could thus also try to perform the analysis-synthesis with this analyzing wavelet.

But it is also possible to generate masking functions in the linear-frequency domain, and simply use STFT and inverse STFT, for example by overlap-add method, to generate a modified speech or background. Although this process is expected to lead lower-quality results than wavelet-based synthesis due to the misfit between the analysis and synthesis methods, it is much faster, and still gives interesting preliminary results on the basic performance of the method.

4.3.4 Relation to adaptive comb filtering methods

Comb filtering first requires a robust F_0 estimation, and according to [99], suffers from the fact that it retains too much interference as it passes through all frequency components close

to the multiples of target F_0. The HTC framework, on the opposite, features an embedded estimation of the F_0, and estimates separately the powers and shapes of the harmonics. HTC also estimates the parameters simultaneously in the time and frequency directions, thus integrating continuity in the model and making it more robust, on the contrary to comb filtering methods. In the multiple-speaker case especially, if the speeches of two speakers are harmonically related, a comb filter method will inevitably fail, while HTC can still perform well, as shown in Section 4.4.3.

4.4 Experimental evaluation for speech enhancement and speaker separation

We performed three types of experiments to confirm the basic effectiveness of our method for speech enhancement, background retrieval and speaker separation.We used the SNR as a quantitative measure of the performance of our algorithm, and tested different settings for the mask functions. We shall stress the fact that the SNR, although it is an easy-to-compute objective value, may not give a full idea of the performance of a CASA system, and may especially differ significantly from a perceptive evaluation. The human ear tends to prefer a stronger masking, even if it introduces artifacts or slightly modifies speech, which is not advantageous in an SNR way.

4.4.1 Speech enhancement

We used the same corpus of voiced speech and interference [48] as in Section 4.2.3. There are 10 interferences: n0, 1 kHz pure tone; n1, white noise; n2, noise bursts; n3, "cocktail party" noise; n4, rock music; n5, siren; n6, trill telephone; n7, female speech; n8, male speech; and n9, female speech. The results are shown in Table 4.6, for different mask settings on various types of interferences. Each value in the table represents the average SNR for one interference mixed with 10 target utterances. Mixture designates the Signal-to-Interference Ratio in the original mixture, and Hu-Wang stands for the state-of-the-art algorithm presented in [99]. The HTC enhanced speech is generated using Eq. (4.1) with p and ϵ as indicated in the first column. The parameters of the HTC model were as in 4.2.2, with a speech model and a noise model, apart from the number of harmonics considered, which was set to 40. The masks were generated in the STFT domain. We note that according to the type of interference, different settings lead to better results. This is due to the fact that these settings change the sharpness of the peaks of the mask, introducing a trade-off

Table 4.6: *SNR results (dB) for the enhanced speech.*

	n0	n1	n2	n3	n4	n5	n6	n7	n8	n9
Mixture	-3.27	-4.08	10.20	4.39	4.05	-5.83	1.90	6.57	10.53	0.75
Hu-Wang	16.34	7.83	16.71	8.32	10.88	14.41	16.89	11.97	14.44	5.27
$p = 1, \epsilon = 0.005$	-3.98	-0.54	15.11	6.63	4.90	-5.76	1.91	7.01	10.65	0.69
$p = 1, \epsilon = 0.1$	-6.04	5.61	11.39	7.79	6.69	-5.31	4.20	7.66	9.86	-0.47
$p = 2, \epsilon = 0.1$	-6.51	7.61	9.56	7.24	6.53	-5.57	4.72	7.06	8.68	-0.98

between Signal-to-Interference Ratio (SIR) and Signal-to-Artifact Ratio (SAR). According to the acoustical properties of the interference to reduce, it is thus possible to use different values for the parameters.

For n1, n2, n3 and n4, the results are promising, with results close to the Hu-Wang algorithm for the first three. It is so far less effective on the other interferences. For n0 and n5, which are interferences with a strong localized signal overlapping with harmonics of the speech utterance, our method failed as the speech model mistook the interference for a harmonic and rose the power of the corresponding Gaussian functions. This should be dealt with in the future, for example by using a stronger constraint on the power of the harmonics and on the shape of the power envelope in the time direction. The analysis of SIR and SAR results tends to show that our algorithm reduces the interference very effectively, but creates artifacts. The question whether these artifacts are perceptually significant or not should be further investigated, as is the improvement of the quality of the synthesized speech, especially using inverse wavelet transforms.

4.4.2 Speech cancellation for background enhancement

The HTC framework can be used not only for speech enhancement, but also for speech cancelling and background retrieval, as explained in section 4.3.2. There are very interesting potential applications to this task, such as the retrieval of speech in the background, or background music retrieval. This last issue is of particular importance: in an acoustical scene where someone is speaking with music playing in the background, being able to "clean" the background music from the speech would ease the automatic recognition of copyrighted material inside television programs for example. Another interesting application is the automatic cancellation of the vocal part inside a music piece to produce karaoke accompaniments at lower cost and from the real song, thus with a better quality than the usual MIDI accompaniments.

Table 4.7: *SNR results (dB) for the retrieved background.*

	n3	n4	n7	n8	n9
Mixture	-4.39	-4.05	-6.57	-10.53	-0.75
$p = 3, \epsilon^3 = 5.10^{-4}$	0.90	-0.01	-2.99	-4.85	-6.44
$p = 3, \epsilon^3 = 2.10^{-3}$	0.39	0.28	-2.62	-5.12	-5.36

The results for the interferences which retrieval is of potential interest for applications are presented in Table 4.7. Mixture designates the Interference-to-Signal Ratio in the original mixture. To our knowledge, no results by previous methods are available on this task. Background is retrieved through Eq. (4.1) with p and ϵ as indicated in the first column. Again, we obtained encouraging results. For an interference constituting of other speech with close average power such as n9, we will see in the next section that using two speech models leads to better results.

4.4.3 Speaker separation

By using two speech models, we showed in Section 4.2.4 that the F_0 contours of concurrent speech by two speakers with close average power can be effectively estimated through HTC. When several speech models q_k are used simultaneously, the ratio of one speech model inside the whole at each time-frequency bin, $q_i(x, t; \mathbf{\Theta}_{opt}) / \sum_k q_k(x, t; \mathbf{\Theta}_{opt})$, can be used as a mask to reconstruct the speech of each speaker. We performed a separation experiment on ten mixtures from Cooke's database (details on the database and HTC experimental setup are given in 4.2.2 and 4.2.3), where utterances by male speakers v0 to v9 are mixed with an utterance by a female speaker n9. The SNR is close to 0 dB for each utterance, and we note that this is a very difficult task as for some of the mixtures the harmonics of the speakers almost constantly overlap. The clean spectrograms of the utterances v0 (male speaker) and n9 (female speaker) can be seen in Fig. 4.1 and Fig. 4.2 respectively, and their mixture in Fig. 4.3. The corresponding spectrograms extracted using HTC from the mixture v0n9 can be seen in Fig. 4.4 and Fig. 4.5 respectively. A constant-Q filterbank transform and inverse transform was used.

The results for the ten mixtures are shown in Table 4.8. For comparison, SNR results obtained using Hu and Wang's state-of-the-art algorithm [99] are also given. This algorithm only focuses on the target source, and results for the second speaker are thus not available. One can see that our method outperforms this algorithm on this task. We also note that, contrary to the algorithm of Hu and Wang, it does not seem to generate musical noise.

Table 4.8: *SNR results (dB) for the speaker separation.*

	v0n9	v1n9	v2n9	v3n9	v4n9	v5n9	v6n9	v7n9	v8n9	v9n9	Average
Mixture SNR	0.30	-1.98	-0.96	2.92	0.63	0.80	-0.05	2.12	1.38	2.72	0.79
Hu-Wang	3.89	2.84	7.59	8.12	3.51	3.24	4.00	6.97	4.84	8.69	5.37
HTC, Speaker 1	7.26	7.40	8.67	10.61	6.64	9.80	7.01	9.58	7.14	12.32	8.64
Mixture NSR	-0.30	1.98	0.96	-2.92	-0.63	-0.80	0.05	-2.12	-1.38	-2.72	-0.79
HTC, Speaker 2	6.33	9.00	9.03	6.82	5.13	8.83	6.20	7.04	5.03	8.49	7.19

4.5 Summary of Chapter 4

In this chapter, we evaluated through experiments the performance of the HTC model introduced in Chapter 3 for various tasks of single channel acoustical scene analysis. We first investigated its accuracy as a pitch determination algorithm in various environments as well as on concurrent speech. On single-speaker clean speech, we obtained good results which we compared with existing methods specifically designed for that task. On co-channel concurrent speech, single-speaker speech mixed with white noise, pink noise, and on a corpus of single-speaker speech mixed with a variety of interfering sounds, we showed that our algorithm outperforms existing methods. We then presented several applications of the HTC framework for single channel speech signal processing problems. We showed its basic effectiveness for speech enhancement and speech cancellation for background retrieval, with potential applications in background music retrieval. We then evaluated its performance on speaker separation, and showed that it outperforms existing methods on this task.

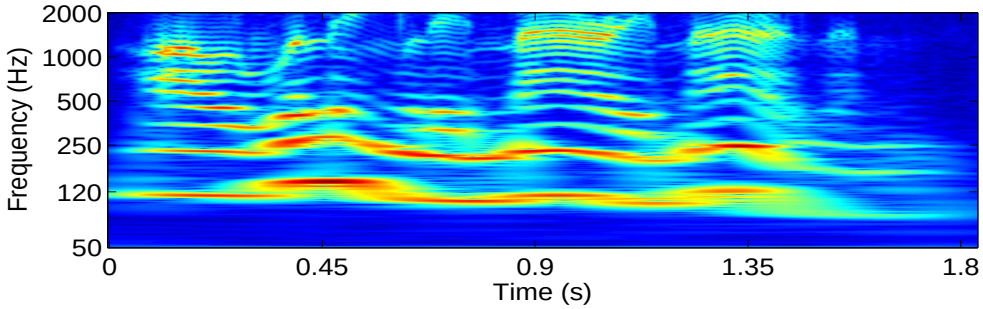


Figure 4.1: *Clean spectrogram of speaker v0*

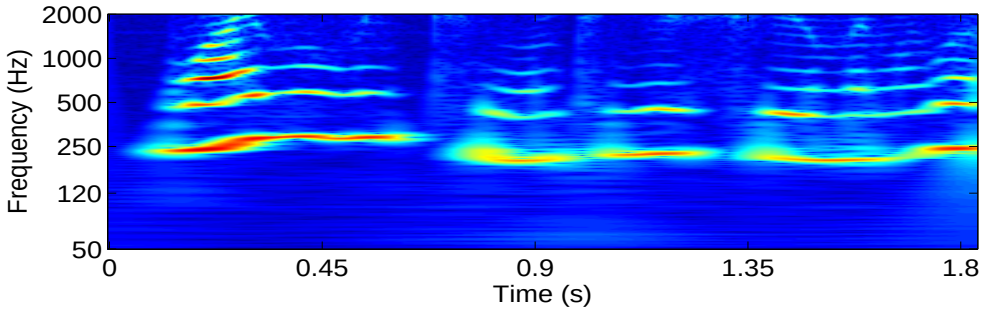


Figure 4.2: *Clean spectrogram of speaker n9*

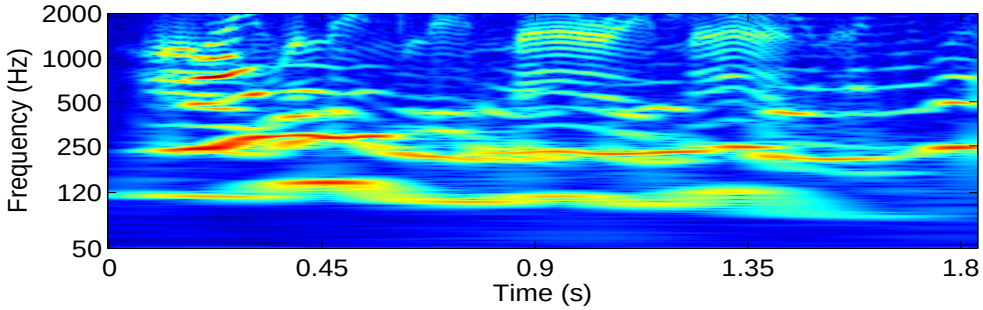


Figure 4.3: *Spectrogram of the mixture v0n9*

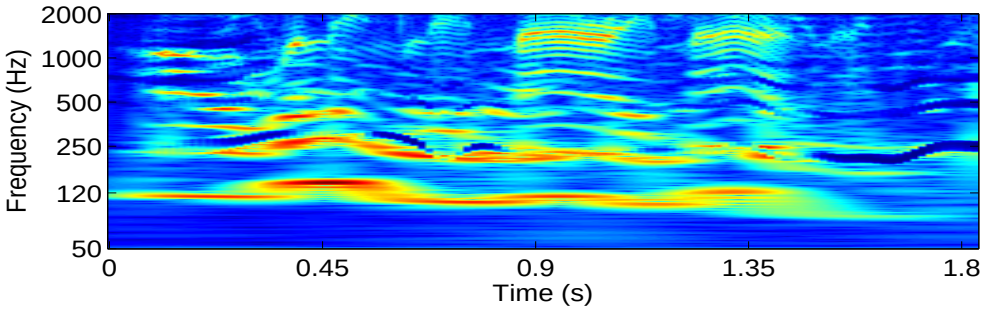


Figure 4.4: *Estimated spectrogram of speaker v0*

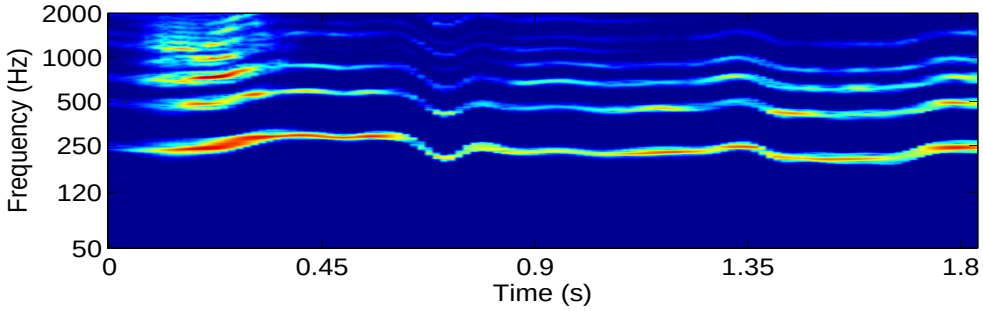


Figure 4.5: *Estimated spectrogram of speaker n9*

Chapter 5

Analysis of acoustical scenes with incomplete data

5.1 Introduction

In Chapter 3 and Chapter 4, we introduced a statistical model for speech signals and showed how it could be used to perform the analysis of acoustical scenes involving speech. We investigate in this chapter how to use such statistical models to analyze acoustical scenes in the presence of incomplete data, and to subsequently recover the missing data.

So far, a particular attention has been given in the audio signal processing, and more specifically CASA, community to solving the so-called "cocktail party problem", the computational implementation of the "cocktail party effect", i.e., the ability of the human auditory system to focus on a single talker among a mixture of conversations and background noises, leading to the development of many methods for multi-pitch estimation, noise canceling or source separation [206]. Less emphasis has been put on the computational realization of another remarkable ability of the human auditory system, auditory induction. Humans are able, under certain conditions, to estimate the missing parts of a continuous acoustic stream briefly covered by noise, to perceptually resynthesize and clearly hear them [111]. Humans are thus able to simultaneously analyze an auditory scene, as in the cocktail party effect, in the presence of gaps, and to reconstruct the signal inside those gaps.

The development of an effective computational counterpart to this ability would lead to many important engineering applications, from polyphonic music recording analysis and restoration to mobile communications robust to both packet-loss and background noise. Attempts to combine scene analysis on incomplete data and reconstruction of the missing

parts are rare, with the notable exception of Ellis [72,73], unfortunately with limited success, and even fewer have been the attempts to do so through a statistical approach.

This chapter aims at developing such a computational counterpart to auditory induction, by simultaneously performing a decomposition of the magnitude wavelet spectrogram of a sound with missing or corrupted samples, and filling in the gaps into that spectrogram. Various approaches have emerged recently which attempt to analyze the structure of the spectrogram of an acoustical scene, such as the HTC framework we introduced in the preceding chapters, while on the other side gap interpolation techniques have been the subject of research for many years [8, 38, 47, 76, 86, 126, 211]. However, only few models so far try to deal with both issues. The framework developed by Reyes-Gomez et al. [166] is an example of such models. While it relies on local regularities of the spectrogram, the framework we introduce can use both local and global regularities.

We show here how statistical models which globally model acoustical scenes can be extended to be used for the analysis of acoustical scenes with incomplete data. After deriving the method for a general class of distortion functions to measure the goodness of fit between the model and the observed data, we show how, for a particular class of functions called Bregman divergences, the introduced method can be interpreted in terms of the EM algorithm, enabling the use of prior distributions on the parameters which can enforce local smoothness regularities. To illustrate simply the concept of this method, we use the non-negative matrix factor 2-D deconvolution algorithm (NMF2D) [176] as a representative method to extract global structures in audio spectrograms and perform a short experimental evaluation to validate our method on a piece of computer generated music. We finally show how to apply this method to the model presented in the preceding chapters, HTC, and how the obtained algorithm can be used to perform F_0 estimation of speech on incomplete data.

5.2 Audio inpainting

5.2.1 Problem setting

We consider the problem of interpolating gaps in audio signals by filling in the gaps in the magnitude spectrogram. We will not consider here the reconstruction of the phase. If the magnitude spectrogram can be accurately reconstructed, other methods could be used to obtain a phase consistent with it, as we shall explain in Chapter 6. We are interested in using local and global regularities in the spectrogram to simultaneously analyze the acoustical scene and fill in gaps that may have occurred into it. We believe that this is close to what

is performed by humans in the auditory induction mechanism, when for example sounds actually missing from a speech signal can in certain conditions be perceptually synthesized by the brain and clearly heard [111]. Our goal is thus to "inpaint" the missing regions of the spectrogram based on global and local regularities, in the same spirit as what is done for image inpainting [26], where diffusion-based (local) and exemplar-based (global) techniques are used [51].

Assuming we have at hand a statistical framework which globally models an acoustical scene, we explain in this chapter how to use it on acoustical scenes with incomplete data. Moreover, we show in Section 5.3.1 how to enforce local smoothness regularities by adding prior distributions on the parameters. We present two examples of statistical frameworks which can be used in this context, Schmidt and Mørup's NMF2D algorithm [141, 176], and the Harmonic-Temporal Clustering (HTC) framework introduced in Chapter 3.

5.2.2 General method and applicability

The general idea is simple: if a framework has been developed to analyze a complete acoustical scene by means of fitting a statistical model to the observed data, we show here that it can be used as well on incomplete data by iterating between analysis steps and reconstruction steps; moreover, if knowing the optimal parameters of the model gives a good estimation of the data where they are missing, the algorithm we present can be used to reconstruct the missing data as well. The procedure goes as follows: during a reconstruction step, the missing data is estimated based on the current value of the model; during an analysis step, the original framework is applied normally on the data completed during the reconstruction step. We show in the following subsection how this iterative algorithm can be interpreted as using an auxiliary function method to optimize the fitting of the statistical model on the regions where data were actually observed.

Situations in which such an incomplete data framework needs to be used are quite varied. One can cite for example situations where a portion of the power spectrum

1. is lost, for example after a packet loss during a network communication,
2. is corrupted, for example by an interference in the background, a degradation of the recording material such as clicks,
3. has been discarded, for example by a binary mask designed to suppress noise or select a particular speaker inside an acoustical scene,
4. or is simply not observed, for example because it lies above the Nyquist frequency or is outside the time interval of analysis.

The important issue here is to make sure that the statistical model used has sufficient "prediction power" in the missing parts. The method we introduce can be used in general to perform the fitting on observed regions even when the original optimization algorithm was only designed to be effective on complete data, such as Gaussian fitting for example. Its capacity of reconstructing the missing-data regions, however, will depend on the design of the model, and especially on the constraints introduced: the only guaranty is to obtain, on the whole domain, a complete model which fits the data were they were observed. Continuity constraints would then for example ensure a smooth transition over missing-data regions. Models such as NMF2D use information from the whole domain to build a lower representation of the acoustical scene, and are thus natural candidate to lead to good reconstruction through the method we propose. Similarly, models such as HTC, through the use of relevant prior distributions, shall also lead to coherent reconstructions, which are inherently guaranteed to respect Bregman's grouping cues due to the original design of the model.

5.2.3 Auxiliary function method

Derivation of the algorithm

Suppose one wants to fit a parametric distribution to an observed contour which is incomplete, in the sense that its values are only known on a subset $I \subset D \subseteq \mathbb{R}^n$, where D is the domain of definition of the problem of interest. Suppose also that if the data were complete, the fitting could be performed (Gaussian distribution fitting, etc.). Then we show that using an iterative procedure based on the auxiliary function method, the fitting to the incomplete data can also be performed.

Let f be the observed contour, and $g(\cdot;\mathbf{\Theta})$ a model parameterized by $\mathbf{\Theta}$ such that the fitting of this model to an observed contour defined on the whole domain D can be performed.

We consider a distortion function $d : \mathcal{S} \times \mathcal{S} \to [0, +\infty)$ where $\mathcal{S} \subseteq \mathbb{R}^n$, such that $d(x, y) \geq 0, \forall x, y \in \mathcal{S}$ and equality holds if and only if $x = y$. As this function d is not required to respect the triangle inequality, it is not necessarily a distance, *sensu stricto*. For such a distortion function, we can introduce a measure of the "distance" between the observed data and the model by integrating d between f and $g(\cdot;\mathbf{\Theta})$ on the subset I:

$$\mathcal{L}(\mathbf{\Theta}) = \int_I d(f(x), g(x;\mathbf{\Theta}))\, dx. \tag{5.1}$$

In this kind of situation, it is often preferable, instead of defining an "incomplete model" whose estimation may be cumbersome, to try to fall back on a complete data estimation

problem. This is what we do here by introducing an auxiliary function. For any function h taking values in $\mathcal{S}$ and defined on $D \setminus I$, let us define

$$\mathcal{L}^+(\mathbf{\Theta}, h) = \mathcal{L}(\mathbf{\Theta}) + \int_{D\setminus I} d(h(x), g(x; \mathbf{\Theta}))\, dx. \tag{5.2}$$

As the second term on the right-hand side is itself derived from the distortion measure, it is non-negative, and thus

$$\mathcal{L}(\mathbf{\Theta}) \leq \mathcal{L}^+(\mathbf{\Theta}, h),\ \forall h. \tag{5.3}$$

Moreover, there is equality in the inequality for $h = g(\cdot; \mathbf{\Theta})$.

The minimization procedure can now be described as follows. After initializing $\mathbf{\Theta}$ for example by performing the distribution fitting on the observed data completed by 0 on $D \setminus I$, one then iteratively performs the following updates:

Step 1 Estimate h such that $\mathcal{L}(\mathbf{\Theta}) = \mathcal{L}^+(\mathbf{\Theta}, h)$:

$$\hat{h} = g(\cdot; \mathbf{\Theta}). \tag{5.4}$$

Step 2 Update $\mathbf{\Theta}$ with $\hat{h}$ fixed:

$$\hat{\mathbf{\Theta}} = \operatorname*{argmin}_{\mathbf{\Theta}} \mathcal{L}^+(\mathbf{\Theta}, \hat{h}). \tag{5.5}$$

The optimization process is illustrated in Fig. 5.1.

Example: Gaussian mixture fitting

In the particular case of the modeling of a non-negative distribution through Gaussian mixtures on $\mathbb{R}^n$, one can use as a measure of fitting the $\mathcal{I}$-divergence.

Let f be the observed contour, and $g(\cdot; \mathbf{\Theta})$ a Gaussian mixture distribution parameterized by $\mathbf{\Theta}$. We consider the $\mathcal{I}$-divergence between f and $g(\cdot; \mathbf{\Theta})$ computed on the interval I

$$\mathcal{I}(\mathbf{\Theta}) = \int_I \left(f(x) \log \frac{f(x)}{g(x; \mathbf{\Theta})} - (f(x) - g(x; \mathbf{\Theta})) \right) dx \tag{5.6}$$

as a measure of the distance between the observed data and the distribution. In the case of Gaussian mixture fitting, the fact that the integral $\int_I g(x; \mathbf{\Theta})\, dx$ can not be computed analytically in general makes the direct minimization of $\mathcal{I}$ with respect to $\mathbf{\Theta}$ arduous.

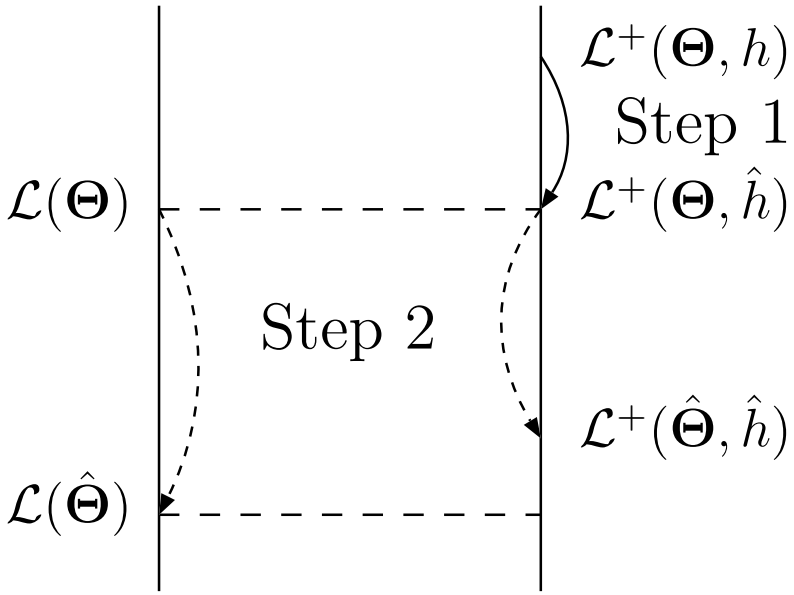


Figure 5.1: *Optimization through the iterative procedure. During the step 1, the auxiliary parameter h is updated to $\hat{h}$ so that $\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}^+(\boldsymbol{\Theta}, \hat{h})$. Then, during the step 2, $\mathcal{L}^+(\boldsymbol{\Theta}, \hat{h})$ is optimized w.r.t. $\boldsymbol{\Theta}$, ensuring that $\mathcal{L}(\hat{\boldsymbol{\Theta}}) \leq \mathcal{L}^+(\hat{\boldsymbol{\Theta}}, \hat{h}) < \mathcal{L}^+(\boldsymbol{\Theta}, \hat{h}) = \mathcal{L}(\boldsymbol{\Theta})$. The minimization of $\mathcal{L}(\boldsymbol{\Theta})$ can thus be performed through the minimization of the auxiliary function $\mathcal{L}^+(\boldsymbol{\Theta}, h)$ alternately w.r.t. h and $\boldsymbol{\Theta}$.*

Following Eq. (5.2), for any non-negative function h defined on $\mathbb{R}^n \setminus I$, let us define

$$\mathcal{I}^+(\boldsymbol{\Theta}, h) = \mathcal{I}(\boldsymbol{\Theta}) + \int_{\mathbb{R}^n \setminus I} \left(h(x) \log \frac{h(x)}{g(x; \boldsymbol{\Theta})} - (h(x) - g(x; \boldsymbol{\Theta})) \right) dx. \tag{5.7}$$

Again, as the second term on the right-hand side is itself an $\mathcal{I}$-divergence, it is non-negative, and thus

$$\mathcal{I}(\boldsymbol{\Theta}) \leq \mathcal{I}^+(\boldsymbol{\Theta}, h), \, \forall h. \tag{5.8}$$

Moreover, there is equality in the inequality for $h = g(\cdot; \boldsymbol{\Theta})$.

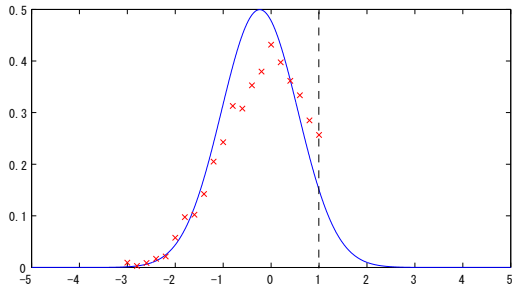
After initializing $\boldsymbol{\Theta}$ for example by performing the distribution fitting on the observed data completed by 0 on $\mathbb{R}^n \setminus I$, one then iteratively performs the following updates:

Step 1 Estimate h such that $\mathcal{I}(\boldsymbol{\Theta}) = \mathcal{I}^+(\boldsymbol{\Theta}, h)$:

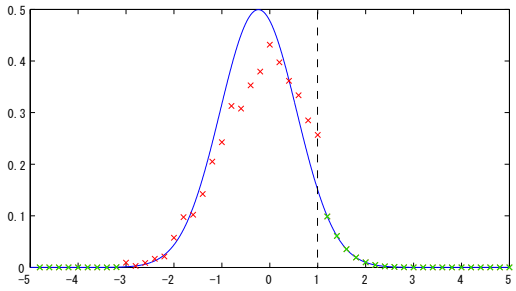
$$\hat{h} = g(\cdot; \boldsymbol{\Theta}). \tag{5.9}$$

Step 2 Update $\boldsymbol{\Theta}$ with $\hat{h}$ fixed:

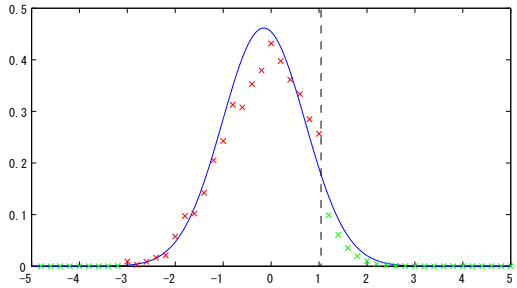
$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \, \mathcal{I}^+(\boldsymbol{\Theta}, \hat{h}). \tag{5.10}$$



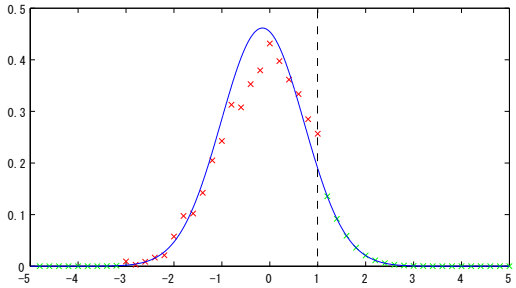
(a) Original incomplete data and initial fitting.



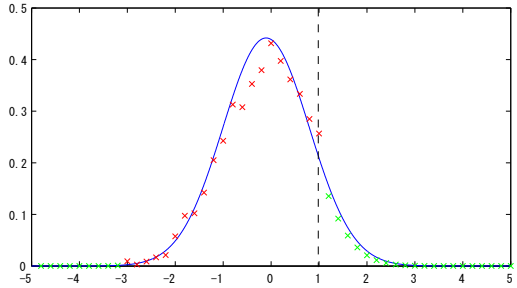
(b) Iteration 1, Step 1.



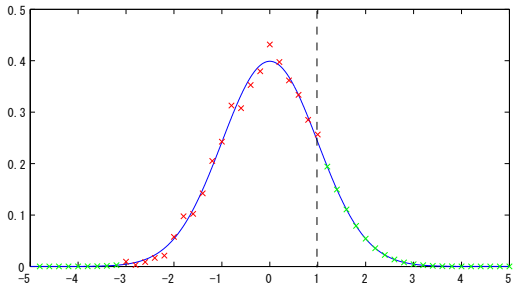
(c) Iteration 1, Step 2.



(d) Iteration 2, Step 1.



(e) Iteration 2, Step 2.



(f) Estimated model after convergence.

Figure 5.2: *Example of fitting of a Gaussian distribution on incomplete data. (a) Original incomplete data (centered normal distribution with variance 1, with multiplicative Gaussian noise, in red) and initial fitting of a Gaussian distribution on the data (in blue), implicitly assuming that the data is 0 outside its interval of definition. (b) After Iteration 1, Step 1: the current model is used as an estimation of the missing data (in green). (c) After Iteration 1, Step 2: Update of the model. (d) After Iteration 2, Step 1: update of the missing data. (e) After Iteration 2, Step 2: update of the model. (f) Estimated model after convergence.*

When applied to the fitting of a Gaussian mixture model to an observed distribution for which the values at certain points have been lost, Step 1 consists in using the value of the current Gaussian mixture model at those points to complete the data, and Step 2 can be performed using usual methods for Gaussian mixture fitting. An example is shown in Fig. 5.2.

5.3 Probabilistic interpretation for Bregman divergences

5.3.1 Relation between Bregman divergence-based optimization and Maximum Likelihood estimation

We follow [16] and [90] to give a brief overview of the concepts of exponential family and Bregman divergence and to present the relation between them. As a complete presentation would take us too far from the purpose of the present discussion, we shall refer to them for more details and for rigorous derivations. We tried however to keep this chapter as self-contained as possible.

Exponential families form a group of probability distributions which comprise many common families of probability distributions such as the normal, gamma, Dirichlet, binomial and Poisson distributions, among others. They are defined as follows.

Definition 5.1. *Let Λ be an open convex subset of $\mathbb{R}^d$ and let $\mathcal{M} = \{P_\beta \mid \beta \in \Lambda\}$ be a family of probability distributions on a sample space $\mathcal{X}$. $\mathcal{M}$ is an exponential family if there exist a function $\zeta = (\zeta_1, \ldots, \zeta_d) : \mathcal{X} \to \mathbb{R}^d$ and a non-negative function $r : \mathcal{X} \to [0, +\infty)$ such that, for all $\beta \in \Lambda$,*

$$P_{\psi,\beta}(X) \triangleq e^{\langle \beta; \zeta(X) \rangle - \psi(\beta)} r(X), \tag{5.11}$$

where $\langle \beta; \zeta(X) \rangle$ is the inner product between β and $\zeta(X)$, and

$$\psi(\beta) = \log \int_{\mathcal{X}} \exp(\langle \beta; \zeta(x) \rangle) r(x)\, dx < +\infty.$$

An exponential family defined in terms of a function $\zeta = (\zeta_1, \ldots, \zeta_d)$ is called a regular exponential family if the representation (5.11) is minimal, i.e., there exists no $\alpha_0, \alpha_1, \ldots, \alpha_d \in \mathbb{R}^{d+1} \setminus \{0\}$ such that for all x with $r(x) > 0$, $\sum_{j=1}^{d} \alpha_j \zeta_j(x) = \alpha_0$.

As an example, we consider the family of Poisson distributions $\{P_\theta \mid \theta \in (0, +\infty)\}$ on the sample space $\mathcal{X} = \mathbb{N}$ defined as $P_\theta(x) = \frac{1}{x!} e^{-\theta} \theta^x$. We see that it is an exponential family, with $\beta = \log \theta$, $\zeta(X) = X$, $\psi(\beta) = e^\beta$, and $r(x) = 1/x!$. The function ζ is not always the identity function as in the Poisson case, as can be seen with the family of normal distributions $\{f_{\mu,\sigma^2} \mid (\mu, \sigma^2) \in \mathbb{R} \times [0, +\infty)\}$ with $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, which can be seen to be an exponential family by setting $\beta = (\mu/\sigma^2, -1/(2\sigma^2))$, $\zeta(X) = (X, X^2)$ and $r(x) = 1/\sqrt{2\pi}$.

It can be shown [16, 90] that exponential families can actually be parameterized by the mean value $\mu(\beta) = \mathbb{E}(\zeta(X))$ of $\zeta(X)$. If we let $\Lambda_{\text{mean}} \triangleq \{\mu \mid \exists \beta \in \Lambda \text{ such that } \mu(\beta) = \mu\}$, then $\mu(\cdot)$ is a 1-to-1 mapping from Λ to Λ_{mean}. Moreover, there exists a function $\phi : \Lambda_{\text{mean}} \to$

$\mathbb{R}$ such that for all $\beta \in \Lambda$ and for all $\mu \in \Lambda_{\text{mean}}$ such that $\mu = \mu(\beta)$,

$$\phi(\mu) + \psi(\beta) = \langle \beta; \mu \rangle, \tag{5.12}$$

from which one can deduce in particular that $\beta(\mu) = \nabla\phi(\mu)$. Altogether, by noticing that

$$\begin{aligned} \langle \beta; \zeta(X) \rangle - \psi(\beta) &= \langle \beta; \mu \rangle - \psi(\beta) + \langle \beta; \zeta(X) - \mu \rangle \\ &= \phi(\mu) + \langle \nabla\phi(\mu); \zeta(X) - \mu \rangle, \end{aligned} \tag{5.13}$$

$P_{\psi,\beta}$ can be rewritten parameterized by $\mu = \mu(\beta)$, leading to the so-called *mean-value parameterization* of the exponential family:

$$P_{\phi,\mu}(X) \triangleq P_{\psi,\beta(\mu)}(X) = e^{\phi(\mu) - \langle \nabla\phi(\mu); \zeta(X) - \mu \rangle} r(X). \tag{5.14}$$

We will call μ the expectation parameter of the exponential family, which will be denoted by $\mathcal{F}_\phi$.

We are now ready to introduce the concept of Bregman divergence and to derive its relation with the exponential families.

Definition 5.2. *Let $\phi : \mathcal{S} \to \mathbb{R}$ be a strictly convex function defined on an open convex set $\mathcal{S} \subseteq \mathbb{R}^d$ such that ϕ is differentiable on $\mathcal{S}$. The Bregman divergence $d_\phi : \mathcal{S} \times \mathcal{S} \to [0, +\infty)$ is defined as*

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y; \nabla\phi(y) \rangle,$$

where $\nabla\phi(y)$ is the gradient vector of ϕ evaluated at y.

Bregman divergences include a large number of useful loss functions such as squared loss, KL-divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, and the $\mathcal{I}$-divergence. They verify a non-negativity property: $d_\phi(x, y) \geq 0, \forall x, y \in \mathcal{S}$, and equality holds if and only if $x = y$.

Banerjee et al. [16] showed that the following informal derivation can be rigorously justified for a wide subclass of Bregman divergences, which includes in particular all the loss functions cited above.

If $P_{\phi,\mu}$ is the probability density function of the regular exponential family $\mathcal{F}_\phi$ (in its mean-value parameterization) associated to the function ϕ defining the Bregman divergence d_ϕ, from (5.14) we have,

$$P_{\phi,\mu}(x) = e^{\phi(\mu) + \langle \nabla\phi(\mu); \zeta(x) - \mu \rangle} r(x)$$

$$= \quad e^{-d_\phi(\zeta(x),\mu)+\phi(\zeta(x))} r(x)$$

and eventually

$$P_{\phi,\mu}(x) = e^{-d_\phi(\zeta(x),\mu)} b_\phi(x) \tag{5.15}$$

where $b_\phi(x) = e^{\phi(\zeta(x))} r(x)$. This relation holds for all $x \in \mathrm{dom}(\phi)$, which can be shown [16] to include the set of all the instances that can be drawn from the distribution $P_{\phi,\mu}$. However, one must be careful when using this relation in certain cases where the inclusion is strict, in particular when the support of the carrier $r(x)$ is strictly smaller than $\mathrm{dom}(\phi)$. Indeed, for all x outside that support, Eq. (5.15) is verified as both members are equal to zero, but it is not informative on the relation between $P_{\phi,\mu}(x)$ and $d_\phi(\zeta(x), \mu)$ as the right-hand side member is zero only because $b_\phi(x)$ is. This is what happens for example for the $\mathcal{I}$-divergence (with $\phi(\mu) = \mu \log \mu - \mu$) for which $\mathrm{dom}(\phi) = \mathbb{R}^+$ (extending the definition of ϕ for $\mu = 0$). The corresponding exponential family is the Poisson family, for which the set of instances and the support of the carrier are only $\mathbb{N}$.

The relation (5.15) builds a bridge between optimization based on Bregman divergences and Maximum-Likelihood (ML) estimation with exponential families. As distribution-fitting problems usually involve only a first-order statistic, we will focus on the case $\zeta(X) = X$. Trying to fit a model $g(\cdot; \boldsymbol{\Theta})$, defined on a domain D with parameter $\boldsymbol{\Theta}$, to an observed distribution f with a measure of distance between the two based on a Bregman divergence d_ϕ then amounts to looking for $\boldsymbol{\Theta}$ minimizing $\int_D d_\phi(f(x), g(x; \boldsymbol{\Theta}))\, dx$. But according to (5.15), this is equivalent (up to some precautions which may have to be taken because of the misfit between the domains of definition of the Bregman divergence and the exponential family evoked above) to maximizing w.r.t. $\boldsymbol{\Theta}$ the log-likelihood $\int_D P_{\phi,g(x;\boldsymbol{\Theta})}(f(x))\, dx$ where the observed data points $f(x)$ at point x are assumed to have been independently generated from $P_{\phi,g(x;\boldsymbol{\Theta})}$.

5.3.2 Relation to the EM algorithm

We consider the framework of Section 5.2.3, with as distortion function a Bregman divergence d_ϕ such that the associated exponential family $\mathcal{F}_\phi$ verifies $\zeta(X) = X$. In the following, we will denote by $\nu_{x,\phi,\boldsymbol{\Theta}}(z)$ the density of a probability distribution from the exponential family $\mathcal{F}_\phi$ with expectation parameter $g(x; \boldsymbol{\Theta})$, which can be written as explained in 5.3.1 directly using the corresponding Bregman divergence:

$$\nu_{x,\phi,\boldsymbol{\Theta}}(z) = e^{-d_\phi(z,g(x;\boldsymbol{\Theta}))} b_\phi(z). \tag{5.16}$$

Classically, to derive the Q-function used in the EM algorithm, one considers the expectation of the log-likelihood $\log P(f|\mathbf{\Theta})$ of the observed data f against the conditional probability of the unobserved data h with respect to the observed data and the model with parameter $\bar{\mathbf{\Theta}}$:

$$\begin{aligned}\log P(f|\mathbf{\Theta}) &= \mathbb{E}(\log P(f|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\ &= \mathbb{E}(\log P(f,h|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} - \mathbb{E}(\log P(h|f,\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\ &= Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) - H(\mathbf{\Theta},\bar{\mathbf{\Theta}}), \end{aligned} \tag{5.17}$$

where h denotes the unobserved data. One notices through Jensen inequality that

$$\forall \mathbf{\Theta}, H(\mathbf{\Theta},\bar{\mathbf{\Theta}}) \leq H(\bar{\mathbf{\Theta}},\bar{\mathbf{\Theta}}),$$

such that if one can update $\mathbf{\Theta}$ such that $Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) > Q(\bar{\mathbf{\Theta}},\bar{\mathbf{\Theta}})$, then $\log P(f|\mathbf{\Theta}) > \log P(f|\bar{\mathbf{\Theta}})$.

Let us now compute the explicit form of $Q(\mathbf{\Theta},\bar{\mathbf{\Theta}})$. First, let us remind here that we showed in 5.3.1 that the optimization based on the Bregman divergence corresponds to an ML problem in which the data are supposed to have been generated independently from probability distributions of an exponential family $\mathcal{F}_\phi$ with expectation parameters $g(x,\mathbf{\Theta})$. Thus, observed and unobserved data are independent conditionally to $\mathbf{\Theta}$, and the Q-function can be written as follows:

$$\begin{aligned} Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) &= \mathbb{E}(\log P(h|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} + \mathbb{E}(\log P(f|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\ &= \int_{\mathbb{R}^n\setminus I} \mathbb{E}(\log P(h(x)|\mathbf{\Theta}))_{P(h(x)|\bar{\mathbf{\Theta}})}\, dx \\ &\qquad + \left(\int P(h|f,\bar{\mathbf{\Theta}})\right) \int_I \log P(f(x)|\mathbf{\Theta})\, dx \\ &= \int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) \log \nu_{x,\phi,\mathbf{\Theta}}(z)\, dz\, dx \\ &\qquad + \int_I \log P(f(x)|\mathbf{\Theta})\, dx \\ &= \int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) \Big(\log b_\phi(z) - d_\phi(z, g(x;\mathbf{\Theta}))\Big)\, dz\, dx \\ &\qquad + \int_I \Big(\log b_\phi(f(x)) - d_\phi(f(x), g(x;\mathbf{\Theta}))\Big)\, dx \\ &= -\int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) d_\phi(z, g(x;\mathbf{\Theta}))\, dz\, dx \\ &\qquad - \int_I d_\phi(f(x), g(x;\mathbf{\Theta}))\, dx + C_1(f,\bar{\mathbf{\Theta}}), \end{aligned} \tag{5.18}$$

where $C_1(f, \bar{\boldsymbol{\Theta}})$ does not depend on $\boldsymbol{\Theta}$. If we now rewrite $d_\phi(z, g(x; \boldsymbol{\Theta}))$ as

$$\begin{aligned} d_\phi(z, g(x;\boldsymbol{\Theta})) &= d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) + \phi(z) - \phi(g(x;\bar{\boldsymbol{\Theta}})) \\ &\quad - \langle z - g(x;\bar{\boldsymbol{\Theta}}); \nabla\phi(g(x;\boldsymbol{\Theta}))\rangle, \end{aligned} \tag{5.19}$$

we can simplify the first term in Eq. (5.18):

$$\begin{aligned} \int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z) d_\phi(z, g(x;\boldsymbol{\Theta}))\, dz &= \Big(\int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz\Big) d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) \\ &\quad - \langle \int (z - g(x;\bar{\boldsymbol{\Theta}}))\nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz; \nabla\phi(g(x;\boldsymbol{\Theta}))\rangle + C_2(f, \bar{\boldsymbol{\Theta}}) \\ &= d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta})) + C_2(f, \bar{\boldsymbol{\Theta}}), \end{aligned} \tag{5.20}$$

where $C_2(f, \bar{\boldsymbol{\Theta}})$ does not depend on $\boldsymbol{\Theta}$. To lead the calculation above, we used the fact that the mass of a probability distribution of an exponential family with expectation parameter $g(x; \bar{\boldsymbol{\Theta}})$ is 1 and its mean is $g(x; \bar{\boldsymbol{\Theta}})$:

$$\int \nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz = 1, \tag{5.21}$$

$$\int z\nu_{x,\phi,\bar{\boldsymbol{\Theta}}}(z)\, dz = g(x;\bar{\boldsymbol{\Theta}}). \tag{5.22}$$

We then obtain for the Q-function

$$\begin{aligned} Q(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) &= -\int_{\mathbb{R}^n \setminus I} d_\phi(g(x;\bar{\boldsymbol{\Theta}}), g(x;\boldsymbol{\Theta}))\, dx \\ &\quad - \int_I d_\phi(f(x), g(x;\boldsymbol{\Theta}))\, dx + C(f, \bar{\boldsymbol{\Theta}}) \\ &= -\mathcal{L}^+(\boldsymbol{\Theta}, g(x;\bar{\boldsymbol{\Theta}})) + C(f, \bar{\boldsymbol{\Theta}}), \end{aligned} \tag{5.23}$$

where $C(f, \bar{\boldsymbol{\Theta}})$ again does not depend on $\boldsymbol{\Theta}$.

Altogether, we find that there is a correspondence between the Q-function and the auxiliary function $\mathcal{L}^+$ that we introduced in Section 5.2.3. Computing the Q-function, i.e., the E-step of the EM algorithm, corresponds to computing the auxiliary function, which is done by replacing the unknown data by the model at the current step. Maximizing the Q-function w.r.t. $\boldsymbol{\Theta}$, i.e., the M-step of the EM algorithm, corresponds to minimizing the auxiliary function w.r.t. $\boldsymbol{\Theta}$. This shows how to derive the auxiliary function in an EM point of view, and enables us for example to consider prior distributions on the parameters and perform a MAP estimation.

5.3.3 Remark on the limitations of this interpretation

We showed that the auxiliary function method in Section 5.2.3 could be derived through the EM algorithm in the special case of the function d being a Bregman divergence d_ϕ such that the associated exponential family verifies $\zeta(X) = X$. We shall note however that one has to pay attention to the support of the probability distributions of the exponential family. Indeed, as noted earlier, it may happen that these distributions have a smaller support than the original set on which the Bregman divergence is defined. This is for example the case for the $\mathcal{I}$-divergence, which is defined on $\mathbb{R}^+$ but is associated to the Poisson distribution, whose support is $\mathbb{N}$. The formulation presented in Section 5.2.3 is thus more general than its EM counterpart, although it does not justify the use of penalty functions as prior distributions on the parameters. In the particular case of the $\mathcal{I}$-divergence, it is actually possible to justify the use of the ML interpretation with real data. This issue is investigated in Appendix A.

5.4 Missing-data non-negative matrix factorization

5.4.1 Overview of the original algorithm

The NMF2D algorithm is an extension of Smaragdis's non-negative matrix factor deconvolution (NMFD) [184], itself an extension of the original non-negative matrix factorization (NMF) [118]. NMF is a general tool which attempts to decompose a non-negative matrix $V \in \mathbb{R}^{\geq 0, M \times N}$ in the product of two usually lower-rank non-negative matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$,

$$V \approx WH.$$

In applications to audio, the horizontal and vertical dimensions of the matrices respectively represent time and frequency (or log-frequency). NMFD extends NMF by introducing a convolution in the time direction, and looks for a decomposition of V as

$$V \approx \Lambda = \sum_\tau W^\tau \overrightarrow{H}^\tau \tag{5.24}$$

where $\rightarrow \tau$ denotes the right shift operator which moves each element in a matrix τ columns to the right, e.g.,

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \overrightarrow{A}^1 = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \\ 0 & 7 & 8 \end{pmatrix}.$$

NMFD thus enables the representation of time structure in the extracted patterns. NMF2D generalizes this approach to the frequency direction through a 2-D convolution. By using

a log-frequency spectrogram, a pitch change corresponds to a shift on the frequency axis. Assuming that the spectral patterns to be modeled are roughly pitch-invariant, NMF2D can thus account for both time and frequency structures. Concretely, the NMF2D model is

$$V \approx \Lambda = \sum_{\tau} \sum_{\phi} \overset{\downarrow \phi}{W^{\tau}} \overset{\rightarrow \tau}{H^{\phi}} \tag{5.25}$$

where $\downarrow \phi$ denotes the down shift operator which moves each element in a matrix ϕ lines down. Up shift and left shift operator can be introduced in the same way. As we mentioned before, applying NMFD or NMF2D to audio signals implies making a sparseness assumption on the signal, as the additivity of magnitudes in the spectral domain is only true if the underlying components of the signal are sparse enough to minimize overlaps.

Lee and Seung [118] introduced efficient algorithms for computing the NMF of a matrix V based on both the least-squares error and the $\mathcal{I}$-divergence, which have been extended by Smaragdis for NMFD [184] and Schmidt and Mørup for NMF2D [141, 176]. These algorithms are based on multiplicative updates. If Λ is defined as in (5.25), we define the objective function as $\mathcal{J}(W, H|V) = ||V - \Lambda||_F^2$ for the least-squares error, where $|| \cdot ||_F$ denotes the Frobenius norm (sum of the squares of all the elements), or $\mathcal{J}(W, H|V) = \sum_{i,j} V_{i,j} \log\left(\frac{V_{i,j}}{\Lambda_{i,j}}\right) - (V_{i,j} - \Lambda_{i,j})$ for the $\mathcal{I}$-divergence. For the least-squares error, the updates can be written as

$$W^{\tau} \leftarrow W^{\tau} \bullet \frac{\sum_{\phi} \overset{\uparrow \phi \rightarrow \tau \, T}{V} H^{\phi}}{\sum_{\phi} \overset{\uparrow \phi \rightarrow \tau \, T}{\Lambda} H^{\phi}}, \quad H^{\phi} \leftarrow H^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow \phi \, T \leftarrow \tau}{W^{\tau}} V}{\sum_{\tau} \overset{\downarrow \phi \, T \leftarrow \tau}{W^{\tau}} \Lambda}, \tag{5.26}$$

while for the $\mathcal{I}$-divergence they become

$$W^{\tau} \leftarrow W^{\tau} \bullet \frac{\sum_{\phi} \left(\frac{V}{\Lambda}\right)^{\uparrow \phi} \overset{\rightarrow \tau \, T}{H^{\phi}}}{\sum_{\phi} 1 \; \overset{\rightarrow \tau \, T}{H^{\phi}}}, \quad H^{\phi} \leftarrow H^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow \phi \, T \leftarrow \tau}{W^{\tau}} \left(\frac{V}{\Lambda}\right)}{\sum_{\tau} \overset{\downarrow \phi \, T}{W^{\tau}} 1}, \tag{5.27}$$

where $\bullet$ denotes the Hadamard product, i.e., element-wise matrix multiplication.

5.4.2 Relation between NMF2D and Specmurt analysis

We note that the philosophy of NMF2D is very close to Specmurt analysis [169], which is defined as the inverse Fourier transform of linear power spectrum with log-scaled frequency. Based on the same assumption that all tones in a polyphonic sound have a common harmonic pattern and that the acoustic signal is sparse enough to assume additivity in the magnitude domain, the sound spectrum can be modeled in the log-frequency domain as the

convolution of a common harmonic structure and the distribution density of the fundamental frequencies of multiple tones. The fundamental frequency distribution can be found by deconvolving the observed spectrum with the assumed common harmonic structure, where the common harmonic structure is given heuristically or quasi-optimized through an iterative algorithm. This deconvolution of the power spectrum as a common harmonic pattern and a fundamental frequency distribution in time and frequency could also be obtained through NMF2D with $\tau = \{0\}$ by enforcing the initial W to have a harmonic structure. The initial setting for W is in general random, but by allowing only rows with indices for example $\{0, \log 2, \log 3, \ldots, \log N\}$, where N is the number of harmonics considered, to have non-zero initial values, the harmonic structure of W is preserved by the multiplicative updates (5.26) and (5.27) along the optimization procedure, and should lead to results close to the iterative algorithm derived in [169] for Specmurt analysis, although the approaches are very different.

Similarly, general NMF2D with $\tau = \{0, \ldots, T\}$ can be considered very close to 2D Specmurt analysis, introduced in [168]. The common harmonic structure is there generalized to take into account the evolution of the spectral structure in the time direction, in the same way as NMFD and NMF2D generalized over NMF. One should thus as well be able to apply the method described in this chapter to adapt Specmurt and 2D specmurt for missing-data problems, although we shall not investigate further this possibility here and only derive the algorithm for NMF2D.

5.4.3 NMF2D on incomplete spectrograms

We consider the wavelet magnitude spectrogram of an acoustical scene represented as a non-negative matrix $V_{i,j}$, defined on a domain of definition $D = [\![1, M]\!] \times [\![1, N]\!]$ (corresponding for example to the time-frequency region $\{x, t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$, sampled in time and frequency). We assume in general that the spectro-temporal patterns to be modeled are roughly pitch-invariant, and that the signals are sparse enough such that the additivity assumption on the magnitude spectrograms holds.

We assume that some regions of the magnitude spectrogram are degraded or missing and are interested in performing simultaneously an analysis of this acoustical scene with the NMF2D algorithm despite the presence of gaps, and a reconstruction of the missing parts.

Even if the data matrix V is incomplete, i.e., if the values $V_{i,j}$ are missing or considered not reliable for some indices $(i, j) \in J \subset D$, due to the fact that the NMF2D update equations (as well as, more generally, NMF update equations) are in fact multiplicative versions of a gradient update, it would actually be possible to still perform the minimization

of the distance taken over the observed data by computing the gradient of this restricted objective function. However, the formulation of the update equations would then become more intricate and less obvious to interpret, and, although the updates could be originally computed simply and efficiently using FFT thanks to their convolutive nature, their missing-data version would require an additional "dirty" trick in order to compute them in the same way (concretely, setting to zero the values of the term against which H or W are convolved in the denominators of (5.26) and (5.27) where data is actually missing before computing their FFT). In any case, using the method introduced in Section 5.2 is cleaner and easier to interpret, more systematic and general. Finally, the simplicity and ease of interpretation of NMF2D make it a good example to illustrate the general principle we presented.

Applying the method introduced in Section 5.2.3 to NMF2D leads to the following algorithm, which can be used to analyze incomplete spectrograms, with both objective functions:

Step 1 $V_{i,j}^{(p+1)} = \begin{cases} \Lambda_{i,j}^{(p)} & \text{if } (i,j) \in J \\ V_{i,j} & \text{if } (i,j) \notin J \end{cases}$

Step 2 Update W and H through (5.26) or (5.27)

5.4.4 Sparseness as a key to global structure extraction

A sparseness term can be added to the NMF2D objective function, in the form of the L^1 norm of the matrix H, leading to the so-called Sparse NMF2D (SNMF2D) [141]. As pointed out by Mørup and Schmidt, there is an intrinsic ambiguity in the decomposition (5.25). The structure of a factor in H can to some extent be put into the signature of the same factor in W and vice versa. Imposing sparseness on H forces the structure to go into W and thus alleviates this ambiguity. In the case of spectrograms with gaps, this is even more critical, and sparseness becomes compulsory. Indeed, without a sparseness term, assuming that the spectral envelopes were time and pitch invariant (which is only approximately true), a perfect reconstruction of the spectrogram with gaps could be obtained with a single frame representing the spectral envelope template in W and the power envelope in the time direction (again, gaps included) in H. The role of sparseness is thus to ensure that global structures are extracted and used throughout the spectrogram, and it will be the key that will enable us to fill in the gaps in the spectrogram.

5.4.5 Use of prior distributions with SNMF2D

The NMF framework can be considered in a Bayesian way based on the correspondence between Bregman divergence based optimization and ML estimation either for the least-

squares error or the $\mathcal{I}$-divergence. Indeed, the NMF objective function can be converted into a log-likelihood [118, 170], to which prior constraints on the parameters can further be added [37, 79].

Sparseness terms involving L^p norms of H can be considered as such, the L^1-norm sparseness term used here corresponding for example to a Laplace distribution.

But one can also introduce Markovian constraints on the parameters to ensure smooth solutions. Using Gamma chains on the coefficients of W and H in the time direction, one can show that analytical update equations can still be obtained and the objective function can be optimized based on the Expectation-Constrained Maximization (ECM) algorithm [138].

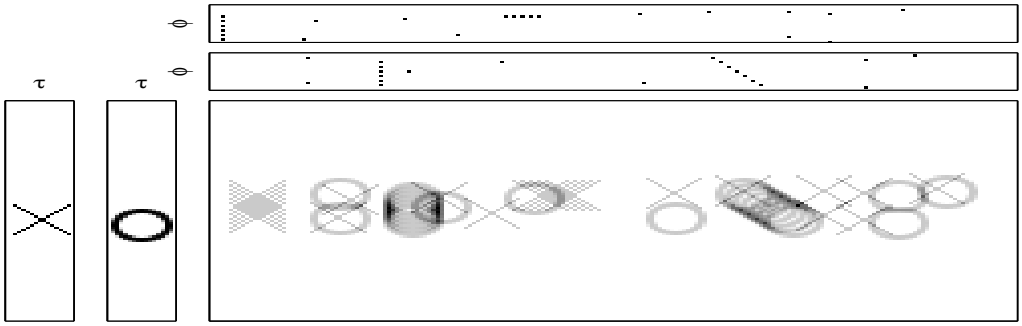
5.4.6 Toy example: reconstructing a 2D image

We first tested our algorithm on simulated data used by Mørup and Schmidt in [141]. The data, shown in Fig. 5.3 (a), were created with W consisting of one cross in the first factor and one circle in the second, convolved with H given in the top of the figure to yield the full data matrix V. The SNMF2D algorithm was used in the same conditions as in [141], with $\tau = \{0, \ldots, 16\}$ and $\phi = \{0, \ldots, 16\}$. The circle and cross templates span roughly 15 frames in both horizontal and vertical directions, while the whole data is 200 frames wide. To construct the incomplete data, we erased 3 frames horizontally and 2 frames every 10 frames vertically, as shown in Fig. 5.3 (b). Note that none of the occurrences of the structures (circle and cross) is fully available. However, in this ideal case where the original data is a strict convolution of the patterns, the proposed algorithm is able to extract the original patterns and their occurrences and to reconstruct the original data, as can be seen in Fig. 5.3 (c) (least-squares update equations) and Fig. 5.3 (d) ($\mathcal{I}$-divergence update equations). This shows that the reconstruction is based on global features of the data.

5.4.7 Audio example: reconstructing gaps in a sound

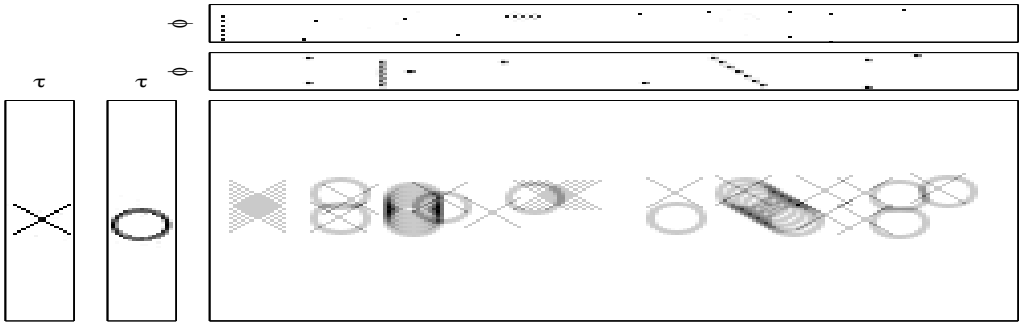
Experimental setting

For auditory restoration experiments, contrary to what is done in [176], we did not use the short time Fourier transform afterwards converted into a log-frequency magnitude spectrogram, but a wavelet transform, which directly gives a log-frequency spectrogram. More precisely, the magnitude spectrogram was calculated from the input signals digitized at a 16 kHz sampling rate using a Gabor wavelet transform with a time resolution of 16 ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest

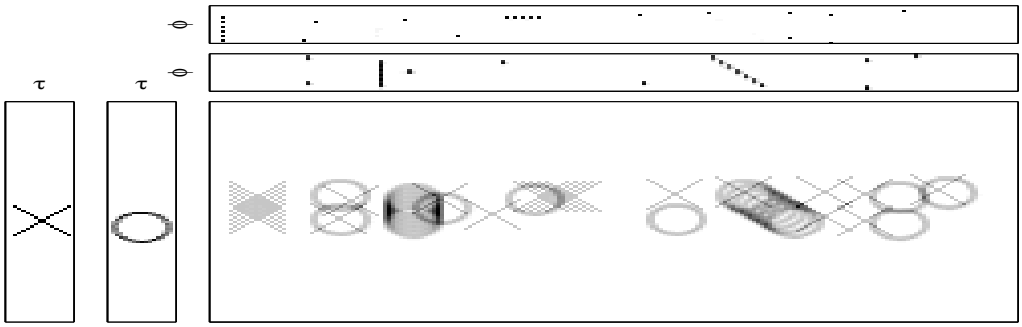


(a) Original simulated data.

(b) Truncated data with truncated regions in black.



(c) Reconstruction using least-squares.



(d) Reconstruction using the $\mathcal{I}$-divergence.

Figure 5.3: *NMF2D with missing data on a toy problem. (a) Original simulated data. W consists of one cross in the first factor and one circle in the second. They are convolved with H given in the top of the figure to yield the full data matrix V. (b) Truncated data. The truncated areas are indicated in black. (c) Estimated factors and reconstructed image using the least-squares algorithm. (d) Estimated factors and reconstructed image using the $\mathcal{I}$-divergence algorithm.*

Table 5.1: *Results of the reconstruction experiment*

	SNR		SSNR	
	in	out	in	out
MI / C	10.7	12.9	10.5	11.7
MI / I	-3.7	13.1	3.4	12.1
SC / C	13.1	13.0	12.3	11.9
SI / I	6.2	10.5	7.3	9.9
I / C	2.2	15.7	2.4	16.9

subband resolution. The frequency range extended from 50 Hz to 8 kHz and was covered by 200 channels, for a frequency resolution of 44 cent.

We used a 4.8 s piece of computer generated polyphonic music containing a trumpet and a piano, already used by Schmidt and Mørup in [176]. Its spectrogram can be seen in Fig. 5.4 (a). The incomplete waveform was built by erasing 80 ms of signal every 416 ms, leading to a signal with about 20 % of data missing. Its spectrogram is shown in Fig. 5.4 (b).

The mask indicating the region J to inpaint was built according to the erased portions of the waveform. With a Gabor wavelet transform, the influence of a local modification of the signal theoretically spans the whole interval. However, as the windows are Gaussian, one can consider that the influence becomes almost null further than about three times the standard deviation. This standard deviation is inversely proportional with the frequency, and the influence should thus be considered to span a longer interval for lower frequencies. Although it leaves some unreliable portions of the spectrogram out of the mask in the lower frequencies, for simplicity, we did not consider here this dependence on frequency, and simply considered unreliable, for each 80ms portion of waveform erased, 6 whole spectrogram frames (corresponding to about 96ms of signal in the highest frequencies). The incomplete spectrogram is shown in Fig. 5.4 (c), with areas to inpaint in black.

The SNMF2D parameters were as follows. As in [176], we used two factors, $d = 2$, since we are analyzing a scene with two instruments, and the number of convolutive components in pitch was set to $\phi = \{0, \ldots, 11\}$, as the pitch of the notes in the data spans three whole notes. For the convolutive components in time, we used empirically $\tau = \{0, \ldots, 31\}$, for a time range of about 500 ms, thus roughly spanning the length of the eighth notes in the music sample. The $\mathcal{I}$-divergence was used as the distortion measure, and the sparseness term coefficient set to 0.001. The algorithm was ran for 100 iterations.

Results

To evaluate the reconstruction accuracy of the spectrogram, we use two measures: Signal to Noise Ratio (SNR) and Segmental SNR (SNR) computed as the median of the individual SNRs of all the frames. We note that computing the SNR directly on the magnitude spectrogram amounts to assuming that the phase is perfectly reconstructed. The results are summarized in Table 5.1, where "in" refers to the measure computed inside the gaps (the inpainted part), "out" to the measure computed outside the gaps (the part more classically reconstructed based on observed data), "M" refers to the proposed Missing-data SNMF2D, "S" to SNMF2D on the whole data with missing data (if any) assumed to be zero, "C" to the magnitude spectrogram of the complete waveform, and "I" to the one of the incomplete waveform. Finally, "WX" refers to the spectrogram reconstructed by applying algorithm W on spectrogram X, and "Y/Z" to the comparison of spectrogram Y with spectrogram Z as a reference. For example, the SNR of "MI/C" is the SNR of the spectrogram reconstructed using our missing-data approach on the spectrogram of the incomplete data w.r.t. the spectrogram of the full waveform.

One can see through MI/I that the proposed algorithm correctly performs its task of reconstructing the observed data ("out"), which is not the case for SI/I, showing that, as expected, SNMF2D cannot perform well if the region where optimization is carried out includes the gaps. The MI/C results show that the formerly erased regions ("in") are correctly inpainted, with a great improvement over the incomplete spectrogram, as seen in I/C, and that our method performs closely to SNMF2D applied on the complete spectrogram, as seen in SC/C.

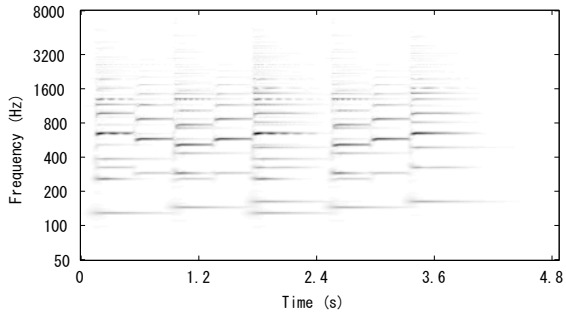
Graphical results are shown in Fig. 5.4 (d), (e), (f), where one can see in particular that the acoustical scene analysis (i.e., the learning of a spectro-temporal pattern for each instrument and the estimation of the pitch and onset time of each note) is performed correctly, and that blind source separation is also performed in spite of the presence of gaps.

5.5 Missing-data HTC

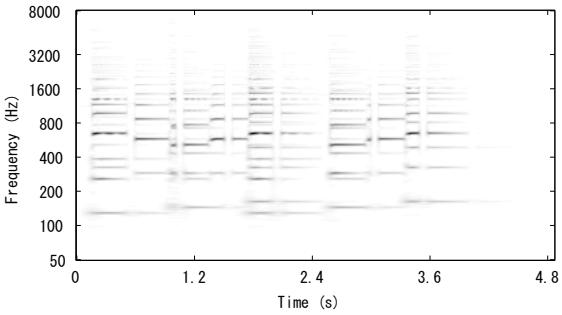
5.5.1 Formulation of the model on incomplete data

The optimization process in HTC, presented in Chapter 3, is nothing else than the fitting of a model (in particular a constrained Gaussian mixture model) to an observed distribution (the wavelet power spectrum of an acoustical scene), using the $\mathcal{I}$-divergence as a measure of the goodness of fit.

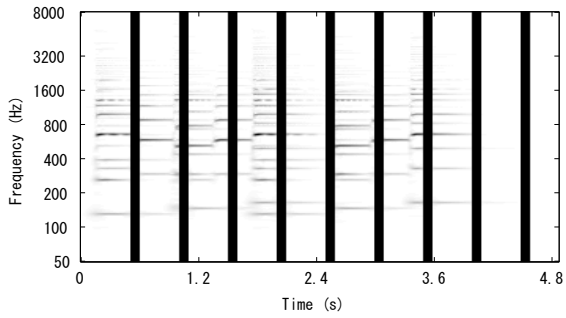
If some parts of the power spectrum are missing or corrupted, or if some parts of the HTC



(a) Spectrogram of the original waveform.

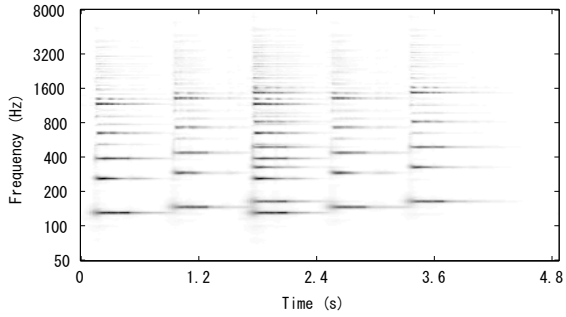


(b) Spectrogram of the truncated waveform.

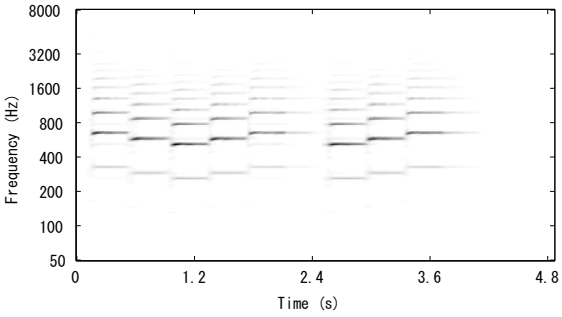


(c) Truncated spectrogram with areas to inpaint in black.

(d) Spectrogram reconstructed using the $\mathcal{I}$-divergence.



(e) Reconstructed and separated spectrogram of the piano part.



(f) Reconstructed and separated spectrogram of the trumpet part.

Figure 5.4: *NMF2D with missing data on the spectrogram of a truncated waveform. (a) Spectrogram of the original waveform (a mixture of piano and trumpet sounds). (b) Spectrogram of the truncated waveform. (c) Truncated spectrogram, with truncated regions indicated in black. (d) Estimated factors and reconstructed spectrogram using the $\mathcal{I}$-divergence algorithm. (e) Reconstructed and separated spectrogram of the piano part. (f) Reconstructed and separated spectrogram of the trumpet part.*

model are partially or entirely lying outside the boundaries of the spectrogram (for example if some harmonics of the model are above the maximum frequency and a prior is used to link the powers of the harmonics, fitting the upper harmonics to zero will bias the optimization), the estimation of the HTC model must be performed under an incomplete-data framework, as in Section 5.2.3. In the same way as we showed there, optimization can be performed in an iterative way by using the values of the model at the previous step as an estimation of the unobserved data. In the case of HTC, this results in a hierarchical algorithm with two levels. At the upper level is the iterative algorithm described above. At the lower level, inside the step 2 of the upper level, the EM algorithm is used as in the classical formulation of the HTC optimization. Let W be the observed part of the spectrogram and $I \subset D$ its domain of definition. The objective function to minimize here is the same as (3.3) but restricted to the domain where the spectrogram is observed:

$$\mathcal{I}(W, Q(\boldsymbol{\Theta})) \triangleq \iint_I \left(W(x,t) \log \frac{W(x,t)}{Q(x,t;\boldsymbol{\Theta})} - \Big(W(x,t) - Q(x,t;\boldsymbol{\Theta}) \Big) \right) dx\, dt. \tag{5.28}$$

We define the auxiliary function as

$$\begin{aligned} \mathcal{I}^+(W, V, Q(\boldsymbol{\Theta})) \triangleq\, & \mathcal{I}(W, Q(\boldsymbol{\Theta})) \\ & + \iint_{D \setminus I} \Bigg(V(x,t) \log \frac{V(x,t)}{Q(x,t;\boldsymbol{\Theta})} \\ & \qquad - \Big(V(x,t) - Q(x,t;\boldsymbol{\Theta}) \Big) \Bigg) dx\, dt. \end{aligned} \tag{5.29}$$

Then membership functions m can be further introduced as in the classical formulation of HTC to build the final auxiliary function $\mathcal{I}^{++}(W, V, Q(\boldsymbol{\Theta}), m)$. These membership functions are non-negative and sum up to 1 for each (x,t): $\sum_{k,n,y} m_{kny}(x,t) = 1$. If we note

$$Z(x,t) = \begin{cases} W(x,t) & \text{if } (x,t) \in I \\ V(x,t) & \text{if } (x,t) \in D \setminus I \end{cases}$$

we define

$$\begin{aligned} \mathcal{I}^{++}(W, V, Q(\boldsymbol{\Theta}), m) \triangleq & \\ & \iint_D \Bigg(\sum_{k,n,y} m_{kny}(x,t) Z(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t) Z(x,t)} \\ & \qquad - \Big(Z(x,t) - Q(x,t;\boldsymbol{\Theta}) \Big) \Bigg) dx\, dt. \end{aligned} \tag{5.30}$$

Using the concavity of the logarithm, one can see that

$$\mathcal{I}^{+}(W, V, Q(\boldsymbol{\Theta})) \leq \mathcal{I}^{++}(W, V, Q(\boldsymbol{\Theta}), m) \tag{5.31}$$

with equality for

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}. \tag{5.32}$$

Altogether, the optimization process can be formulated as follows.

Step 1 Estimate V such that $\mathcal{I}(W, Q(\boldsymbol{\Theta})) = \mathcal{I}^{+}(W, V, \boldsymbol{\Theta})$:

$$\hat{V}(x,t) = Q(x,t;\boldsymbol{\Theta}), \forall (x,t) \in D \setminus I. \tag{5.33}$$

Step 2 Update $\boldsymbol{\Theta}$ with $\hat{V}$ fixed:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, \mathcal{I}^{+}(W, \hat{V}, \boldsymbol{\Theta}). \tag{5.34}$$

To do so, perform one iteration of the classical formulation of HTC:

E-Step

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}, \tag{5.35}$$

M-Step

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \Big(\mathcal{I}^{++}(W, \hat{V}, \boldsymbol{\Theta}, \hat{m}) - \log P(\boldsymbol{\Theta}) \Big) \tag{5.36}$$

where $P(\boldsymbol{\Theta})$ is a prior distribution on the parameters.

5.5.2 Optimization of the model

Closed-form update equations

Analytical update equations for the M-step are derived in [108]. However, when the F_0 contour is modeled using cubic spline functions, which is relevant for speech or musical instruments whose pitch can vary continuously, the spline parameters were updated in Chapter 3 and Chapter 4 not globally but one after the other. The corresponding optimization procedure, called the Expectation-Constrained Maximization algorithm (ECM) [138], does not ensure the minimization in the M-step but nonetheless guarantees the decrease of the objective function. This spline parameter update was thus not optimal but yet led to very good results in F_0 estimation accuracy. However, it suffered from some instability problems

in long regions with low harmonic energy (silence or unvoiced parts of speech for example). When dealing with missing-data problems, such issues become critical, and we thus need to use better update equations for the spline parameters, which we present in detail in the following. Contrary to the update equations previously described in Chapter 3, the ones presented here are global analytical update equations which lead to the minimum of the auxiliary function in the M-step. They ensure a greater stability of the spline model and a better F_0 estimation accuracy, as shown in the next section.

Spline contour

The analysis interval is divided into subintervals $[t_i, t_{i+1})$ of equal length ϵ. The parameters of the spline contour model are then the values z_i of the F_0 at each bounding point t_i. Assuming that the second derivative vanishes at the bounds of the analysis interval leads to the so-called natural splines. Under this assumption, one can explicitly compute offline a matrix M linking the values z_i'' of the second derivative of the contour at t_i with the values z_i, such that $\mathbf{z}'' = M\mathbf{z}$. An analytical expression for the contour $\mu(t;\mathbf{z})$ is obtained as in (3.18) (the matrix M depends on the boundary conditions: there we used zero derivatives, and here we use zero second derivatives). One can notice that the expression of $\mu(t;\mathbf{z})$ is actually linear in $\mathbf{z}$:

$$\mu(t;\mathbf{z}) = \mathbf{A}(t)^T\mathbf{z} \tag{5.37}$$

where $\mathbf{A}(t)$ is a column vector such that, for $t \in [t_i, t_{i+1})$,

$$\begin{aligned}\mathbf{A}(t) = \frac{1}{t_{i+1}-t_i}\Big((t_{i+1}-t)\mathbf{e}_i + (t-t_i)\mathbf{e}_{i+1} \\ -\tfrac{(t-t_i)(t_{i+1}-t)}{6}\big[(t_{i+2}-t)\mathbf{M}_i^T + (t-t_{i-1})\mathbf{M}_{i+1}^T\big]\Big)\end{aligned} \tag{5.38}$$

where $\mathbf{M}_j$ denotes the j-th row of the matrix M and $\mathbf{e}_j$ denotes the j-th vector of the canonical basis. We note furthermore that $\mathbf{A}(t) = \nabla_{\mathbf{z}}\,\mu(t;\mathbf{z})$.

Optimization of the objective function

During the M-step of the EM algorithm, one wants to minimize $\mathcal{J}(\mathbf{\Theta}) = \mathcal{I}^{++}(W, \hat{V}, \mathbf{\Theta}, \hat{m}) - \log P(\mathbf{\Theta})$ with respect to $\mathbf{\Theta}$. We can compute the gradient with respect to $\mathbf{z}$:

$$\begin{aligned}\nabla_{\mathbf{z}}\mathcal{J} &= -\iint_D \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \mu(t,\mathbf{z}) - \log n)\mathbf{A}(t)\,dx\,dt - \nabla_{\mathbf{z}}\log P(\mathbf{\Theta}) \\ &= -\iint_D \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \mathbf{A}(t)^T\mathbf{z} - \log n)\mathbf{A}(t)\,dx\,dt - \nabla_{\mathbf{z}}\log P(\mathbf{\Theta})\end{aligned}$$

where $\ell_{kny}(x,t) = m_{kny}(x,t)Z(x,t)$. Note that the term $\iint_D Q(x,t;\mathbf{\Theta})\,dx\,dt$ in (5.30) does not contribute to the gradient w.r.t. $\mathbf{z}$ as the spline parameters do not influence the normalization of the model.

Let

$$\begin{aligned}\phi(t) &= \int \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}(x - \log n)\,dx,\\ \gamma(t) &= \int \sum_{k,n,y} \frac{\ell_{kny}(x,t)}{\sigma_k^2}\,dx.\end{aligned}$$

Then

$$\nabla_{\mathbf{z}}\mathcal{J} = -\int \phi(t)\mathbf{A}(t)\,dt + \left(\int \gamma(t)\mathbf{A}(t)\mathbf{A}(t)^T\,dt\right)\mathbf{z} - \nabla_{\mathbf{z}}\log P(\mathbf{\Theta}).$$

One can then obtain the Hessian matrix:

$$H_{\mathbf{z}}\mathcal{J} = \int \gamma(t)\mathbf{A}(t)\mathbf{A}(t)^T\,dt - H_{\mathbf{z}}\log P(\mathbf{\Theta}). \tag{5.39}$$

If one uses a Markov assumption on the spline parameters with Gaussian distributions for the state transitions, the prior distribution becomes

$$P(\mathbf{z}) = P(z_0)\prod_{j=1}^{|\mathbf{z}|} P(z_j|z_{j-1}),$$

with z_0 following a uniform distribution and

$$P(z_j|z_{j-1}) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(z_j - z_{j-1})^2}{2\sigma_s^2}}.$$

Then

$$\begin{aligned}\nabla_{\mathbf{z}}\log P(\mathbf{\Theta}) &= -\frac{1}{\sigma_s^2}\begin{pmatrix} 1 & -1 & & & 0\\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1\\ 0 & & & -1 & 1\end{pmatrix}\mathbf{z}\\ &= H_{\mathbf{z}}\log P(\mathbf{\Theta})\mathbf{z}.\end{aligned} \tag{5.40}$$

Putting to 0 the gradient w.r.t. $\mathbf{z}$, one can find the update equation for $\mathbf{z}$:

$$\mathbf{z} = (H_{\mathbf{z}}\mathcal{J})^{-1}\int \phi(t)\mathbf{A}(t)\,dt. \tag{5.41}$$

The convexity can be studied by looking at $H_{\mathbf{z}}\mathcal{J}$ in Eq. (5.39). The first term is indeed non-negative, as $\gamma(t) \geq 0, \forall t$. For the second term, coming from the prior distribution, we recognize a tridiagonal matrix, for which the principal minors can be easily calculated. If $T = (t_{ij})$ is a tridiagonal matrix and α_n its n-th principal minor, then

$$\alpha_n = t_{n,n}\alpha_{n-1} + t_{n,n-1}t_{n-1,n}\alpha_{n-2}. \tag{5.42}$$

In our case, we see that the principal minors of $H_{\mathbf{z}} \log P(\mathbf{\Theta})$ are all non-positive. The matrix $-H_{\mathbf{z}} \log P(\mathbf{\Theta})$ is thus positive semi-definite. Altogether, $H_{\mathbf{z}}\mathcal{J}$ is at least positive semi-definite, and the update (5.41) thus corresponds to a minimum.

5.5.3 F_0 estimation on incomplete data with HTC

Importance of F_0 estimation accuracy

Ensuring a very good accuracy for the F_0 estimation is not only important as a necessary step for computational auditory induction by HTC, but it is also in itself a primary issue. Indeed, being able to estimate the F_0 accurately is important as well for some previous audio interpolation methods such as Maher's sinusoidal model based method [130], in which the harmonics before and after the gap need to be linked, or Vaseghi and Rayner's extended AR model [196], which takes advantage of the long-term correlation structure of the signals by introducing extra predictor parameters around the pitch period.

Relevance of HTC's F_0 contour model

When applied to speech, HTC is based on a spline F_0 contour. The Markovian prior presented above is used on the parameters of the contour to ensure that it will not move too abruptly. This Markovian prior penalizes the deviation of a spline parameter from the linear interpolation of its neighbors. Altogether, HTC's F_0 contour model is somewhere between a spline interpolation and a linear interpolation, depending on the strength of the matching between the HTC source model and the observed data.

Attempting to use this model for reconstruction of incomplete data implies that the F_0 contour inside the gap is close to an interpolation based on the values of the contour outside the gap. To confirm the relevance of this interpolation, we thus need to ensure that, assuming that the F_0 estimation on complete parts of the data is accurately performed, the F_0 of the missing parts of the data is accurately performed as well.

We thus evaluated as a preliminary experiment the accuracy of the F_0 contour obtained by interpolating the reference F_0 values outside the gaps on the whole interval using both

natural splines and linear interpolation. This can be considered as an evaluation of what can be expected by HTC.

We then conducted experiments to confirm the accuracy of the proposed method for F_0 estimation. The goal here was first to confirm that the new spline update equations indeed outperform the former update equations on single-speaker F_0 estimation in clean environment for complete data, then to evaluate the F_0 accuracy with parts of the data missing.

Experimental setting

The general conditions of the experiments were exactly the same as in Section 4.2.1. For the incomplete data, we prepared two sets of data by replacing segments of the utterances by silence. The first set, which we shall denote by Erase-50ms, was produced by erasing 50 ms every 250 ms of data (the obtained utterances would then be successions of 50 ms of silence, 200 ms of speech, 50 ms of silence, etc.), thus leading to utterances with approximately 20 % of the data missing. The second set, which we shall denote by Erase-100ms, was produced by erasing 100 ms every 500 ms of data (the obtained utterances would then be successions of 100 ms of silence, 400 ms of speech, 100 ms of silence, etc.), thus leading again to utterances with approximately 20 % of the data missing. We shall note that gaps from 30 ms to 50 ms are already considered very long gaps in the audio restoration literature [76, 86, 130].

Preliminary experiment on F_0 interpolation

We first performed a preliminary experiment based on the reference F_0 and the reliability mask derived in [44]. The reliability mask was used to determine the voiced regions of the speech utterance, and a global contour over the whole utterance was derived by interpolating the values of the F_0 reference which were both inside the reliability mask and outside the erased segments of the data. We used both linear interpolation and cubic spline interpolation. We then computed the gross error rates of the interpolated F_0 values inside the gaps (by construction the values outside the gaps are equal to the reference and thus no error can occur there). Results for both incomplete data sets can be seen in Table 5.2. Spline interpolation does not perform as well as linear interpolation due to the large variations that can occur depending on the slope of the contour at the beginning or end of a voiced region. This is precisely what the Markovian prior in HTC's F_0 contour model aims to avoid.

Accuracy on complete data

We first used the classical HTC formulation on complete data, using the new spline update equations. Here, only step 2 of the algorithm devised in 5.5.1 is thus used (iteration of

Table 5.2: *Gross error rates (%) for F_0 interpolation inside the gaps based on reference F_0*

Data set	Cubic Splines	Linear interpolation
Erase-50ms	1.2	0.5
Erase-100ms	6.9	2.9

Table 5.3: *Gross error rates for F_0 estimation on complete data (clean single-speaker speech)*

Method	Gross error (%)
YIN	2.2
HTC (former spline update)	3.5
HTC (proposed spline update)	1.2

equations (5.35) and (5.36)).

The results can be seen in Table 5.3, with for comparison the results obtained with HTC using the former spline update equations as well as the ones obtained with the state-of-the-art algorithm YIN [44]. We note that we obtained 2.2 % gross error rate for YIN using the code made available by its authors, as opposed to 1.3 % reported in the original paper. We can see that HTC with the newly proposed spline update equations now performs comparably to YIN.

Accuracy on incomplete data

The wavelet transforms were performed on the truncated waveforms of the data sets Erase-50ms and Erase-100ms introduced above. The regions $D \setminus I$ which are to be considered missing in the spectrogram were defined as the frames corresponding to the erased parts of the waveform. The influence of the erased portion is larger for low frequencies, but we neglect this and consider missing a whole frame regardless of the frequency bin.

In such situations where part of the data is irrelevant, one might think that algorithms which perform F_0 estimation more locally should be used, using interpolation between the preceding and following voiced portions to obtain F_0 values inside the gap. If the estimation can be accurately performed outside the gaps, such a method should lead very good results. However, one needs to note that if such algorithms are used, a robust Voice Activity Detection (VAD) must be performed as well to determine which points should be used in the interpolation. A poor VAD accuracy could lead to very bad results in the interpolation process. To illustrate this and as a comparison with HTC, we used the algorithm YIN to perform F_0 estimation outside of the gaps, and used a linear interpolation to obtain values

Table 5.4: *Gross error rates for F_0 estimation on incomplete data with HTC and YIN (clean single-speaker speech).*

	HTC (YIN)	
Data set	Error inside the gaps (%)	Total error (%)
Erase-50ms	7.9 (10.7)	3.5 (6.7)
Erase-100ms	20.9 (14.4)	6.4 (7.3)

inside the gaps. The positions of the gaps were given, and the voiced regions were determined using the aperiodicity measure given by YIN, with a threshold of 0.2. The obtained F_0 contour was extended with constant values before the first detected voiced region and after the last detected voiced region. The results given here were obtained using linear interpolation, but cubic spline interpolation gave similar results.

Results for HTC and YIN are given in Table 5.4, with gross error rates for the whole file as well as for the erased segments only.

We can see that the performance of HTC degrades as the gaps become longer, while remaining satisfactory. HTC performs better than the algorithm based on YIN for the total accuracy as well as for the accuracy inside the gaps with 50 ms erased segments, while the algorithm based on YIN performs better inside the gaps with 100 ms erased segments.

Altogether, this shows that HTC's F_0 estimation accuracy is very good even in the presence of long gaps, and that, although other F_0 estimation algorithms could be used as well, it is not obvious, regardless of their performance on complete data, whether they can be turned into effective algorithms on incomplete data, due in particular to the importance of a robust VAD for the interpolation to be effective.

5.6 Summary of Chapter 5

We presented a computational framework to model auditory induction, i.e., the human auditory system's ability to estimate the missing parts of a continuous auditory stream briefly covered by noise, by extending acoustical scene analysis methods based on global statistical models such as HTC and SNMF2D to handle unobserved data. We related the method to the EM algorithm, enabling the use of priors on the parameters. We illustrated on a simple example how the proposed framework was able to simultaneously perform acoustical scene analysis and gap interpolation in a music piece with SNMF2D, and how a robust F_0 estimation could be performed on incomplete data with HTC.

Chapter 6

Consistency and inconsistency in complex STFT spectrograms

6.1 Introduction

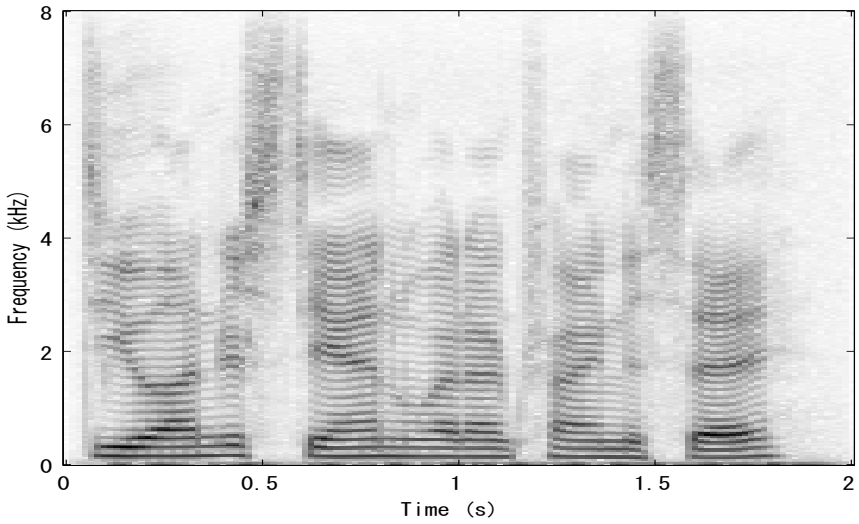
In the previous chapters, we have presented algorithms to perform the analysis of acoustical scenes in the time-frequency power (or magnitude) domain: speech enhancement and speaker separation in Chapter 4, missing-data reconstruction in Chapter 5. More generally, many acoustical signal processing techniques, developed for a wide range of applications such as source separation [176, 185, 203], noise canceling [33], time-scale and pitch-scale modifications or more generally audio modification [116], involve a processing of the magnitude spectrogram, whether it be a short-time Fourier transform (STFT) spectrogram, or a spectrogram obtained using other transforms such as constant-Q transforms for example.

Although discarding the phase has the advantage that we do not need to worry about its (arguably intricate) modeling, it also obviously means that we are throwing away some information and losing the correspondence between the time domain (the waveform) and the complex time-frequency domain (for example the complex STFT spectrogram, which we shall focus on in this chapter). This raises several issues: first, the additivity of signals in a mixture is not true in the power domain, as cross-terms are usually non-zero; second, phase information is missing, and needs to be reconstructed in some way if resynthesis of a time-domain signal is desired; third, phase information may actually be relevant and worth being exploited.

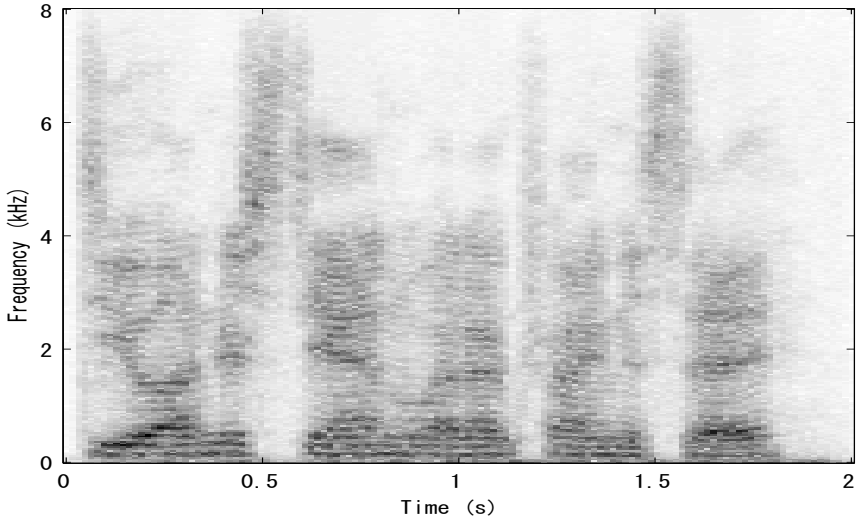
Dealing with these three issues can be a motivation to work either in the time domain, as we shall investigate in Chapter 7, or in the complex time-frequency domain: under some

conditions on the analysis and synthesis windows which we shall review in this chapter, there is indeed a perfect equivalence between a time-domain signal and its complex spectrogram constructed using the STFT. Moreover, the STFT being a linear transform, the additivity of waveforms in a mixture still holds true in the complex time-frequency domain. However, as the STFT representation is obtained from overlapping frames of a waveform, it is redundant and has a particular structure. Thus, starting from a set of complex numbers in the complex time-frequency domain, it is not guaranteed whether there exists a signal in the time domain whose STFT is equal to that set of complex numbers. Therefore, if we were to work in the complex time-frequency domain, for example performing source separation based on some hypothetical model of the complex spectrogram of a sound, we would need to ensure that the separated complex spectrograms are all proper spectrograms (we shall call them "consistent spectrograms"), i.e., that they all are the STFT of some time-domain signal. Carefully studying the structure of complex STFT spectrograms and finding a way to ensure that sets of complex numbers indeed respect this structure is thus a crucial issue.

On the other side, we can choose to work in the magnitude or the power domain, implicitly assuming that additivity in that domain is approximately true. If resynthesis is desired, we need to adjoin some phase to the estimated magnitude to reconstruct a signal, and this phase needs to be coherent with the magnitude in order to produce a perceptually satisfactory sounding signal. However, phase information is usually not available. In many situations, such as time-frequency-domain-based methods for time-scale modification [116] or for reconstruction of missing parts of an acoustic signal, for example in the computational auditory induction framework we presented in Chapter 5, phase must be partially or totally reconstructed. Sometimes, as in source separation, based on a sparseness assumption on the repartition of acoustic energy in the time-frequency space, the phase of a mixture can be used as a rough estimation of the phase when reconstructing each extracted component using the estimated magnitude spectrograms. However, in both cases, incoherences between the phase and the magnitude spectrogram from which we want to reconstruct a signal lead in general to perceptually disturbing artifacts. Moreover, the magnitude spectrogram of the reconstructed signal may actually be very different from the one we intended to reconstruct a signal from, as illustrated in Figure 6.1: Fig. 6.1 (a) shows the magnitude of the STFT spectrogram of the utterance "Do you understand what I'm trying to say?" by a male speaker, taken from [111], and Fig. 6.1 (b) shows the magnitude of the STFT spectrogram of the signal obtained as the inverse STFT of the original magnitude spectrogram with random phase. An effective method for phase reconstruction would thus have many applications and broaden the range



(a) Spectrogram of the original utterance



(b) Spectrogram of signal obtained by randomizing the phase

Figure 6.1: *Illustration of the influence of an incorrect phase on the magnitude spectrogram of the resynthesized signal. (a) Magnitude spectrogram of the original utterance "Do you understand what I'm trying to say?" by a male speaker. (b) Magnitude spectrogram of the signal reconstructed by performing the inverse STFT of the original spectrogram combined with a random phase.*

of situations where magnitude spectrogram based techniques can be applied. In order to be able to reconstruct the phase of a spectrogram such that it is as coherent as possible with a given magnitude, one first needs to understand what "coherent" means in such a context, i.e., to understand the structure of complex STFT spectrograms, and then to find a way to estimate a phase such that its combination with a given magnitude leads to a set of complex numbers which respects that structure as well as possible. Moreover, even if resynthesis is not necessary, understanding the structure of complex STFT spectrograms would enable us

to ensure that the estimated magnitude spectrogram is such that there exists a consistent complex spectrogram of which it is the magnitude, which could help avoid irrealistic solutions when designing a Wiener filter or a binary mask in the magnitude or power domain.

In all these situations, we thus need to quantify the consistency of a set of complex numbers as an STFT spectrogram. In the following, we will call consistent STFT spectrogram a set of complex numbers which has been obtained as the STFT spectrogram of a real signal, and inconsistent STFT spectrogram a set of complex numbers which cannot be obtained as such. In this chapter, we derive explicit consistency constraints for STFT spectrograms as the kernel of a simple linear operator in the complex time-frequency domain with coefficients depending on the window length, the frame shift and the analysis and synthesis windows used to build the spectrogram or which the spectrogram is assumed to have been obtained from. The norm of the image of a set of complex numbers by this linear operator defines a consistency criterion, which can for example be used as a prior distribution on complex spectrograms when performing separation tasks in the complex time-frequency domain, or as an objective function on the phase when trying to recover the most coherent phase for a given magnitude spectrogram.

We will first review in Section 6.2 the perfect reconstruction conditions on the analysis and synthesis windows. Then, in Section 6.3, we will derive the consistency constraints for STFT spectrograms, explain how to define a consistency criterion and how to simplify it based on a local approximation. In Section 6.4, we will introduce an algorithm for phase reconstruction, based on the optimization of an objective function derived from the consistency criterion, and show how it can be used to develop a flexible real-time time-scale modification algorithm. Finally, in Section 6.5, we will explain how inconsistent spectrograms can be used to perform audio encryption, the synthesis window acting as a decoding key.

6.2 Perfect reconstruction constraints on the window functions

Let $(x(t))_{t\in\mathbb{Z}}$ be a digital signal. We review here the conditions for perfect reconstruction of the signal through STFT and inverse STFT [9, 89]. Let N be the window length, R the window shift, W the analysis window function and S the synthesis window function. We suppose that W and S are zero outside the interval $0 \leq t \leq N-1$. We assume that the window length N is an integer multiple of the shift R, and we note $Q = N/R$. The STFT at frame m is defined as the discrete Fourier transform (DFT) of the windowed short-time

signal $W(t-mR)x(t)$ (with the phase origin at the start of the frame, $t = mR$).

The inverse STFT procedure consists in Fourier-inverting each frame of the STFT spectrogram, multiplying each obtained (periodic) short-time signal by the synthesis window and summing together all the windowed short-time signals. On a particular frame $mR \leq t \leq mR + N - 1$, this leads to a reconstructed signal $y(t)$ given by

$$\begin{aligned} y(t) &= S(t-mR)W(t-mR)x(t) \\ &\quad + \sum_{q=1}^{Q-1} S(t-(m-q)R)W(t-(m-q)R)x(t) \\ &\quad + \sum_{q=1}^{Q-1} S(t-(m+q)R)W(t-(m+q)R)x(t) \end{aligned}$$

where the three terms on the right-hand side are respectively the contribution of the inverse transforms of frame m, the overlapping frames on the left and the overlapping frames on the right. As the contributions of frames with an index difference larger than Q do not overlap, by equating $y(t) = x(t)$ for all t, we obtain as in [89] the following necessary conditions for perfect reconstruction:

$$1 = \sum_{q=0}^{Q-1} W(t-qR)S(t-qR), \forall t. \tag{6.1}$$

6.3 Characterization of consistent STFT spectrograms

6.3.1 Derivation of the consistency constraints

Let $(H(m,n))_{0\leq m\leq M-1, 0\leq n\leq N-1}$ be a set of complex numbers, where m will correspond to the frame index and n to the frequency band index, and W and S be analysis and synthesis windows verifying the perfect reconstruction conditions (6.1) for a frame shift R. To stress the dependence of the STFT and inverse STFT on the analysis and synthesis window functions, we shall specify them as a subscript when talking about the STFT and inverse STFT as mathematical operators. STFT_W will thus denote the STFT with analysis window W, and iSTFT_S the inverse STFT with synthesis window S. When not indicated, we will assume that we use W and S respectively.

For the set H to be a consistent STFT spectrogram, it needs to be the STFT spectrogram of a signal $x(t)$. But by perfect reconstruction, this signal can be none other than the result of the inverse STFT of the set $(H(m,n))$. A necessary and sufficient condition for H to be a consistent spectrogram is thus for it to be equal to the STFT of its inverse STFT.

The point here is that, for a given window length N and a given frame shift, the operation $\mathrm{iSTFT}_S \circ \mathrm{STFT}_W$ from the space of real signals to itself is the identity, while $\mathrm{STFT}_W \circ \mathrm{iSTFT}_S$ from $\mathbb{C}^{M\times N}$ to itself is not.

Let us derive consistency constraints for STFT spectrograms based on this consideration, by explicitly stating that a spectrogram must be equal to the STFT of its inverse STFT, or in other words that it needs to be in the kernel of the linear operator from $\mathbb{C}^{M\times N}$ to itself defined by

$$\mathcal{F}(H) = \mathrm{STFT}_W \circ \mathrm{iSTFT}_S(H) - H. \tag{6.2}$$

If we focus on a single frame, this leads to the following computation. For convenience of notation, we introduce the shifted index $k = t - mR$. Let us first work out the contribution of frame m. Its inverse DFT is given by

$$h_m(k) = \frac{1}{N}\sum_{n=0}^{N-1} H(m,n) e^{j2\pi n \frac{k}{N}} \tag{6.3}$$

which is first windowed by the synthesis window $S(k)$ to recover a short-time signal $l_m(k) = S(k)h_m(k)$ that will later be overlap-added to its neighbors to obtain the inverse STFT signal $x(t)$.

Similarly, for frame $m+q$ we obtain

$$l_{m+q}(k) = \frac{1}{N} S(k-qR) \sum_{n=0}^{N-1} H(m+q,n) e^{j2\pi n \frac{k-qR}{N}}. \tag{6.4}$$

The short-time signals $l_{m+q}(k)$ are added, leading to the inverse STFT of H for $mR \le t \le mR + N - 1$. This signal is then windowed by the analysis window $W(k)$, and the DFT is computed to obtain the STFT. By equating the result to the original set $H(m,n)$, we obtain a set of equations which are the conditions we are looking for. For $0 \le n' \le N-1$,

$$\begin{aligned} H(m,n') \;=\; & \frac{1}{N}\sum_k W(k) e^{-j2\pi k \frac{n'}{N}} \\ & \Big\{ S(k) \sum_{n=0}^{N-1} H(m,n) e^{j2\pi n \frac{k}{N}} \\ & \quad + \sum_{q=1}^{Q-1} S(k+qR) \sum_{n=0}^{N-1} H(m-q,n) e^{j2\pi n \frac{k+qR}{N}} \\ & \quad + \sum_{q=1}^{Q-1} S(k-qR) \sum_{n=0}^{N-1} H(m+q,n) e^{j2\pi n \frac{k-qR}{N}} \Big\}. \end{aligned} \tag{6.5}$$

By introducing the coefficients

$$\alpha_q^{(R)}(p) = \frac{1}{N}\sum_k W(k)S(k+qR)e^{-j2\pi p\frac{k+qR}{N}} - \delta_p\delta_q, \tag{6.6}$$

where $-(N-1) \leq p \leq N-1$ and δ_i is the Kronecker delta ($\delta_0 = 1$ and $\delta_i = 0$ for $i \neq 0$), we can rewrite this set of equations as a linear system and obtain the consistency constraints we are looking for.

Proposition 6.1. *For an analysis window W and a synthesis window S verifying the perfect reconstruction conditions (6.1) for a frame shift R, a set of complex numbers $H \in \mathbb{C}^{M\times N}$ is a consistent spectrogram if and only if, $\forall m \in [\![0, M-1]\!]$, $\forall n' \in [\![0, N-1]\!]$,*

$$\begin{aligned} 0 \;=\; \sum_{n=0}^{N-1} \Big[& \alpha_0^{(R)}(n'-n)H(m,n) \\ & + \sum_{q=1}^{Q-1} e^{j2\pi\frac{qR}{N}n'}\alpha_q^{(R)}(n'-n)H(m-q,n) \\ & + \sum_{q=1}^{Q-1} e^{-j2\pi\frac{qR}{N}n'}\alpha_{-q}^{(R)}(n'-n)H(m+q,n)\Big], \end{aligned} \tag{6.7}$$

or more concisely:

$$\sum_{q=-(Q-1)}^{Q-1} e^{j2\pi\frac{qR}{N}n'}\Big(\alpha_q^{(R)} * H\Big)(m-q,n') = 0, \tag{6.8}$$

where the convolution acts on the second parameter of H and the coefficients $\alpha_q^{(R)}$ are defined by (6.6).

The above proposition summarizes in simple mathematic terms the fact that a consistent STFT spectrogram must be equal to the STFT of its inverse STFT, as the left-hand term in Eq. (6.8) is $\mathcal{F}(H)$.

This idea of deriving consistency constraints which are independent of the signal has been investigated in image processing by Ando for the derivation of consistent gradient operators [11]. Ono and Ando also derived a class of filterbanks such that the time derivative of the amplitude and phase are represented as the real and imaginary parts of a holomorphic function in the time-frequency domain [147, 148]. This can be understood as a consistency constraint in the continuous domain. Similar ideas have also been investigated in the theory of reproducing kernels for continuous wavelet transforms [213].

We note that the coefficients $\alpha_q^{(R)}$ can be simply computed from the image by the operator $\mathcal{F}$ of a set of complex numbers where all bins have value zero except one which has value

one. Indeed, let $U \in \mathbb{C}^{(2Q-1)\times N}$ be a set of complex numbers $U(q,p) = \delta_q \delta_p$, where for convenience of notation the time frame index q is assumed to span $[\![-(Q-1), Q-1]\!]$ and the frequency bin index p is considered modulo N. The elements of U are zero everywhere except $U(0,0) = 1$. Then, we easily see from Eq. (6.7) that

$$\mathcal{F}(U)(q,p) = e^{j2\pi \frac{qR}{N} p} \alpha_q^{(R)}(p), \tag{6.9}$$

from which we can obtain the coefficients $\alpha_q^{(R)}$.

6.3.2 Consistency criterion

Equation (6.8) represents the relation between a set of complex numbers and the STFT of its inverse STFT. Its left member is $\mathcal{F}(H)$, i.e., the difference between H and the STFT of its inverse STFT, which we shall call the residual in the following. The L^2 norm of $\mathcal{F}(H)$ is equal to zero for a consistent STFT spectrogram as stated in (6.8), and can be considered as a criterion on the consistency of a set of complex numbers considered as an STFT spectrogram. This defines a consistency criterion as follows:

$$\mathcal{I}(H) = ||\mathcal{F}(H)||^2 = \sum_{m,n} \Big| \sum_{q=-(Q-1)}^{Q-1} e^{j2\pi \frac{qR}{N} n} \Big(\alpha_q^{(R)} * H \Big)(m-q, n) \Big|^2. \tag{6.10}$$

6.3.3 Consistency as a cost function

We can think of using the consistency criterion (6.10) as a cost function, for example when working on source separation or spectrogram modification algorithms in the complex time-frequency domain. In such situations where one attempts to estimate complex spectrograms following certain properties from an input signal, ensuring that the estimated sets of complex numbers are consistent spectrograms is both likely to lead to better, because more consistent, results, and to ease the optimization process by reducing the dimension of the problem, as it will tend to discard solutions which are not coherent. Two examples of methods which will hopefully benefit from the consistency criterion we introduced are the harmonic percussive source separation (HPSS) framework [149] and the complex NMF framework recently introduced by Kameoka [109]. The HPSS framework attempts to separate the harmonic and percussive parts of a music signal by modeling the STFT spectrogram of a music signal as the sum of two components which respectively minimize continuity cost functions in the time and frequency directions. The model is so far defined in the power domain, but could obviously benefit from a formulation in the complex domain which would avoid the approximate additivity assumption of the powers. Ensuring that the separated harmonic

and percussive parts are as consistent as possible is expected to lead to more meaningful and perceptually satisfactory solutions. The complex NMF framework attempts to represent a complex spectrogram as a combination of typical non-negative spectral templates, their non-negative sparse activations, and complex phase terms. One of the applications of this framework is source separation in the complex time-frequency domain, and again introducing a consistency cost function on the separated complex spectrograms is likely to lead to more meaningful and perceptually better results, and to ease the optimization process.

In Section 6.4, we will investigate the use of the consistency criterion introduced above to define an objective function on phase when the magnitude is fixed.

6.3.4 Approximated consistency criterion

By looking at the actual values of the coefficients $\alpha_q^{(R)}(p)$ involved in the definition of the residual $\mathcal{F}(H)$, we notice that most of the weight is actually concentrated near $(0, 0)$, as can be seen in Fig. 6.2 for a window length $N = 512$ and a frame shift $R = 256$, with a Hanning analysis window and a rectangular synthesis window. One can thus approximate the consistency criterion by using only $(2l + 1) \times (2Q - 1)$ coefficients instead of the total $N \times (2Q - 1)$, where $l \ll N$:

$$\hat{\mathcal{I}}(H) = \sum_{m,n} \Big| \sum_{q=-(Q-1)}^{Q-1} e^{j2\pi \frac{qR}{N} n} \sum_{|p|\leq l} \alpha_q^{(R)}(p) H(m - q, n - p) \Big|^2, \tag{6.11}$$

with the frequency bin index in H considered modulo N.

6.3.5 Optimization of the analysis/synthesis windows

As the criterion (6.11) is an approximate version of the criterion (6.10) based on the observation that the weight in the coefficients $\alpha_q^{(R)}(p)$ is concentrated in small values of p (modulo N), finding analysis and synthesis windows which concentrate as much weight as possible in a given range of coefficients can lead to a better approximation. We investigated this idea and performed an optimization of the analysis/synthesis windows for a 50 % overlap and for $l = 2$, to maximize the L^2 norm of the 5×3 coefficients considered, assuming the analysis and synthesis windows were equal and symmetric. Quite remarkably, the window we obtained was very similar to the square root of the Hanning window, as can be shown in Fig. 6.3 for a window length of 512 samples. We thus used the square root of the Hanning window, also known as the sine window, in the experiments we conducted. The central coefficients for this window are shown in Fig. 6.4, and the central 5×3 values $\alpha_q^{(R)}(p)$ with

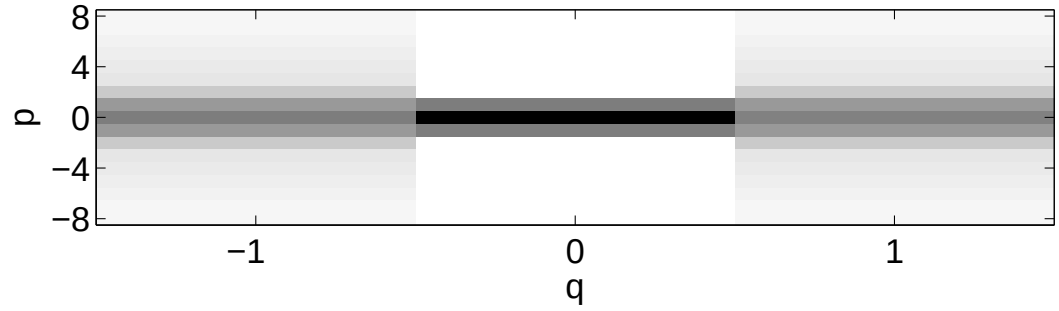


Figure 6.2: *Magnitude of the central coefficients* $\alpha_q^{(R)}(p)$ *for* $N = 512$, $R = 256$, *a Hanning analysis window and a rectangular synthesis window.*

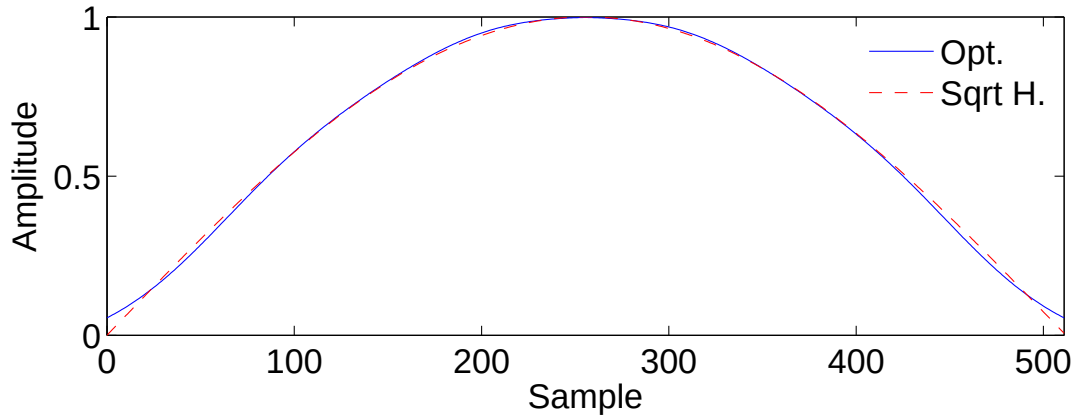


Figure 6.3: *Comparison of the optimized window and the square root Hanning window for* $N = 512$, $R = 256$ *and* $l = 2$.

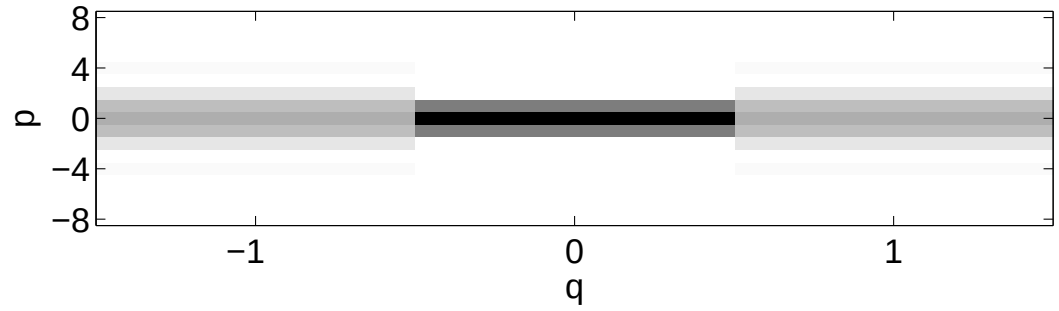


Figure 6.4: *Magnitude of the central coefficients* $\alpha_q^{(R)}(p)$ *for* $N = 512$, $R = 256$, *and a square root Hanning analysis and synthesis window.*

$p \in [\![-2; 2]\!]$ and $q \in [\![-1; 1]\!]$ are given by

$$\begin{array}{|c|c|c|} \hline -0.0530536 & 0.0000000 & -0.0530536 \\ \hline 0.1250000j & -0.2500000 & -0.1250000j \\ \hline 0.1591529 & 0.5000000 & 0.1591529 \\ \hline -0.1250000j & -0.2500000 & 0.1250000j \\ \hline -0.0530536 & 0.0000000 & -0.0530536 \\ \hline \end{array} . \tag{6.12}$$

The particular symmetries of the coefficients $\alpha^{(R)}$ will be studied in Section 6.4.5.

6.4 Phase reconstruction for a modified STFT spectrogram

We consider here the application of the consistency criterion defined in Section 6.3.2 to develop an algorithm for reconstructing the most coherent phase given a modified magnitude STFT spectrogram.

The iterative STFT algorithm [89] introduced by Griffin and Lim is the reference for such algorithms. Its principle is to find the consistent STFT spectrogram with magnitude closest to a given modified magnitude spectrogram. Here, we propose a flexible phase reconstruction algorithm based on an objective function derived from the consistency criterion. Contrary to the iterative STFT algorithm which looks for a signal whose magnitude STFT spectrogram is closest to the given magnitude spectrogram, the algorithm we propose looks for a phase such that the spectrogram obtained by associating it with that magnitude spectrogram is as consistent as possible, or in other words changes as less as possible when going to the time domain and back to the time-frequency domain. It can still be considered conceptually close to the iterative algorithm in that the inverse STFT of the combination of the estimated phase and the modified magnitude spectrogram is the kind of signal the iterative algorithm tries to estimate, but a crucial difference is that it operates directly in the time-frequency domain and thus does not require to go at each iteration from the time-frequency domain to the time domain and vice versa through FFTs on the whole frequency band. This enables us to select at each iteration independently which time-frequency bin's phase to update, giving an extra flexibility to our algorithm which should enable it to be included in a wide range of signal processing algorithms, and also allowing, together with a focus on local phase coherence conditions, to reduce the computational cost.

6.4.1 Objective function for phase reconstruction problems

In the problem of phase reconstruction, we are given a set of real non-negative numbers $A_{m,n}$ which are supposedly the amplitude part of an STFT spectrogram, for example obtained through modifications of the power spectrum of a sound. The goal is to estimate the phase $P_{m,n}$ to adjoin to A such that $A_{m,n}e^{jP_{m,n}}$ is as close as possible to be a consistent STFT spectrogram.

Based on the derivation of Section 6.3.2, this amounts to minimizing the consistency

criterion $\mathcal{I}$ w.r.t. the phase P, with the amplitude A given, defining the following objective function:

$$\tilde{\mathcal{I}}(P) = \sum_{m,n} \Big| \sum_{q=-(Q-1)}^{Q-1} e^{j2\pi \frac{qR}{N} n} \alpha_q^{(R)} * A_{m-q,n} e^{jP_{m-q,n}} \Big|^2, \tag{6.13}$$

If an estimation of the phase, for example the phase of the mixture when dealing with source separation, is available, it can be used as initial setting for P.

In [89], Griffin and Lim presented the iterative STFT algorithm, which consists in iteratively updating the phase $P_{m,n}^{(k)}$ at step k by replacing it with the phase of the STFT of its inverse STFT while keeping the magnitude A. The algorithm is illustrated in Fig. 6.5, where $x^{(k+1)}$ denotes the inverse STFT of $A_{m,n} e^{jP_{m,n}^{(k)}}$, $\hat{x}^{(k+1)}$ the STFT of $x^{(k+1)}$, and $P_{m,n}^{(k+1)}$ the phase of $\hat{x}^{(k+1)}$, $P_{m,n}^{(k+1)} = \angle \hat{x}^{(k+1)}$.

They showed that this procedure estimates a real signal x which minimizes (at least locally) the distance

$$d(x, A) = \sum_{m,n} \Big| |\hat{x}|_{m,n} - A_{m,n} \Big|^2, \tag{6.14}$$

i.e., the squared error between the magnitude of the STFT $\hat{x}$ of x, and the magnitude spectrogram A. As can be seen in Fig. 6.5, we shall note that the objective function $\tilde{\mathcal{I}}$ measures a slightly different quantity from the distance (6.14), but that the iterative STFT algorithm also converges to a minimum of (6.13). Indeed, both distances become equivalent near the convergence, as one can show that $d(x^{(k+1)}, A) \leq \tilde{\mathcal{I}}(P^{(k)}) \leq d(x^{(k)}, A)$ [89]. However, the objective function $\tilde{\mathcal{I}}$ we introduced has the advantages to be explicit and defined directly in the time-frequency domain, and in its general version (6.10) not to be limited to phase reconstruction problems with fixed magnitude. Its explicitness will for example allow us to take advantage of sparseness, or to optimize the phase only where it needs to be estimated when some regions can be relied on. The objective function (6.13) here enables us to derive a simplified algorithm for phase reconstruction, as we shall now explain.

6.4.2 Direct optimization of $\tilde{\mathcal{I}}$

The iterative STFT algorithm, as mentioned above, can be used to minimize $\tilde{\mathcal{I}}$. However, this can be considered as an indirect minimization, and it is worth looking at a direct minimization of $\tilde{\mathcal{I}}$ through classical optimization methods. This will indeed provide us with the freedom to modify/approximate the objective function on one hand, and to select how each bin will be dealt with on the other. For example, if only some parts of the spectrogram must have their phase reconstructed, e.g., after their magnitude has been reconstructed

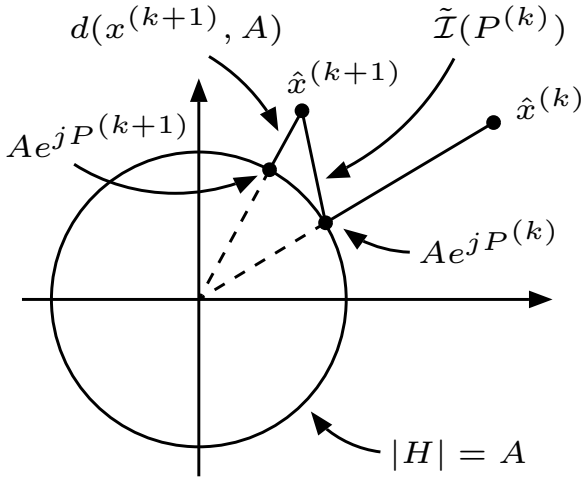


Figure 6.5: *Illustration of the iterative STFT algorithm and the relation between the objective function $\tilde{\mathcal{I}}$ and the distance $d(x, A)$.*

by some other means, iterative STFT does not allow to keep the other parts unchanged and reconstruct the phase only where it is necessary while taking into account boundary conditions between the regions. This can be simply performed with the framework we develop here by updating only the bins whose phase is considered not reliable. This remark is actually not limited to the optimization of the phase: minimizing the consistency criterion (6.10) with respect to both the magnitude and phase of some bins while keeping others fixed could be used for example as a way to reconstruct bins which have been discarded by a binary mask from the bins which have been determined as reliable. Finally, one could also imagine introducing weights depending on frequency to emphasize stronger consistency in certain frequency regions based on perceptual criteria, or on magnitude to give more importance to the reduction of inconsistency around bins with large magnitude. The bin selection presented in Section 6.4.4 is a discrete version of this last idea.

We note that an interesting probabilistic approach to phase reconstruction which also relies on the direct optimization of an objective function has been proposed by Achan et al. [7]. The objective function is defined in the time domain and allows to account for prior distributions on the waveform. This approach is however more dedicated to speech signals.

6.4.3 Approximate objective function and phase coherence

Here, we will make the following two approximations. We will first neglect the influence of $P_{m,n}$ in all the terms $\mathcal{F}(H)(m', n')$ other than $\mathcal{F}(H)(m, n)$, which is the one where it is multiplied by $\alpha_0^{(R)}(0)$. The motivation behind this first approximation is that the coefficient

$\alpha_0^{(R)}(0)$ dominates over the other coefficients. By assuming the other phase terms fixed, we will then update each bin's phase $P_{m,n}$ so that $\alpha_0^{(R)}(0)A_{m,n}e^{jP_{m,n}}$ is in opposite direction with the terms coming from the neighboring bins, while keeping its amplitude $A_{m,n}$ fixed. This corresponds to performing a coordinate descent method [214]. More precisely, the update for bin (m, n') is

$$P_{m,n'} \leftarrow -s\angle\Big(\sum_{(p,q)\neq(0,0)} e^{j2\pi\frac{qR}{N}n'}\alpha_q^{(R)}(p)H(m-q,n'-p)\Big), \tag{6.15}$$

where $s = 1$ if $\alpha_0^{(R)}(0) > 0$ and $s = -1$ if $\alpha_0^{(R)}(0) < 0$.

Second, following the approximation of the consistency criterion explained in Section 6.3.4 motivated by the concentration of the coefficients $\alpha_q^{(R)}(p)$ near $(0, 0)$ (with p considered modulo N), we further approximate the update equations (6.15) by using only $(2l + 1) \times (2Q-1)$ central coefficients. This approximation is motivated here as well by the importance of local phase coherences, in particular the so-called "horizontal" and "vertical" coherences, to obtain a perceptually good reconstructed signal, and can be considered close to phase locking techniques [110, 116, 157]. Horizontal coherence refers to phase consistency within each frequency channel, i.e., to the fact that in frequency band n, phase roughly evolves at a speed corresponding to n, and vertical coherence refers to phase consistency across channels, in particular to the fact that in a time frame m, the phases at bins n and $n + 1$ are roughly equal.

This approximation enables us to compute directly the update of each bin through the summation of a few terms, instead of the whole convolution which would be involved if using all the terms. The update becomes:

$$P_{m,n'} \leftarrow -s\angle\Big(\sum_{\substack{(n,q)\neq(0,0)\\|n|\leq l}} e^{j2\pi\frac{qR}{N}n'}\alpha_q^{(R)}(n)H(m-q,n'-n)\Big), \tag{6.16}$$

where frequency indices are understood modulo N. For $l = 2$ and a 50 % overlap, for example, we only consider 5×3 coefficients.

6.4.4 Taking advantage of sparseness

As evoked above, using a direct optimization of the objective function $\tilde{\mathcal{I}}$ enables us to select which bins to update. This can be the key to deal with problems where only a part of the spectrogram has to have its phase reconstructed, but it can also in general be used to lower the computational cost. Indeed, we can use the sparseness of the acoustic signal to limit

the updates to bins with a significant amplitude, or progressively decrease the amplitude threshold above which the bins are updated, starting with the most significant bins and refining afterwards. This idea can be related to the peak picking techniques in [110, 116].

6.4.5 Further simplifications

The number of operations involved in the computation of the updates (6.16) can be further reduced by noticing symmetries in the coefficients $\alpha_q^{(R)}$. First, without any assumption on the analysis and synthesis windows, it is obvious from (6.6) that

$$\alpha_q^{(R)}(-p) = \overline{\alpha_q^{(R)}(p)}. \tag{6.17}$$

When the analysis and synthesis windows are symmetric and such that $W(0) = 0$, the coefficients have still more symmetries. Indeed, we notice that, from (6.6),

$$\begin{aligned}
\alpha_q^{(R)}(p) &= \frac{1}{N}\sum_{k=0}^{N-1} W(k)S(k+qR)e^{-j2\pi p\frac{k+qR}{N}} \\
&= \frac{1}{N}\sum_{k=0}^{N-1} W(N-k)S(N-(k+qR))e^{-j2\pi p\frac{k+qR}{N}} \\
&= \frac{1}{N}\sum_{k'=1}^{N} W(k')S(k'-qR)e^{j2\pi p\frac{k'-qR}{N}} \\
&= \overline{\alpha_{-q}^{(R)}(p)},
\end{aligned} \tag{6.18}$$

as the difference between the last two lines is $\frac{1}{N}(W(N)S(N-qR) - W(0)S(-qR))e^{-j2\pi p\frac{qR}{N}}$, which is zero under the above assumptions.

Based on these symmetries and on the fact that, for complex numbers a, b and c, the computation of the quantity $ab + \overline{a}c$ can be performed using only 4 real multiplications instead of the 8 real multiplications required for the general sum of two products of two complex numbers, we can reduce the number of multiplications involved in the computation.

6.4.6 Time-scale modification

Need for an efficient frequency-domain algorithm

Many methods for time-scale and pitch-scale modification of acoustic signals have been proposed, and the interest on this subject intensified in recent years with the increase in the commercial application of such techniques. So far, most commercial implementations rely on time-domain methods, usually variations on Synchronous Overlap and Add (SOLA)

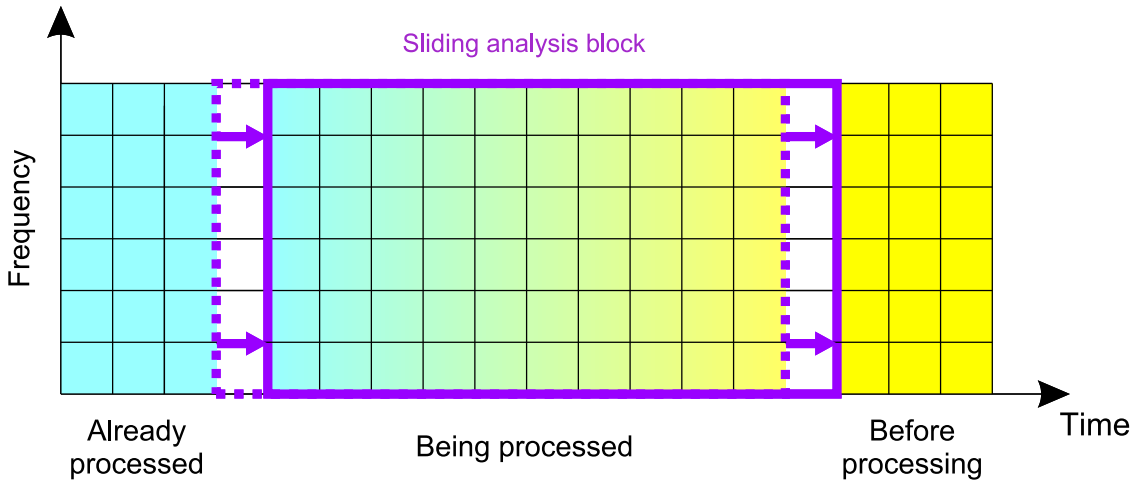


Figure 6.6: *Illustration of the sliding-block analysis principle.*

or Pitch Synchronous Overlap and Add (PSOLA) techniques [143]. Their advantages are a low computational cost and good quality results for small modification factors (smaller than ±20 % or ±30 %) and monophonic sounds. For larger factors, polyphonic sounds or non-pitched signals, however, the quality of the results drops severely. On the other hand, frequency-domain methods, such as the phase vocoder [64], are not limited to such constraints, but they involve a much higher computational cost and introduce artifacts of their own [116]. These artifacts have been shown to be mainly connected to phase incoherences, and special care must thus be taken when estimating the phases in the modified signal's STFT spectrogram. The iterative STFT algorithm of Griffin and Lim has been proposed as a way to correct such phase incoherences, although the computational cost and the slow speed of convergence have been obstacles to its adoption in commercial applications. The algorithm we introduced is a flexible alternative to iterative STFT, and by an active use of sparseness and the reduction of the number of multiplications involved at each step, should lead to a lower computational cost.

Sliding-block analysis for real-time processing

Inspired by an idea in [149], we derive a real-time optimization scheme for the objective function introduced above based on a sliding-block analysis. As illustrated in Fig. 6.6, the spectrogram is not processed all at once, but progressively from left to right, making it possible to change the parameters while sound is being played. In the particular case of time-scale modification, this leads to the following procedure. The waveform to be time-scaled is read N samples at a time, where N is the window length. The STFT transform of this incoming frame is computed and adjoined to the frames of STFT spectrogram already

computed, at the extreme right. If the block size is set to b and the frame shift to R, at a given time, we keep $b + 2Q$ frames, where $Q = N/R$ is the number of overlapping frames: the b central frames are updated using the algorithm derived above, through update equations (6.16), while the Q already processed frames on the left and the Q yet to be processed frames on the right are kept fixed and only used in the computations of the updates of the b central frames. Once the update has been performed, the frames are shifted to the left, and the frame which just exited the central block is inverse-DFTed and overlap-added, after windowing by the synthesis window, to the already computed part of the time-scaled waveform. The determination of the start of the next N sample part of incoming signal to be read is made in accordance with the time scale modification factor f such that the average shift for the incoming signal is fR, while keeping an integer shift at each step. The procedure is then iterated. Altogether the input waveform is read with window shifts of fR samples on average and synthesized with window shifts of R samples, resulting in a resynthesized signal whose length is $1/f$ times the length of the original signal. The number of iterations performed on each frame is equal to the block size.

This way of building intermediate frames of the spectrogram is arguably much more reliable than interpolating neighboring spectrogram frames. An initial estimation of the phase is also obtained in this way, which can be used as a starting point for the optimization by both the iterative STFT algorithm and the method we propose. Other initialization schemes for the phase have been proposed, for example by Slaney et al. [182] and Zhu et al. [215]. They apply not only to time-scale modification but more generally to phase reconstruction based on magnitude spectrograms, and have been proven by their respective authors to lead to better results than simply using zero phase as initial phase. Incorporating them in our framework could lead to better results as well.

Experimental evaluation

We implemented the proposed method and the iterative STFT algorithm and compared their convergence speed on the time-scale modification of the first 23 s of Chopin's Nocturne No. 2. The time-scale modification factor was set to 0.7, the frame length to 1024 and the frame shift to 512, for a final length of approximately 32 s. We used a 5×3 approximation of the coefficients $\alpha_q^{(R)}(p)$ for our algorithm. We ran our algorithm in two different experimental conditions, one where all the bins are updated at each iteration, and the other relying on sparseness, where at iteration k, only bins whose amplitude is larger than Ae^{-Bk} are updated. Here, we used $A = 1$ and $B = 0.005$, which were determined experimentally. The objective

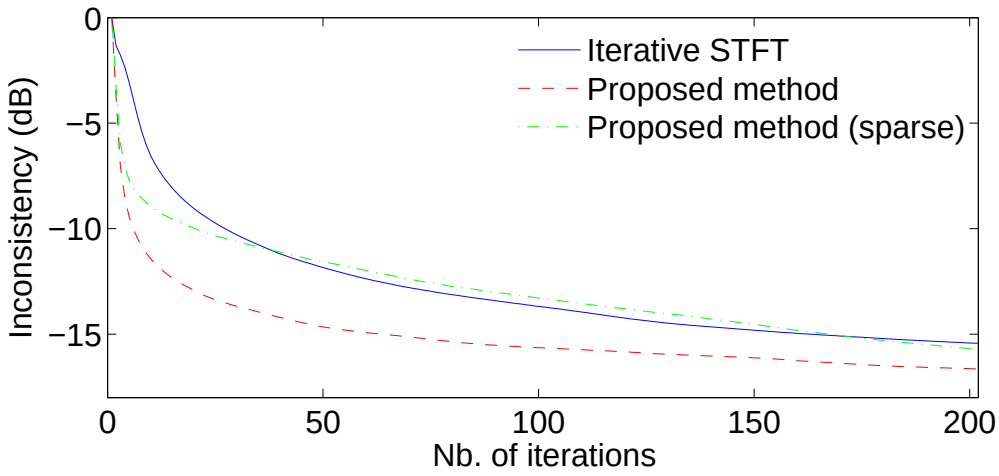


Figure 6.7: *Comparison of the evolution of the inconsistency measure $\mathcal{I}(H)$ w.r.t. the number of iterations for the iterative STFT algorithm and the proposed method.*

function $\mathcal{I}(H)$ is used as a measure of convergence, and represented in decibels, with the initial value as a reference.

A comparison of the speed of convergence in terms of the decrease of $\mathcal{I}(H)$ w.r.t. the number of iterations (or, equivalently, the block size) is shown in Fig. 6.7. One can see that, although our algorithm is based on an approximation of the original objective function $\mathcal{I}(H)$, it outperforms the iterative STFT algorithm in terms of speed of convergence as the number of iterations required to reach a given decrease of inconsistency. The sparse version of our method has a slower convergence speed, close to the iterative STFT algorithm. This could be expected as only a part of the bins are updated. We also compared the computation times of the three methods, measuring only the time required by the phase reconstruction part of the algorithm as the other parts are identical. With our implementations, for 200 iterations, the iterative STFT algorithm took 29.1 s, our method with full updates 24.2 s, and our method using sparseness 2.1 s. Looking at the amount of time required to reach a certain decrease in inconsistency illustrates how our method, by combining speed of convergence and lower computational cost, leads to a much shorter computational time than iterative STFT. Table 6.1 shows the computation times to reach a given decrease of inconsistency in dB, for the time-scale modification of the first 23 s of Chopin's Nocturne no.2 with a factor of 0.7, in the same conditions as described above.

In terms of flexibility, speed of convergence and computation time, our algorithm thus outperforms the iterative STFT algorithm.

Table 6.1: *Computation time (s) required to reach a certain decrease of the inconsistency measure $\mathcal{I}$*

Time to reach	-10dB	-13dB	-15dB
Iterative STFT (G&L)	3.9s	10.9s	23.8s
Proposed method	0.6s	2.4s	7.6s
Proposed method (sparse)	0.2s	0.8s	1.6s

6.5 Audio encryption based on inconsistent STFT spectrograms

We can design a family of encryption codes based on any perfect reconstruction analysis/synthesis window couple by using a jammer in the STFT space. The key idea here is that, starting from any set of complex numbers $H \in \mathbb{C}^{M\times N}$, one can build an inconsistent STFT spectrogram with non-zero "energy" such that its inverse STFT is identically zero. In other words, for a given frame shift R and any perfect reconstruction analysis/synthesis window couple, there exists a family of sets of complex numbers in $\mathbb{C}^{M\times N}$ whose inverse STFT is identically silence. Indeed, starting from $H \in \mathbb{C}^{M\times N}$, let x_H be its inverse STFT signal. The STFT X_H of x_H is a consistent spectrogram, whose inverse STFT is also x_H. Thus the inverse STFT of $\mathcal{F}(H) = X_H - H$ is identically 0, although in general the L^2 norm of $X_H - H$ is not. In algebraic terms, this can be simply summarized as

$$\mathrm{Im}(\mathrm{STFT}_W \circ \mathrm{iSTFT}_S - \mathrm{Id}) \subset \mathrm{Ker}(\mathrm{iSTFT}_S), \tag{6.19}$$

the subspace on the left being in general not reduced to $\{0\}$. The important point to note here is that only the inverse STFT with the synthesis window S which was used in building X_H will lead to silence, the inverse STFT with any other window leading in general to a non-zero signal.

We can apply this procedure to a set of random complex numbers to obtain a very "noisy" inconsistent spectrogram which, with the correct synthesis window, leads to silence when inverse STFT-ed. Now, if this inconsistent noisy spectrogram, multiplied by a large coefficient, is added to the coherent spectrogram (built using the same window couple) of a speech or music sound for example, we obtain a set of complex numbers in which the power coming from the speech or music sound is masked and hardly detectable. An example of the magnitude of such a spectrogram is shown in Fig. 6.9, with the square root Hanning window as analysis/synthesis window, a window length $N = 512$ and window shift $R = 256$. The

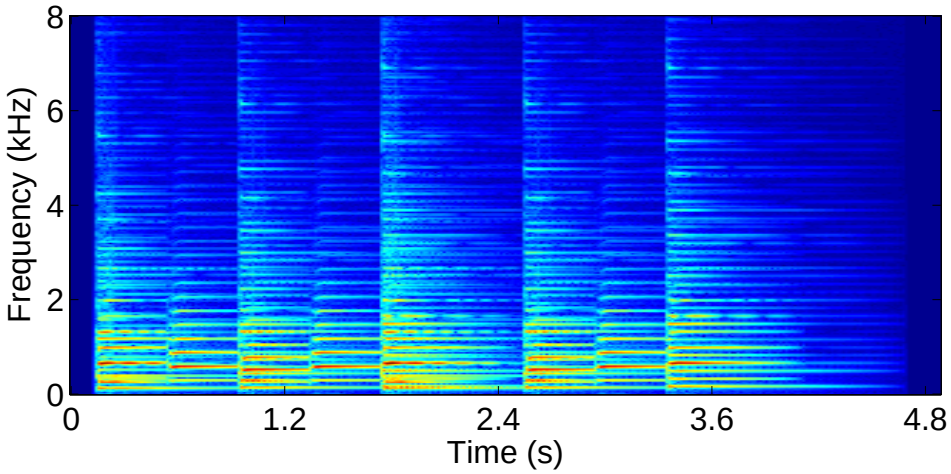


Figure 6.8: *Magnitude of the original spectrogram.*

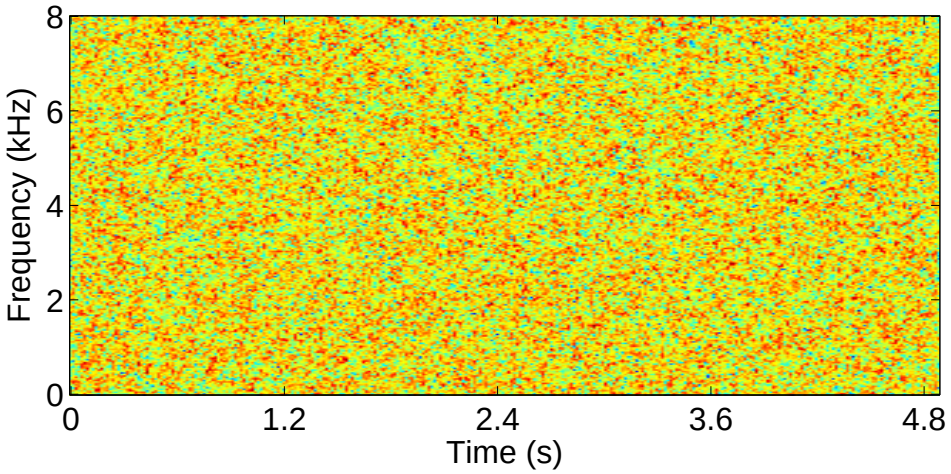


Figure 6.9: *Magnitude of the inconsistent spectrogram.*

random set of complex numbers was generated by randomly modifying the phase of the spectrogram of a Gaussian white noise signal with standard deviation 1. Multiplied by a coefficient 100, it was added to the spectrogram of a computer generated music piece consisting of a mixture of piano and trumpet with a 16kHz sampling rate [176], shown in Fig. 6.8.

If the inverse STFT is performed with a different window than the one used to build the jammer spectrogram, the obtained signal is more or less noise. This can be seen in Fig. 6.10 where the inverse STFT of the spectrogram in Fig. 6.9 is computed using the Hanning window, leading to a Signal to Noise Ratio (SNR) of about -30 dB, while the inverse STFT using the same Hanning window of the original spectrogram without addition of the jammer still leads to an SNR of about $+18$ dB. However, if the correct synthesis window is used, the jammer part of the spectrogram cancels off and the original speech or music sound is

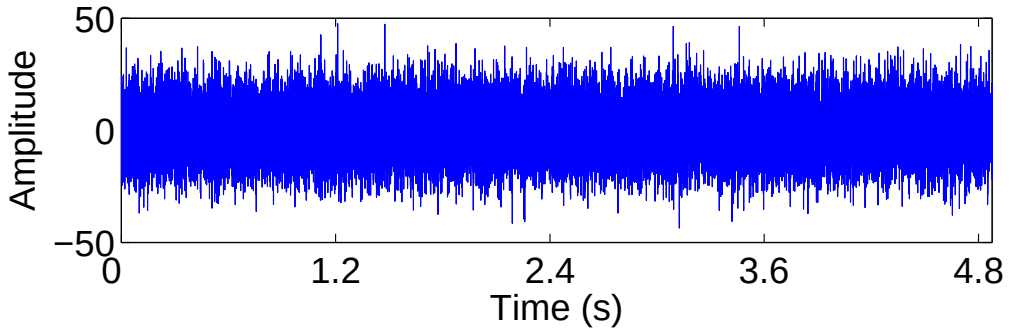


Figure 6.10: *Waveform of the inverse STFT of the inconsistent spectrogram using a Hanning window.*

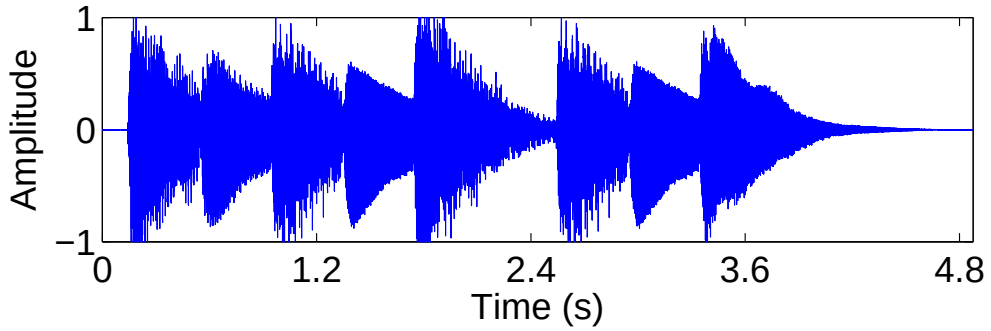


Figure 6.11: *Waveform of the inverse STFT of the inconsistent spectrogram using the correct square root Hanning window.*

perfectly recovered (up to quantization error), as shown in Fig. 6.11. The synthesis window function thus acts as a key to decrypt the spectrogram and retrieve the hidden message.

Another way to produce interesting results, whose potential should be further investigated, is to start with an audio signal (white noise was used above, but speech or music can also be used), randomly change the phase of its STFT, and use the obtained set of complex numbers as a root for the procedure. The "hidden" sound can be heard for the correct window, while for other windows a distorted version of the root will be heard.

Although such issues as dynamic range limitation or codebreaking by a window optimization based on the minimization of the output power should be considered, the potential applications of this technique as an encryption system is worth investigating.

6.6 Summary of Chapter 6

In this chapter, we introduced a general framework to assess for the consistency of complex STFT spectrograms, and explained how it could be used to introduce cost functions in the complex time-frequency domain for a wide range of algorithms and to define an objective

function on phase when working in the time-frequency magnitude or power domain. We developed a flexible phase reconstruction algorithm which can take advantage of the sparseness of the input signal and can take into account regions where phase is reliable. We applied this algorithm for time scale modification, and explained how to perform a real-time processing based on a sliding-block analysis. Finally, we showed how inconsistent STFT spectrograms could be used to hide sounds in the spectrogram domain.

Chapter 7

Adaptive template matching in the time domain

7.1 Introduction

In Chapter 6, we exposed the inherent limitations of working in the magnitude or power domain and the issues that they raise, namely non-additivity in mixtures, necessity to recover the phase information for resynthesis and impossibility to exploit phasre-related cues, and argued that these issues needed to be either avoided or solved. There, we presented a framework to deal with these issues based on the structure of complex STFT spectrograms, introducing a consistency criterion which can be used both when working in the complex time-frequency domain to ensure that what is done is consistent with the underlying structure of the space, or when working in the magnitude or power domain to reconstruct the phase information. In this chapter, we investigate another way to deal with these issues by working in the time domain. Additivity of waveforms holds true, and there is no resynthesis problem. Moreover, the model we present exploits the recurrence of elementary waveforms inside a signal, and is thus able to take advantage of cues directly related to the phase information.

Another motivation for the development of the model introduced in this chapter is that of unsupervised learning based on the data. We have so far put emphasis on how to design algorithms relying on statistical models to analyze acoustical scenes, but we have not yet addressed the problem of the acquisition of such models. The model of speech signals we introduced in Chapter 3 has been tailored, based mainly on prior knowledge on the particularities of speech and on the human auditory organization process. It is interesting, for two reasons, to wonder how one can extract regularities directly from the waveform.

First, the human brain needed to build up the models it is using from distorted, mixed, and incomplete stimuli, in an unsupervised way, and modeling this process could both lead to new insights on the way the brain may work and to new theoretical results in machine learning theory. Second, although using tailored models and constraints enables one to use prior knowledge, the results we can expect are limited to the quality and the appropriateness of that prior knowledge. It may be more rewarding to try to learn the most appropriate model directly from the data. Moreover, trying to learn the models which are best fitted to a certain type of signals may in turn give us information on the structure of that signal. The model we introduce in this chapter performs this kind of data-driven learning by extracting, in an unsupervised way, relevant constituents recurring in a waveform.

In Section 7.2, we explain the motivations for the choices we made in the design of the model. After reviewing in Section 7.3 the original semi-NMF model, we introduce in Section 7.4 our framework, called shift-invariant semi-NMF, and derive update equations to optimize its parameters, which are waveform templates and their amplitudes. In Section 7.5, we give a proof of convergence for the update equations of the amplitudes which takes into account a sparseness prior. In Section 7.6, we explain how the templates can also be optimized through gradient descent, explicating on the way the general relation between natural gradient descent with unit L^2-norm constraint and gradient descent on the objective function were the unit-norm constraint is explicitly included through normalization. We then evaluate the performance of our algorithm in Section 7.7, both on synthetic and real data. We finally discuss some issues in Section 7.8.

7.2 Motivations for the design of the model

It is often the case that an observed waveform is the superposition of elementary waveforms, taken from a limited set and added with variable latencies and variable but positive amplitudes. Examples are a music waveform, made up of the superposition of stereotyped instrumental notes, or extracellular recordings of nerve activity, made up of the superposition of spikes from multiple neurons. In these examples, the elementary waveforms include both positive and negative excursions, but they usually contribute with a positive weight. Additionally, the elementary events are often temporally compact and their occurrence temporally sparse. Conventional template matching uses a known template and correlates it with the signal; events are assumed to occur at times where the correlation is high. Multiple template matching raises combinatorial issues that are addressed by Matching Pursuit [133]. However

these techniques assume a preexisting dictionary of templates. We wondered whether one can estimate the templates directly from the data, together with their timing and amplitude.

Over the last decade a number of blind decomposition methods have been developed that address a similar problem: given data, can one find the amplitudes and profiles of constituent signals that explain the data in some optimal way. This includes sparse coding (see for example [56, 57, 155] for applications to audio signals), independent component analysis (ICA), non-negative matrix factorization (NMF), and a variety of other blind source separation algorithms. The different algorithms all assume a linear superposition of the templates, but vary in their specific assumptions about the statistics of the templates and the mixing process. These assumptions are necessary to obtain useful results because the problem is under-constrained.

Sparse coding and ICA do not fit our needs because they do not implement the constraint that components (templates) are added with positive weights. NMF constrains weights to be non-negative but requires templates to also be non-negative. We will use instead the semi-NMF algorithm of Ding et al. [63, 122] that allows factoring a matrix into a product of a non-negative and an arbitrary matrix. To accommodate time shifts we modify it following the ideas of Mørup et al. [142] who presented a shift-invariant version of the NMF algorithm, that also includes sparsity constraints. We begin with the conventional formulation of the NMF modeling task as a matrix factorization problem and then derive in the subsequent section the case of a 1D sequence of data. NMF models a data matrix X as a factorization,

$$\hat{X} = AB\,, \tag{7.1}$$

with $A \geq 0$ and $B \geq 0$ and finds these coefficients such that the square modeling error $||X - \hat{X}||^2$ is minimized. Matrix A can be thought of as component amplitudes and the rows of matrix B are the component templates. Semi-NMF drops the non-negative constraint for B, while shift-NMF allows the component templates to be shifted in time. In the NMF algorithm, there is an update equation for A and an update equation for B. Semi-NMF and shift-NMF each modifies one of these equations, fortunately not the same, so their updates can be interleaved without interference.

7.3 Review of semi-NMF

Assume we are given N observations or segments of data with T samples arranged as a matrix X_{nt}. (The segments can also represent different epochs, trials, or even different

channels.) The goal is to model this data as a linear superposition of K component templates B_{kt} with amplitudes A_{nk}, i.e.,

$$\hat{X}_{nt} = \sum_k A_{nk} B_{kt} = A_n^k B_{kt} \,. \tag{7.2}$$

The second expression here uses Einstein notation: indices that appear both as superscript and subscript within a product are to be summed. In contrast to matrix notation, all dimensions of an expression are apparent, including those that are absorbed by a sum, and the notation readily extends to more than two dimensions, which we will need when we introduce delays. We use this notation throughout the chapter and include explicit sum signs only to avoid possible confusion.

Now, to minimize the modeling error

$$E = ||X - \hat{X}||_2^2 = \sum_{nt} \left(X_{nt} - A_n^k B_{kt}\right)^2 \,, \tag{7.3}$$

the semi-NMF algorithm iterates between finding the optimum B for a given A, which is trivially given by the classic least-squares solution,

$$B_{kt} = \left(A_k^n A_{k'n}\right)^{-1} A_n^{k'} X_t^n \,, \tag{7.4}$$

and improving the estimate of A for a given B with the multiplicative update

$$A_{nk} \leftarrow A_{nk} \sqrt{\frac{\left(X_n^t B_{kt}\right)^+ + A_n^{k'} \left(B_{k'}^t B_{kt}\right)^-}{\left(X_n^t B_{kt}\right)^- + A_n^{k'} \left(B_{k'}^t B_{kt}\right)^+}} \,. \tag{7.5}$$

In these expressions, k' is a summation index; $(M)^{-1}$ stands for matrix inverse of M; and, $(M)^+ = \frac{1}{2}(|M| + M)$ and $(M)^- = \frac{1}{2}(|M| - M)$ are to be applied on each element of matrix M. The multiplicative update (7.5) ensures that A remains non-negative in each step; while, baring constraints for B, the optimum solution for B for a given A is found in a single step with (7.4).

7.4 Shift-invariant semi-NMF

7.4.1 Formulation of the model for a 1D sequence

Consider now the case where the data is given as a 1-dimensional time sequence X_t. In the course of time, various events of unknown identity and variable amplitude appear in

this signal. We describe an event of type k with a template B_{kl} of length L. Time index l represents now a time lag measured from the onset of the template. An event can occur at any point in time, say at time sample n, and it may have a variable amplitude. In addition, we do not know *a priori* what the event type is and so we assign to each time sample n and each event type k an amplitude $A_{nk} \geq 0$. The goal is to find the templates B and amplitudes A that explain the data. In this formulation of the model, the timing of an event is given by a non-zero sample in the amplitude matrix A. Ideally, each event is identified uniquely and is well localized in time. This means that for a given n the estimated amplitudes are positive for only one k, and neighboring samples in time have zero amplitudes. This new model can be written as

$$\hat{X}_t = \sum_n A_n^k B_{k,t-n} \tag{7.6}$$

$$= \sum_n \sum_l A_n^k \delta_{n,t-l} B_{kl} \,. \tag{7.7}$$

The Kronecker delta δ_{tl} was used to induce the desired shifts n. We can dispense with the cumbersome shift in the index if we introduce

$$\tilde{A}_{tkl} = \sum_n A_{nk} \delta_{n,t-l} \,. \tag{7.8}$$

The tensor $\tilde{A}_{tkl}$ represents a block Toeplitz matrix, with K blocks of dimension $T \times L$. Each block implements a convolution of the k-th template B_{kl} with amplitudes signal A_{nk}. With this definition the model is written now simply as:

$$\hat{X}_t = \tilde{A}_t^{kl} B_{kl} \,, \tag{7.9}$$

with $A_{nk} \geq 0$. We will also require a unit-norm constraint on the K templates in B, namely, $B_k^l B_{kl} = 1$, to disambiguate the arbitrary scale in the product of A and B.

7.4.2 Optimization criterion with sparseness prior

Under the assumption that the data represent a small set of well-localized events, matrix A should consist of a sparse series of pulses, the other samples having zero amplitude. To favor solutions having this property, we use a generalized Gaussian distribution as prior probability for the amplitudes. Assuming Gaussian white noise, the new cost function given

by the negative log-posterior reads (up to a scaling factor),

$$E = \frac{1}{2}||X - \hat{X}||_2^2 + \beta\,||A||_\alpha^\alpha \tag{7.10}$$

$$= \frac{1}{2}\sum_t \left(X_t - \tilde{A}_t^{kl} B_{kl}\right)^2 + \beta \sum_{kl} A_{kl}^\alpha\,, \tag{7.11}$$

where $||\cdot||_p$ denotes the L^p norm (or quasi-norm for $0 < p < 1$). The shape parameter α of the generalized Gaussian distribution controls the odds of observing low versus high amplitude values and should be chosen based on the expected rate of events. For our data we mostly choose $\alpha = 1/4$. The parameter β is a normalization constant which depends on the power of the noise, σ_N^2, and the power of the amplitudes, σ_A^2, with $\beta = \frac{\sigma_N^2}{\sigma_A^\alpha}\big(\Gamma(3/\alpha)/\Gamma(1/\alpha)\big)^{\alpha/2}$.

7.4.3 A update

The update for A which minimizes this cost function is similar to update (7.5) with some modifications. In (7.5), amplitudes A can be treated as a matrix of dimensions $T \times K$ and each update can be applied separately for every n. Here the problem is no longer separable in n and we need to treat A as a $1 \times TK$ matrix. B is now a $TK \times T$ matrix of shifted templates defined as $\tilde{B}_{nkt} = B_{k,t-n}$. The new update equation is similar to (7.5), but differs in the term BB^T:

$$A_{nk} \leftarrow A_{nk} \sqrt{\frac{\left(\tilde{X}_n^l B_{kl}\right)^+ + A^{n'k'}\left(\tilde{B}_{n'k'}^t \tilde{B}_{nkt}\right)^-}{\left(\tilde{X}_n^l B_{kl}\right)^- + A^{n'k'}\left(\tilde{B}_{n'k'}^t \tilde{B}_{nkt}\right)^+ + \alpha\beta A_{nk}^{\alpha-1}}}\,. \tag{7.12}$$

The summation in the BB^T term is over t, and is 0 most of the time when the events do not overlap. We also defined $\tilde{X}_n^l = X_{n+l}$, and the time index in the summation $\tilde{X}_n^l B_{kl}$ extends only over lags l from 0 to $L-1$. To limit the memory cost of this operation, we implemented it by computing only the non-zero parts of the $TK \times TK$ matrix BB^T as $2L-1$ blocks of size $K \times K$. The extra term in the denominator of (7.12) is the gradient of the sparseness term in (7.11). We give a proof of convergence for this equation in Section 7.5.

7.4.4 B update

The templates B that minimize the square modeling error, i.e., the first term of the cost function (7.11), are given by a least-squares solution which now writes:

$$B_{kl} = \left(\tilde{A}_{kl}^t \tilde{A}_{tk'l'}\right)^{-1} \tilde{A}_t^{k'l'} X^t\,. \tag{7.13}$$

The matrix inverse is now over a matrix of LK by LK elements. Note that the sparseness prior will act to reduce the magnitude of A. Any scaling of A can be compensated by a corresponding inverse scaling of B so that the first term of the cost function remains unaffected. The unit-norm constraint for the templates B therefore prevents A from shrinking arbitrarily.

7.4.5 Normalization

The normalization constraint of the templates B can be implemented using Lagrange multipliers, leading to the constrained least-squares solution:

$$B_{kl} = \left(\tilde{A}^t_{kl}\tilde{A}_{tk'l'} + \Lambda_{kl,k'l'}\right)^{-1} \tilde{A}^{k'l'}_t X^t \,. \tag{7.14}$$

Here, $\Lambda_{kl,k'l'}$ represents a diagonal matrix of size $KL \times KL$ with K different Lagrange multipliers as parameters that need to be adjusted so that $B^l_k B_{kl} = 1$ for all k. This can be done with a Newton-Raphson root search of the K functions $f_k(\Lambda) = B^l_k B_{kl} - 1$. The K dimensional search for the Lagrange multipliers in Λ can be interleaved with updates of A and B. For simplicity however, in our first implementation we used the unconstrained least-squares solution ($\Lambda = 0$) and renormalized B and A every 10 iterations.

7.4.6 Modeling multiple sequences sharing common templates

We can consider a model where several data samples (or exemplars, trials) are encoded with common templates B but each having their own set of amplitudes. This model is a simple extension of the one discussed above, and can be used to perform batch training of the templates. We note that performing training in such a way is slightly different from concatenating all the files together and using the usual model. If we use concatenation, the algorithm might try to model what happens at the boundary, although no particular meaning should be sought for there. Moreover, the complexity of the algorithm roughly grows as $T \log T$ where T is the data length, thus for C data samples the complexity grows as $CT \log CT$ for the concatenation version, while it grows as $CT \log T$ for the batch learning version. Although this may not make a huge difference, it reflects the fact that the algorithm tries to do something (uselessly) more involved by considering the boundaries between data.

The model now reads

$$\hat{X}_{ct} \;=\; \sum_{nk} A_{cnn} B_{k,t-n} \tag{7.15}$$

$$= A_c^{nk} \tilde{B}_{nkt} \tag{7.16}$$

$$= \tilde{A}_{ct}^{kl} B_{kl}. \tag{7.17}$$

The update equations for A are unchanged as each A_c is optimized independently from the others, based only on data sample X_c. The update equations for the templates B can be simply obtained from the existing ones, as

$$B_{kl} = \left(\tilde{A}_{kl}^{ct} \tilde{A}_{ctk'l'} \right)^{-1} \tilde{A}_{ct}^{k'l'} X^{ct}, \tag{7.18}$$

which differs from the original updates only by an extra summation over c in both the numerator and denominator.

7.5 Convergence proof for the A update

We prove here that the updates (7.12) and (7.13) converge to a local minimum. If normalization is taken care of in the B updates, we are left with the convergence proof for the A updates. Ding et al. [63] already showed that this was the case without the sparseness term in Eq. (7.12). We will use an inequality derived by Kameoka to prove the convergence of sparseness-based NMF and complex NMF [109], and modify the convergence proof of Ding et al. to take into account the sparseness term. The additional term in the update equations was also used by Mørup to enforce sparseness in the NMF algorithm using an L^1 norm [142], but he only proved the convergence of his algorithm for a version without that sparseness term.

We first briefly review Ding et al.'s convergence proof using our notations. As usual in convergence proofs for NMF-like multiplicative updates, the proof is based on an auxiliary function method [119].

If we write

$$E(A) = \frac{1}{2} \sum_t \left(X_t - A^{kn} \tilde{B}_{nkt} \right)^2 + \beta \sum_{nk} A_{nk}^{\alpha} \tag{7.19}$$

with $\tilde{B}$ fixed, the goal is to find an auxiliary function $\tilde{E}(A, A')$ such that

$$E(A) \leq \tilde{E}(A, A'),\ E(A) = \tilde{E}(A, A) \tag{7.20}$$

for any A and A'. One can then easily see that a sequence of updates

$$A^{(t+1)} = \underset{A}{\operatorname{argmin}}\, \tilde{E}(A, A^{(t)}) \tag{7.21}$$

leads to a monotonic decrease in $E(A^{(t)})$.

We can rewrite $E(A)$ as

$$\begin{aligned} E(A) &= \frac{1}{2}\sum_t X_t^2 - \sum_{nkt} X_t A_{kn} \tilde{B}_{nkt} \\ &\quad + \frac{1}{2}\sum_{tnkn'k'} A_{kn}\tilde{B}_{nkt}A_{k'n'}\tilde{B}_{n'k't} + \beta\sum_{nk} A_{nk}^{\alpha} \\ &= \frac{1}{2}X^t X_t - A^{kn}(X^t\tilde{B}_{nkt}) \\ &\quad + \frac{1}{2}A^{nk}A^{n'k'}(\tilde{B}_{n'k'}^t\tilde{B}_{nkt}) + \beta\sum_{nk} A_{nk}^{\alpha}. \end{aligned} \tag{7.22}$$

If we now write $P_{nk} = X^t\tilde{B}_{nkt}$ and $Q_{nkn'k'} = \tilde{B}_{n'k'}^t\tilde{B}_{nkt}$, we can rewrite $E(A)$ as

$$\begin{aligned} E(A) &= \frac{1}{2}X^t X_t - A^{kn}P_{nk}^{+} + A^{kn}P_{nk}^{-} \\ &\quad + \frac{1}{2}A^{nk}A^{n'k'}Q_{nkn'k'}^{+} - \frac{1}{2}A^{nk}A^{n'k'}Q_{nkn'k'}^{-} \\ &\quad + \beta\sum_{nk} A_{nk}^{\alpha} \end{aligned} \tag{7.23}$$

Ding et al. showed that the following inequalities hold:

$$A^{nk}A^{n'k'}Q_{nkn'k'}^{+} \leq \sum_{nk}\frac{(A'^{n'k'}Q_{nkn'k'}^{+})A_{nk}^2}{A'_{nk}} \tag{7.24}$$

$$A^{kn}P_{nk}^{-} \leq \sum_{nk} P_{nk}^{-}\frac{A_{nk}^2 + A'^2_{nk}}{2A'_{nk}} \tag{7.25}$$

$$A^{kn}P_{nk}^{+} \geq \sum_{nk} P_{nk}^{+}A'_{nk}(1 + \log\frac{A_{nk}}{A'_{nk}}) \tag{7.26}$$

$$A^{nk}A^{n'k'}Q_{nkn'k'}^{-} \geq \sum_{nkn'k'} Q_{nkn'k'}^{-}A'_{nk}A'_{n'k'}(1 + \log\frac{A_{nk}A_{n'k'}}{A'_{nk}A'_{n'k'}}) \tag{7.27}$$

for any non-negative A and A'.

Combining these inequalities, one can derive an auxiliary function which leads to a local convergence proof for the shift-invariant semi-NMF updates without sparseness constraint. The sparseness term however needs special attention, and an inequality of its own. We use here an inequality derived by Kameoka [109], which states that

$$A_{nk}^{\alpha} \leq \frac{\alpha A'^{\alpha-2}_{nk}}{2}A_{nk}^2 + (1 - \frac{\alpha}{2})A'^{\alpha}_{nk},\ \forall\alpha,\ 0 < \alpha < 2. \tag{7.28}$$

Thanks to this inequality, the general Gaussian sparseness term can be replaced by a second-

order term. Altogether, we can define an auxiliary function for $E(A)$ as

$$\begin{aligned}\tilde{E}(A,A') &= \frac{1}{2}X^tX_t \\ &\quad - \sum_{nk} P^+_{nk}A'_{nk}(1+\log\frac{A_{nk}}{A'_{nk}}) \\ &\quad + \frac{1}{2}\sum_{nk} P^-_{nk}\frac{A^2_{nk}+A'^2_{nk}}{A'_{nk}} \\ &\quad + \frac{1}{2}\sum_{nk} \frac{(A'^{n'k'}Q^+_{nkn'k'})A^2_{nk}}{A'_{nk}} \\ &\quad - \frac{1}{2}\sum_{nkn'k'} Q^-_{nkn'k'}A'_{nk}A'_{n'k'}(1+\log\frac{A_{nk}A_{n'k'}}{A'_{nk}A'_{n'k'}}) \\ &\quad + \beta\sum_{nk}\frac{\alpha A'^{\alpha-2}_{nk}}{2}A^2_{nk} \\ &\quad + TK(1-\frac{\alpha}{2})A'^{\alpha}_{nk}. \end{aligned} \tag{7.29}$$

By differentiating $\tilde{E}(A,A')$ w.r.t. A_{nk} and setting it to 0, we obtain the following equation on A_{nk}:

$$P^+_{nk}\frac{A'_{nk}}{A_{nk}} + \frac{A'_{nk}}{A_{nk}}(A'^{n'k'}Q^-_{nkn'k'}) = P^-_{nk}\frac{A_{nk}}{A'_{nk}} + \frac{A_{nk}}{A'_{nk}}(A'^{n'k'}Q^+_{nkn'k'}) + \alpha\beta\frac{A_{nk}}{A'_{nk}}A'^{\alpha-1}_{nk}. \tag{7.30}$$

Multiplying both members of this equation by $A_{nk}A'_{nk}$ and grouping terms together, we obtain the following update equation for A in terms of A',

$$A_{nk} = A'_{nk}\sqrt{\frac{P^+_{nk}+A'^{n'k'}Q^-_{nkn'k'}}{P^-_{nk}+A'^{n'k'}Q^+_{nkn'k'}+\alpha\beta A'^{\alpha-1}_{nk}}}. \tag{7.31}$$

Updating A through (7.21) thus leads to

$$A_{nk} \leftarrow A_{nk}\sqrt{\frac{P^+_{nk}+A^{n'k'}Q^-_{nkn'k'}}{P^-_{nk}+A^{n'k'}Q^+_{nkn'k'}+\alpha\beta A^{\alpha-1}_{nk}}} \tag{7.32}$$

which is none other than (7.12).

7.6 Updating B using the natural gradient

7.6.1 Optimization under unit-norm constraints

We explained above how the update of B could be performed in two steps, through a least-squares solution to (7.11) followed by normalization to ensure that the templates B respect the unit-norm constraint. Although this may sound like the most efficient thing

to do to optimize the templates B, one may fear that, if extra care is not taken about normalization as explained in Section 7.4.5, the least-squares solution (7.13) makes too good a job at fitting the data by sending the templates B very far from the unit-norm manifold. The subsequent normalization may then result in an increase in the error. A slower but more cautious update, which would take into account as much as possible the unit-norm normalization on the templates, may thus be worth investigating. Gradient descent through the natural gradient [10] and classical gradient descent w.r.t. non-normalized templates B on an objective function in which the templates appear explicitly normalized [142] are natural candidates for such an optimization process. It turns out that for a unit-norm normalization with the L^2-norm, for which the constrained manifold is Riemannian and natural gradient can be defined, these two procedures are actually equivalent.

In general, let $\mathcal{I}(f)$ be an objective function with parameter $f \in \mathbb{R}^n$, and let $\mathcal{J}(f) = \mathcal{I}(\frac{f}{||f||})$ where $||f||$ is any vector norm. In the following, we shall note $v = \frac{\partial ||u||}{\partial u}\Big|_{u=f}$ and $\tilde{f} = \frac{1}{||f||} f$. Then

$$\frac{\partial \mathcal{J}}{\partial f_i} = \sum_j \frac{\partial \tilde{f}_j}{\partial f_i} \cdot \frac{\partial \mathcal{I}}{\partial f_j}(\tilde{f}), \tag{7.33}$$

and using

$$\frac{\partial \tilde{f}_j}{\partial f_i} = \begin{cases} \frac{1}{||f||} - \frac{f_i v_i}{||f||^2} & \text{if } j = i \\ -\frac{f_j v_i}{||f||^2} & \text{if } j \neq i \end{cases} \tag{7.34}$$

we obtain

$$\frac{\partial \mathcal{J}}{\partial f_i} = \frac{1}{||f||} \frac{\partial \mathcal{I}}{\partial f_i}(\tilde{f}) - v_i \sum_j \frac{f_j}{||f||^2} \cdot \frac{\partial \mathcal{I}}{\partial f_j}(\tilde{f}), \tag{7.35}$$

and finally

$$\nabla \mathcal{J} = \frac{1}{||f||} \left(\nabla \mathcal{I}(\tilde{f}) - (\tilde{f} \cdot \nabla \mathcal{I}(\tilde{f}))v \right), \tag{7.36}$$

or written in matrix form

$$\nabla \mathcal{J} = \frac{1}{||f||} \left(\mathrm{Id} - v\tilde{f}^T \right) \nabla \mathcal{I}(\tilde{f}). \tag{7.37}$$

For the L^2 norm, we have $v = \frac{f}{||f||}$, and Eq. (7.37) simplifies to

$$\nabla \mathcal{J} = \frac{1}{||f||} \left(\mathrm{Id} - vv^T \right) \nabla \mathcal{I}(\tilde{f}). \tag{7.38}$$

which is equal up to a scaling factor to the tangent gradient of $\mathcal{I}$ taken at point $\tilde{f}$. Introduced in [65], the tangent gradient is indeed given in general for a unit-norm constraint (with arbitrary norm) by

$$\nabla^{(T)} \mathcal{I}(w) = \left(\mathrm{Id} - \frac{vv^T}{||v||_2^2} \right) \nabla \mathcal{I}(w), \tag{7.39}$$

where $v = \frac{\partial ||u||}{\partial u}\Big|_{u=w}$. Note that even for a unit-norm constraint with any arbitrary norm, it is still the L^2 norm of v which appears in the expression of the tangent gradient. For the L^2-norm constraint, we have $||v||_2 = 1$, and thus

$$\nabla \mathcal{J}(f) = \frac{1}{||f||} \nabla^{(T)} \mathcal{I}(\tilde{f}). \tag{7.40}$$

In the particular case of the L^2-norm constraint as well, the tangent gradient is itself equivalent to the natural gradient [65]. Altogether, if the parameter vector f is normalized after each gradient update based on Eq. (7.38), the subsequent optimization procedure is equivalent to a natural gradient descent on the original objective function $\mathcal{I}$.

7.6.2 Update equations for the 1D model

Replacing in Eq. (7.6) the templates B by a normalized version leads to

$$\check{X}_t = \sum_n A_n^k \frac{B_{k,t-n}}{\sqrt{\sum_l B_{k,l}^2}}. \tag{7.41}$$

Computing the gradient w.r.t. B of the objective function

$$\check{E}(B) = \frac{1}{2} ||X - \check{X}||_2^2 + \beta \, ||A||_\alpha^\alpha \tag{7.42}$$

leads to the same gradient term as computing the natural gradient for the objective function E defined in Eq. (7.11) with unnormalized templates, as explained above.

A simple computation leads to

$$\frac{\partial \check{E}}{\partial B_{k'l'}} = -\sum_t (X_t - \check{X}_t) \{ \tilde{A}_{tk'l'} - B_{k'l'} \sum_l \tilde{A}_{tk'l} B_{k'l} \} \tag{7.43}$$

where we have assumed $\sum_l B_{k,l}^2 = 1, \forall k$. If we note $R_t = X_t - \check{X}_t$ the modeling error and $L_{k't} = \sum_l \tilde{A}_{tk'l} B_{k'l}$ the contribution of template $B_{k'}$ to the model, then we can rewrite the gradient as

$$\frac{\partial \check{E}}{\partial B_{k'l'}} = -\sum_t R_t \{ \tilde{A}_{tk'l'} - B_{k'l'} L_{k't} \}. \tag{7.44}$$

Introducing the $K \times L$ partial cross-correlation matrix

$$M_{kl} = \sum_t R_t \tilde{A}_{tkl} = \sum_t R_t A_{t-l,k}, \tag{7.45}$$

we have

$$\sum_t R_t L_{k't} = \sum_l M_{k'l} B_{k'l}, \tag{7.46}$$

and the gradient can be finally rewritten as

$$\frac{\partial \check{E}}{\partial B_{k'l'}} = -M_{k'l'} + B_{k'l'} \sum_l M_{k'l} B_{k'l}. \tag{7.47}$$

Thanks to this last expression, the gradient can be efficiently computed. The templates B can then be updated at step p through

$$B_{kl}^{(p+1)} \leftarrow B_{kl}^{(p)} - \mu_p \frac{\partial \check{E}}{\partial B_{kl}}, \tag{7.48}$$

where μ_p is a step size parameter.

Gradient update equations for the batch learning model introduced in Section 7.4.6 can also be obtained, here again with an extra summation on c.

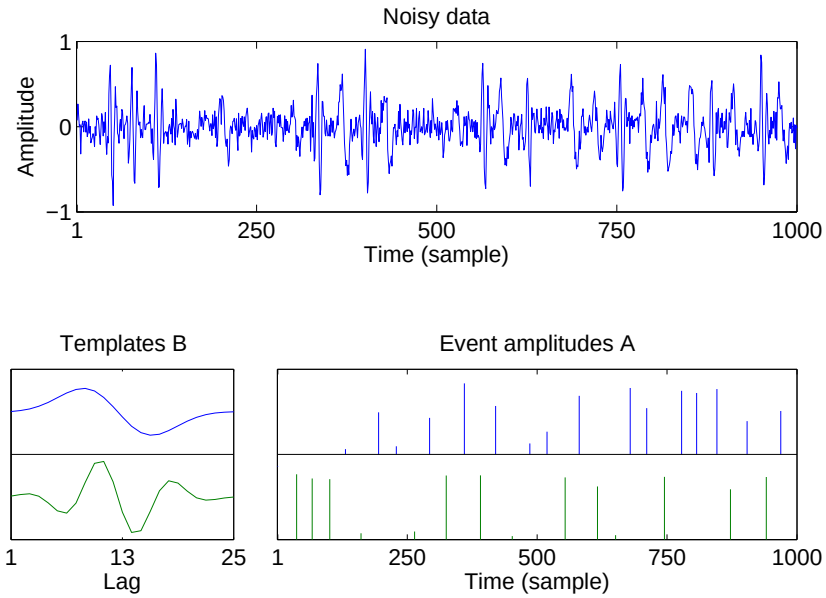
7.7 Performance evaluations

We evaluated the algorithm on synthetic and real data. Synthetic data are used to provide a quantitative evaluation of performance as a function of SNR and the similarity of different templates. The algorithm is then applied to extracellular recordings of neuronal spiking activity, where we evaluate its ability to recover two distinct spike types that are typically superimposed in this data, and to music data, where we show that it can be used to decompose a drum loop into its components, simultaneously learning their time course and detecting their onset timings.

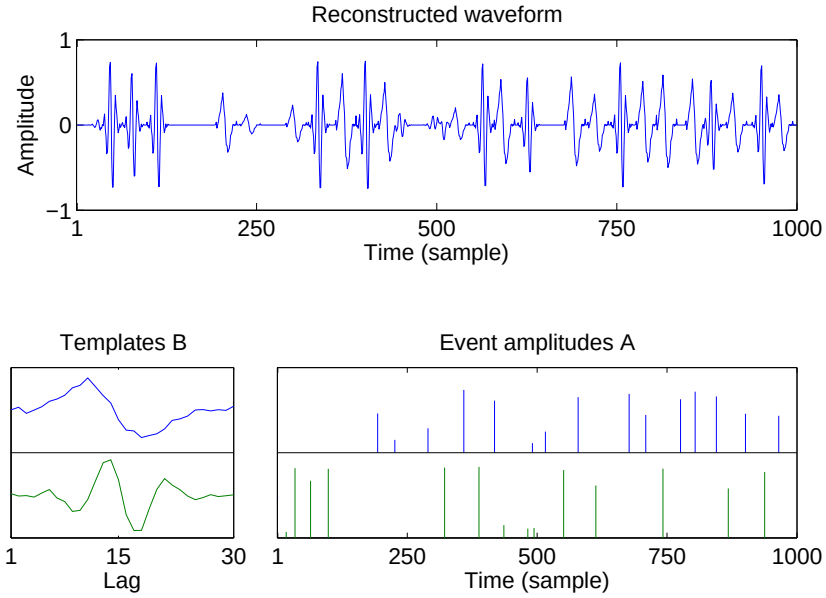
7.7.1 Quantitative evaluation on synthetic data

The goal of these simulations is to measure performance based on known truth data. We report detection rate, false-alarm rate, and classification error. In addition we report how accurately the templates have been recovered. We generated synthetic spike trains with two types of "spikes" and added Gaussian white noise. Figure 7.1 shows an example for SNR $= \sigma_A/\sigma_N = 2$ (or 6 dB). The two sets of panels show the templates B (original in Fig. 7.1 (a) and recovered in Fig. 7.1 (b)), amplitudes A (same as above) and noisy data X (Fig. 7.1 (a)) and estimated $\hat{X}$ (Fig. 7.1 (b)). The figure shows the model parameters which resulted in a minimum cost. Clearly, for this SNR the templates have been recovered accurately and their occurrences within the waveform have been found with only a few missing events.

Performance as a function of varying SNR is shown in Figure 7.2. Detection rate is measured as the number of events recovered over the total number of events in the original



(a) Noisy synthetic data (top) and synthetic parameters (bottom).



(b) Reconstructed waveform (top) and estimated parameters (bottom).

Figure 7.1: *Example of synthetic spike trains and estimated model parameters at an SNR of 2 (6 dB). The parameters are the templates B and weight matrices A.*

data. False alarms occur when noise is interpreted as actual events. Presence or absence of a recovered event is determined by comparing the original pulse train with the reconstructed pulse train A (channel number k is ignored). Templates in this example have a correlation time (3 dB down) of 2-4 samples and so we tolerate a misalignment of events of up to ± 2 samples. We simulated 30 events with amplitudes uniformly distributed in $[0, 1]$. The

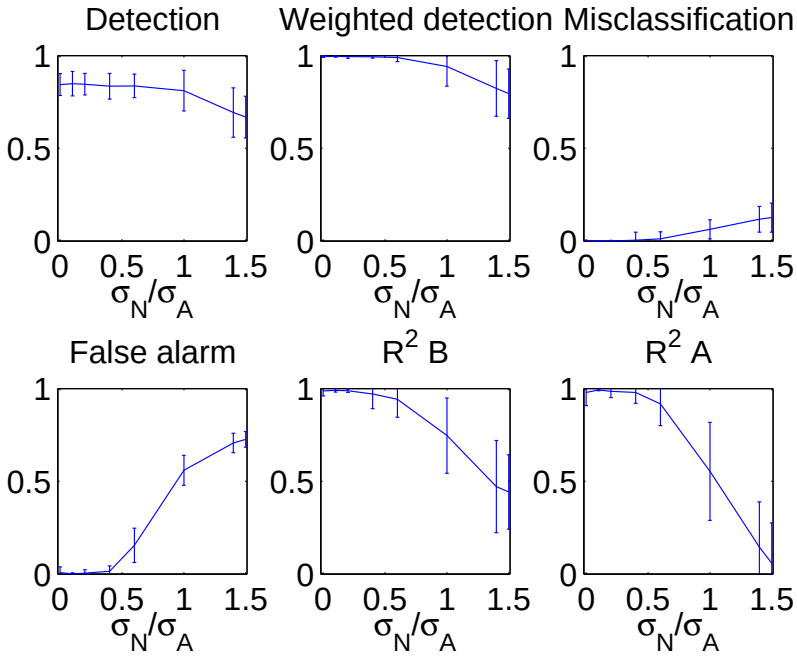


Figure 7.2: *Performance as a function of SNR. Error bars represent standard deviation over 100 repetitions with varying random amplitudes and random noise. Top left: detection rate. Top center: weighted detection rate. Top right: misclassification rate (rate of events attributed to the wrong template). Bottom left: false alarm rate (detected events which do not correspond to an event in the original data). Bottom center: R^2 of the templates B. Bottom right: R^2 of the amplitudes A.*

algorithm tends to miss smaller events with amplitudes comparable to the noise amplitude. To capture this effect, we also report a detection rate that is weighted by event amplitude. Some events may be detected but assigned to the wrong template. We therefore report also classification performance. Finally, we report the goodness of fit as R^2 for the templates B and the continuous valued amplitudes A for the events that are present in the original data.

Note that the proposed algorithm implements implicitly a clustering and classification process. Obviously, the performance of this type of unsupervised clustering will degrade as the templates become more and more similar. Figure 7.3 shows the same performance numbers as a function of the similarity of the templates (without additive noise). A similarity of 0 corresponds to the templates shown as examples in Figure 7.1 (these are almost orthogonal with a cosine of 74°), and similarity 1 means identical templates. Evidently the algorithm is most reliable when the target templates are dissimilar.

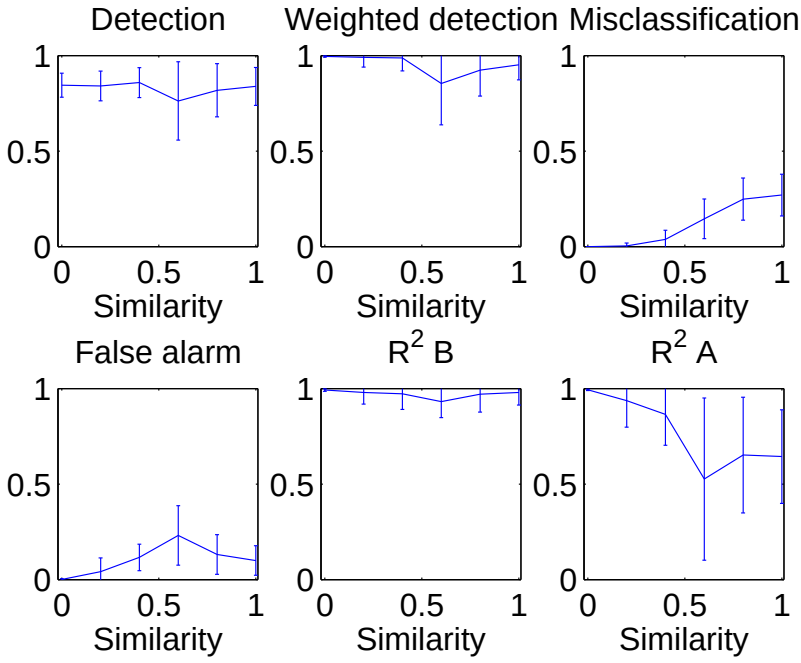


Figure 7.3: *Performance as a function of similarity. Error bars represent standard deviation over 100 repetitions with varying random amplitudes. Top left: detection rate. Top center: weighted detection rate. Top right: misclassification rate (rate of events attributed to the wrong template). Bottom left: false alarm rate (detected events which do not correspond to an event in the original data). Bottom center: R^2 of the templates B. Bottom right: R^2 of the amplitudes A.*

7.7.2 Analysis of extracellular recordings

The original motivation for this algorithm was to analyze extracellular recordings from single electrodes in the guinea pig cochlear nucleus. Spherical and globular bushy cells in the anteroventral cochlear nucleus (AVCN) are assumed to function as reliable relays of spike trains from the auditory nerve, with "primary-like" responses that resemble those of auditory nerve fibers. Every incoming spike evokes a discharge within the outgoing axon [105]. However, recent observations give a more nuanced picture, suggesting that the post-synaptic spike may sometimes be suppressed according to a process that is not well understood [12].

Extracellular recordings from primary-like cells within AVCN with a single electrode typically show a succession of events made up of three sub-events: a small pre-synaptic spike from the large auditory nerve fiber terminal, a medium-sized post-synaptic spike from the initial segment of the axon where it is triggered (the IS spike), and a large-sized spike produced by back-propagation into the soma and dendrites of the cell (the soma-dendritic or

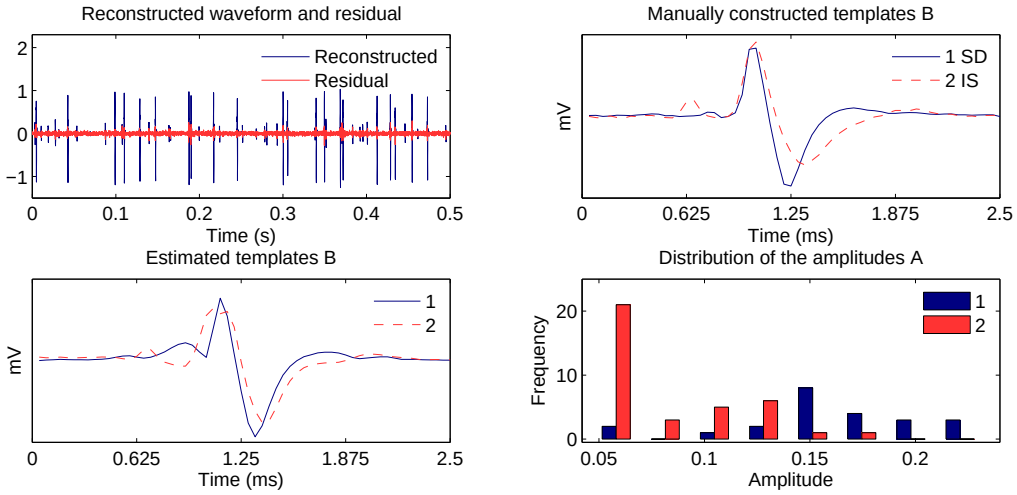


Figure 7.4: *Experimental results on extracellular recordings. Top left: reconstructed waveform (blue) and residual between the original data and the reconstructed waveform (red). Top right: templates B estimated manually from the data. Bottom left: estimated templates B. Bottom right: distribution of estimated amplitudes A. The SD spikes (blue) generally occur with larger amplitudes than the IS spikes (red).*

SD spike) (Fig. 7.4). Their relative amplitudes depend upon the position of the electrode tip relative to the cell. Our aim is to isolate each of these components to understand the process by which the SD spike is sometimes suppressed. The events may overlap in time (in particular the SD spike always overlaps with an IS spike), with varying positive amplitudes. They are temporally compact, on the order of a millisecond, and they occur repeatedly but sparsely throughout the recording. The assumptions of our algorithm are met by these data, as well as by multi-unit recordings reflecting the activity of several neurons (the "spike sorting problem").

In the portions of our data that are sufficiently sparse (spontaneous activity), the components may be separated by an ad-hoc procedure: (a) trigger on the high-amplitude IS-soma complexes and set to zero, (b) trigger on the remaining isolated IS spikes and average to derive an IS spike template (the pre-synaptic spike is treated as part of the IS spike), (c) find the best match (in terms of regression) of the initial portion of the template to the initial portion of each IS-SD complex, (d) subtract the matching waveform to isolate the SD spikes, realign, and average to derive an SD spike template. The resulting templates are shown in Fig. 7.4 (top right). This ad-hoc procedure is highly dependent on prior assumptions, and we wished to have a more general and "agnostic" method to apply to a wider range of situations.

Figure 7.4 (bottom) shows the result of our automated algorithm. The automatically recovered spike templates seem to capture a number of the key features. Template 1, in blue, resembles the SD spike, and template 2, in red, is similar to the IS spike. The SD spikes are larger and have sharper peaks as compared to the IS spikes, while the IS spikes have an initial peak at 0.7 ms leading the main spike. The larger size of the extracted spikes corresponding to template 1 is correctly reflected in the histogram of the recovered amplitudes. However the estimated spike shapes are inaccurate. The main difference is in the small peak preceding the template 1. This is perhaps to be expected as the SD spike is always preceded in the raw data by a smaller IS spike. The expected templates were very similar (with a cosine of 38° as estimated from the manually extracted spikes), making the task particularly difficult.

7.7.3 Analysis of music data

We also tested our model on the extraction of audio waveforms. The waveforms of audio signals are in general much longer than the spikes encountered in the preceding section, and their extraction thus involves much larger computational cost. Their shape is also much more complex than that of the short spikes we extracted in the previous section. To test our algorithm on such data, we performed a decomposition and separation experiment on waveforms constituted of overlapping drum sounds. The data are semi-synthetic, in that we constructed them with real drum sound waveforms repeated identically, possibly with varying amplitudes or with additive noise. This kind of situation could be encountered for example in music recorded with electronic drums for example, where sample sounds are triggered by a human player.

The drum sounds were a bass drum sound and a snare drum sound taken from the RWC music database [87]. The original sounds were downsampled from 44.1 kHz to 4 kHz to lower the computational cost, and their length was cut to 1000 samples (0.25 s). The input data was then constructed by combining several occurrences of these two templates at random times. The length of the input data was set to 8000 samples (2 s), and the total number of activations was 10, each activation being either attributed to the bass drum or the snare drum with probability 0.5. The activation time were selected by dividing the total time interval into equal length sub-intervals and adding to the left bound of each interval a random time lag uniformly distributed between zero and one fifth of the length of each sub-interval. The actual overlap between two consecutive sounds was thus between 25 % and 50 %. The amplitudes were randomly determined using a uniform distribution on $[0.5; 1.5]$.

The original data is shown in Fig. 7.5 (a). The sound templates are shown in the bottom left, and their corresponding activations in the bottom right. The bass drum sound is on top, in blue, and the snare drum sound is at the bottom, in green. In a second experiment, we added Gaussian white noise to the data of the first experiment at an SNR of 0.1 (20 dB). The original noisy data is shown in Fig. 7.6 (a).

The number of templates to be extracted was set to $K = 2$, and their length to $L = 1200$ samples to cope with the loss of extremal parts of the templates due to their centering, as we will describe in Section 7.7.4. In both experiments, the amplitudes A were updated at each step, while the templates B were only updated every 10 steps, and the sparseness prior shape parameter α was again set to $1/4$. For the experiment on clean data, the initial sparseness coefficient β was set to 0.3, while for the experiment on noisy data, it was set to 0.5. The results are shown in Fig. 7.5 (b) and Fig. 7.6 (b) for each of the two experimental settings respectively.

Although the bass drum and snare drum sound templates are quite different, with a cosine measure of 76°, the situation is still very difficult as the snare drum sound is made of a double stroke, as can be seen in the bottom left part of Fig. 7.5 (a). The correlation time of the snare drum template is about 40 samples, making it both harder to learn its shape and to precisely detect its activation times. At around 20 samples, the auto-correlation is only decreased to about 75 % of its maximum value (-1.25 dB). Moreover, the energy of the snare drum sound was lower than that of the drum sound, and it was thus harder to learn and detect, especially in the presence of noise. However, our algorithm still performed rather well on these data. The waveform was reconstructed with an SNR of 12.7 dB in the clean data case, and 11.9 dB in the noisy case. Both with and without additive Gaussian noise, the R^2 of the estimated bass drum template was about 0.97 (or an SNR of 15.8 dB without noise and 14.9 dB with noise), and that of the estimated snare drum template was about 0.90 (SNR of 9.9 dB) without additive noise and 0.81 with additive noise (SNR of 6.5 dB). Allowing for ± 2 shifts in the time direction, the detection rate was 100 % without noise and 90 % with noise: one snare drum activation was shifted by about 20 samples, probably due to the high correlation of the snare drum sound template with itself for a 20 sample time lag. The R^2 of the activations was about 0.999 for the bass drum in both situations, while it was about 0.993 for the snare drum without noise and about 0.963 with noise, allowing 20 sample shifts. There was only one false alarm, for the bass drum sound in the experiment without noise, but the corresponding amplitude was very small, as can be seen in Fig. 7.5 (b) in the top part of the bottom right figure.

7.7.4 Implementation details

As with the original NMF and semi-NMF algorithms, the present algorithm is only locally convergent. To obtain good solutions, we restart the algorithm several times with random initializations for A (drawn independently from the uniform distribution in $[0, 1]$) and select the solution with the maximum posterior likelihood or minimum cost (7.11). In addition to these multiple restarts, we use a few heuristics that are motivated by the desired results of spike detection and drum sound decomposition. We can thus prevent the algorithm from converging to some obviously suboptimal solutions:

Re-centering the templates: We noticed that local minima with poor performance typically occurred when the templates B were not centered within the L lags. In those cases the main peaks could be adjusted to fit the data, but the portion of the template that extends outside the window of L samples could not be adjusted. To prune these suboptimal solutions, it was sufficient to center the templates during the updates while shifting the amplitudes accordingly.

Pruning events: We observed that spikes tended to generate non-zero amplitudes in A in clusters of 1 to 3 samples. After convergence we compact these to pulses of 1-sample duration located at the center of these clusters. Spike amplitude was preserved by scaling the pulse amplitudes to match the sum of amplitudes in the cluster.

Re-training with a less conservative sparseness constraint: To ensure that templates B are not affected by noise we initially train the algorithm with a strong penalty term (large β effectively assuming strong noise power σ_N^2). Only spikes with large amplitudes remain after convergence and the templates are determined by only those strong spikes that have high SNR. After extracting templates accurately, we retrain the model amplitudes A while keeping the templates B fixed assuming now a weaker noise power (smaller β).

As a result of these steps, the algorithm converged frequently to good solutions (approximately 50 % of the time on the simulated data). The performance reported here represents the results with minimum error after 6 random restarts.

7.8 Discussion

Alternative models: The present 1D formulation of the problem is similar to that of Mørup et al. [142] who presented a 2D version of this model that is limited to non-negative templates. We have also derived a version of the model in which event timing is encoded explicitly as time delays τ_n following [140]. We are omitting this alternative formulation here

for the sake of brevity.

Alternative priors: In addition to the generalized Gaussian prior, we tested also Gaussian process priors [162] to encourage orthogonality between the k sequences and refractoriness in time. However, we found that the quadratic expression of a Gaussian process competed with the L^α sparseness term. The combination of both criteria could be investigated by allowing for correlations in the generalized Gaussian. The corresponding distributions are known as elliptically symmetric densities [77] and the corresponding process is called a spherically invariant random processes [161].

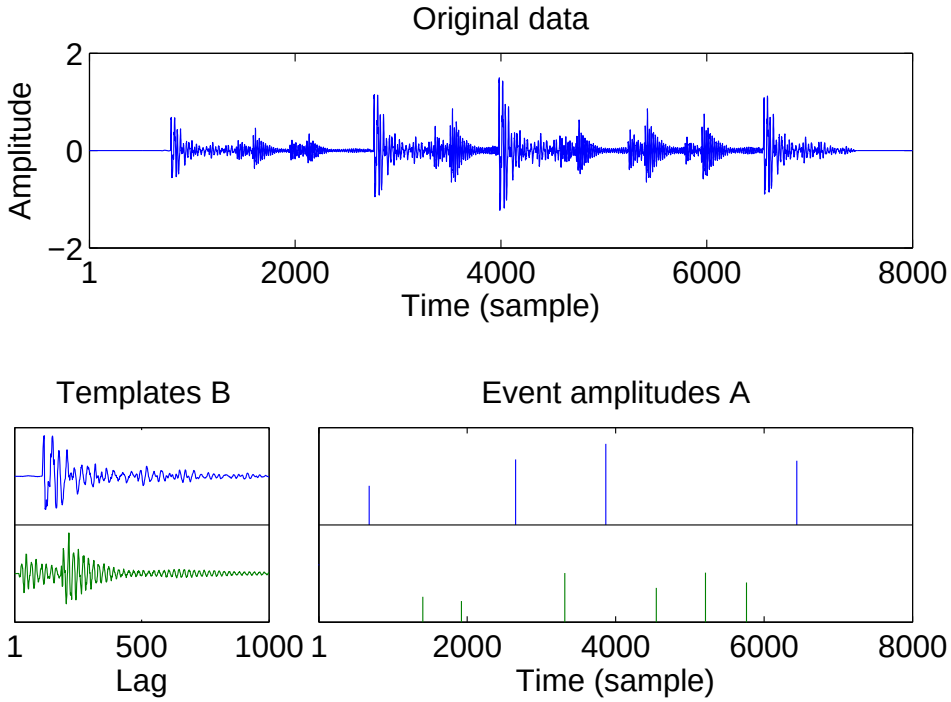
Sparseness and dimensionality reduction: As with many linear decomposition methods, a key feature of the algorithm is to represent the data within a small linear subspace. This is particularly true for the semi-NMF algorithm since, provided a sufficiently large K and without enforcing a sparsity constraint, the positivity constraint on A actually amounts to no constraint at all (identical templates with opposite sign can accomplish the same as allowing negative A). For instance, without sparseness constraint on the amplitudes, a trivial solution in our examples above would be a template B_{1l} with a single positive spike somewhere and another template B_{2l} with a single negative spike, and all the time course encoded in A_{n1} and A_{n2}.

MISO identification: The identifiability problem is compounded by the fact that the estimation of templates B in this present formulation represents a multiple-input single-output (MISO) system identification problem. In the general case, MISO identification is known to be under-determined [25]. In the present case, the ambiguities of MISO identification may be limited due to the fact that we allow only for limited system length L as compared to the number of samples N. Essentially, as the number of examples increases with increasing length of the signal X, the ambiguity in B is reduced.

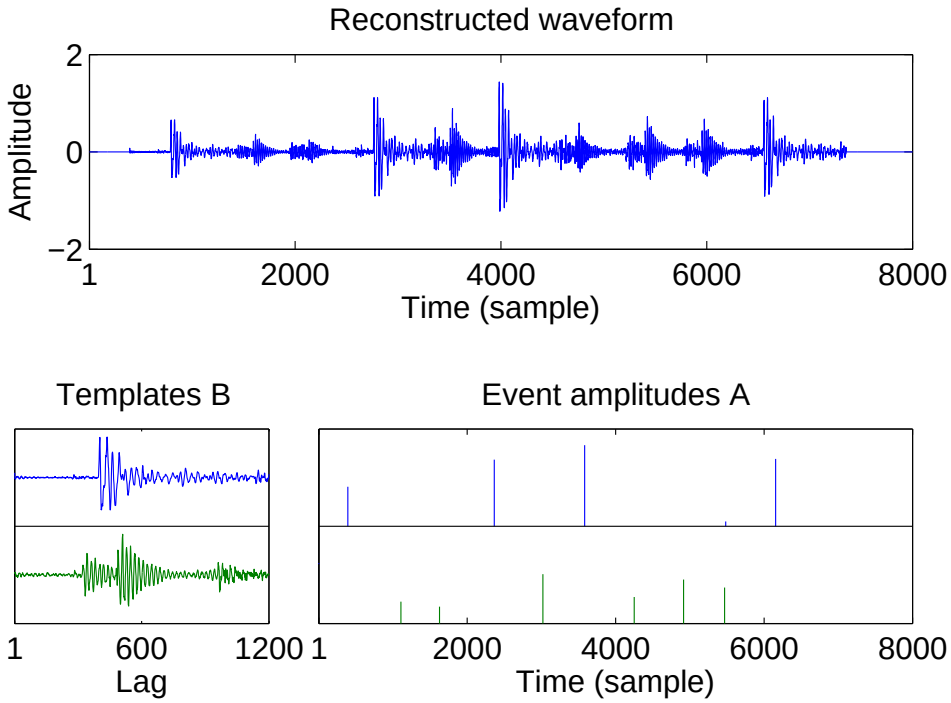
7.9 Summary of Chapter 7

In this chapter, we presented a model in the time domain, called shift-invariant semi-NMF, for the decomposition of waveforms in a limited number of elementary constituents, added with variable latencies and variable but positive amplitudes. We further introduced a sparseness prior on the amplitudes, to ensure that the extracted constituents capture meaningful information reappearing at various time instants. We explained how to optimize its parameters, giving a convergence proof including the sparseness term. We studied the performance of the model on synthetic data, and showed that the model could be used to effectively

recover recurring templates and their activation times from a mixed waveform, for audio signals (separation of overlapping drum sounds) as well as for extracellular recordings (spike sorting with overlapping spikes).

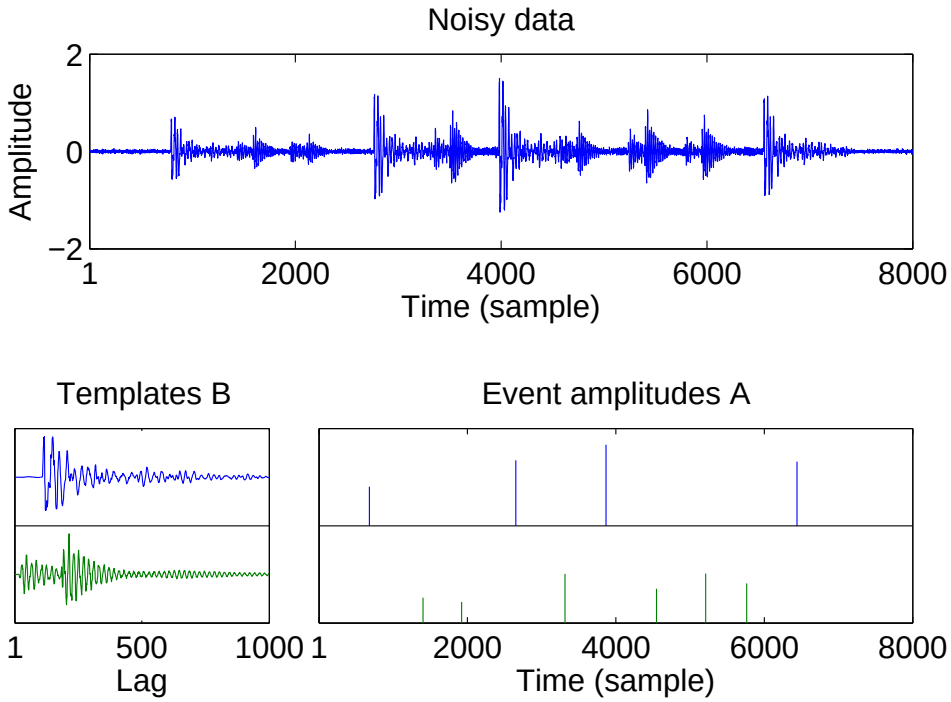


(a) Drum loop waveform (top) and original parameters (bottom).

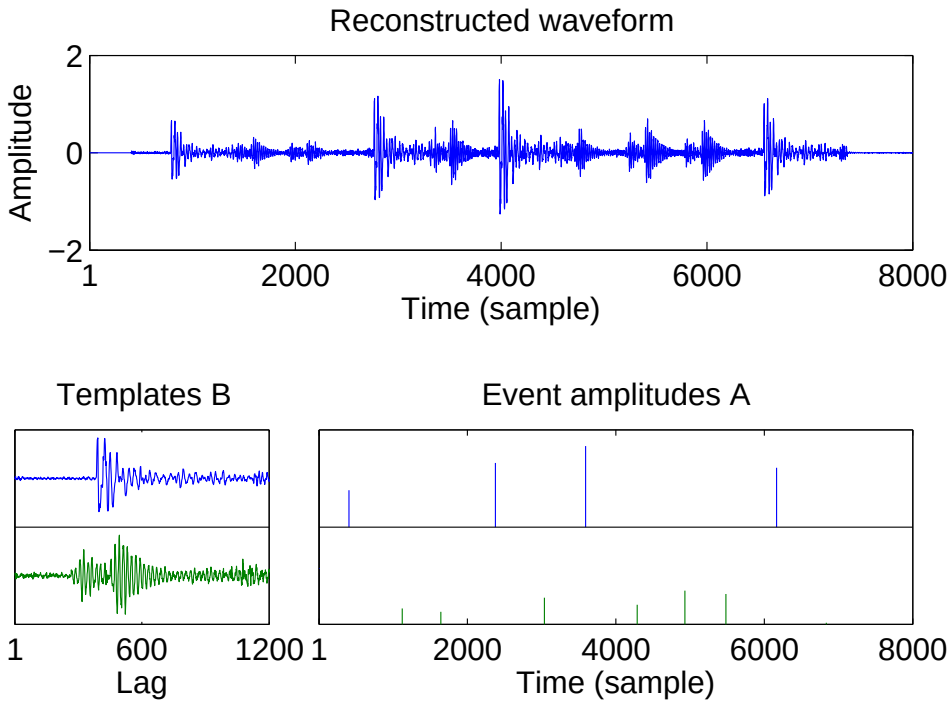


(b) Reconstructed waveform (top) and estimated parameters (bottom).

Figure 7.5: *Example of decomposition of a drum loop waveform. The amplitudes of the activations in the original waveform are randomly determined.*



(a) Drum loop waveform (top) and original parameters (bottom).



(b) Reconstructed waveform (top) and estimated parameters (bottom).

Figure 7.6: *Example of decomposition of a drum loop waveform with additive Gaussian noise at an SNR of 0.1 (20 dB). The amplitudes of the activations in the original waveform are randomly determined.*

Chapter 8

Data-driven learning of FIR filterbanks

8.1 Introduction

In Chapter 7, we explained that acquiring models from the data in an unsupervised way is an important problem, and proposed there a time-domain framework for the extraction of elementary templates inside a waveform. We also noted in the introduction that one of the advantages of working in the time domain is to not rely on a particular choice of analysis parameters. Finding "natural" parameters which suit a particular signal for time-frequency analysis is indeed an important issue, and data-driven optimization approaches for their determination have recently been investigated in speech processing (e.g., [94, 152]). The analogy with the adaptation of the auditory system on one side and with the importance of extracting robust cues to start the bootstrap process of the acquisition of languages on the other side are the ideas we focused on to formulate the framework we develop in this chapter.

The task of finding words in running speech is difficult because of the lack of obvious acoustic markers at word boundaries, such as onset transients or silent pauses. Adult speakers of a language might conceivably solve the problem by matching templates of words stored in a lexicon to the incoming acoustic stream. Infants do not have this option, as they lack a lexicon. How can one acquire a lexicon without knowledge of how to segment speech into the appropriate chunks to store in this lexicon? Among other hypotheses, it has been suggested [46] that the modulation structure related to prosodic and segmental organization of

speech might allow the infant's developing perceptual system to identify initial anchors that facilitate the acquisition of a more complete set of speech part boundaries.

The concepts of *temporal envelope* and *modulation spectrum* are gaining momentum in auditory science (e.g., [54]), speech science (e.g., [88]), automatic speech recognition (e.g., [95]), and evaluation of auditory impairments (e.g., [125]). "Modulation" can be defined as a part of the temporal structure of the acoustic waveform that is not well captured by standard spectral representations based on the Fourier Transform of the raw waveform. Modulation features extend over wider temporal spans (and thus lower frequencies) than represented in the audio spectrum. They describe the shape of the temporal *amplitude envelope* of the stimulus waveform, rather than the waveform itself. The pitch of a sound and most aspects of its timbre (for example vowel timbre) are usually assumed to reflect the audio spectrum, whereas the perception of roughness, rhythm and long-term temporal structure, for example of speech, are associated with modulation. Processing of temporal envelope structure is assumed to be distinct from that of temporal "fine structure" (e.g., [125]), although there is some overlap in the region of pitch. Both are presumably important for the perception of speech, and a number of studies have attempted to tease apart their respective roles using vocoded speech in which either envelope or fine structure is degraded (e.g., [125, 179]).

Modulation thus seems to play a central role for auditory perception, and if we were to consider the possibility of a tuning of the initial acoustic processing by exposure to the regularities of speech, it would make sense to assume that during the course of development and/or evolution, the human ear and brain adapted for modulation analysis through a data-driven learning process. We shall design a mathematical framework to investigate this hypothesis.

Perception of modulation presumably arises from the analysis of neural activity within each channel from the cochlea. Sensitivity to modulation has been ascribed to the existence of a "modulation filterbank" [54] implemented within the auditory brainstem or midbrain (e.g., [62]) or cortex [71]. We focus our study on the optimization of the combination of peripheral and central filterbanks to best extract the modulation structure of the input data. This is relevant for the hypothesis that such a criterion might in part drive the design of the human auditory system.

8.2 The temporal envelope

Intuitively, the temporal envelope of a signal is a smooth function that bounds the amplitude of its oscillations. We expect the envelope to remain positive and vary slowly, while the

carrier or fine structure makes faster positive and negative excursions. It is straightforward to synthesize a waveform based on such a description, but harder to demodulate an existing signal into envelope and fine structure. The task may seem trivial ("draw" a line connecting waveform peaks), but it is hard to perform in full generality.

Demodulation usually involves two steps: a non-linearity to produce positive values, and temporal smoothing to give them an "envelope-like" time-course. Popular non-linearities are a full-wave rectifier (absolute value), half-wave rectifier (analogous to cochlear transduction), or square (instantaneous power), possibly followed by a logarithmic transform (dB scale). Smoothing usually involves some form of low-pass filtering. We choose instantaneous power as the non-linearity. For a waveform $s(t)$, we will note

$$v(t) = s(t)^2 \tag{8.1}$$

and define its "temporal envelope" $w(t)$ as

$$w(t) = \mathcal{L}_{\omega_c}(v)(t) = \mathcal{L}_{\omega_c}\left(s^2\right)(t), \tag{8.2}$$

where $\mathcal{L}_{\omega_c}$ denotes a low-pass filter with cut-off frequency ω_c.

This quantity is not very relevant perceptually if derived directly from the acoustic waveform, as one can argue that the ear has access only to channels filtered by the cochlea. Accordingly, it is common to apply Eq. (8.2) to the outputs of a filterbank, for example a cochlear model or some other type of spectro-temporal analysis. This produces in effect a *spectro-temporal envelope*, or array of frequency-specific temporal envelopes.

Output $u_j(t)$ of channel j of the initial filterbank is related to the acoustic input $s(t)$ by convolution:

$$u_j(t) = f_j * s(t) = \sum_{k=0}^{K} f_j(k)s(t-k) \tag{8.3}$$

where $f_j(t)$ is the impulse response of the jth filter (approximated here as a K-tap finite-impulse response filter). The spectro-temporal envelope at time and frequency indices t and j can then be defined as:

$$w_j(t) = \mathcal{L}_{\omega_c}(v_j)(t) = \mathcal{L}_{\omega_c}\left((f_j * s)^2\right)(t). \tag{8.4}$$

Our goal here is to optimize this initial filterbank under a certain criterion, which we will describe in the next section, such that it is "suited" for modulation analysis.

8.3 Description of the model

8.3.1 Objective

We are looking for a filterbank which would be adapted to extract the modulation present in a signal. The idea is to maximize the "modulation energy" of the filter outputs, defined as the energy $||w_j||$ (with $||\cdot||$ denoting the L^2 norm) of the temporal envelope w_j obtained after rectifying and smoothing (here we low-pass at 20 Hz) the output signal u_j. In order to avoid trivial solutions such as several filters converging to the same optimal filter, we also introduce an orthogonality constraint on the filters.

8.3.2 Formulation of the objective function

Let us denote by $s(t)$ the input signal, and let $F = (f_1, \ldots, f_N)$ be a $K \times N$ matrix representing the filterbank to optimize, such that $F_{ij} = f_j(i)$. Each of its columns corresponds to an FIR filter of order K. We suppose that F verifies

$$F^T F = \mathrm{Id}, \tag{8.5}$$

i.e., F lies on the Stiefel manifold $V_N(\mathbb{R}^K)$ of ordered N-tuples of orthonormal vectors of $\mathbb{R}^K$. This means simply that the filters are normalized and mutually orthogonal.

Our optimization problem can now be stated as the maximization of the *total modulation energy* $\mathcal{I}(F) = \sum_j ||w_j||$, where w_j is defined as in Eq. (8.4), with respect to F under the condition that F lies on the Stiefel manifold. The objective function to maximize is thus

$$\mathcal{I}(F) = \sum_j \sqrt{\int \left(\mathcal{L}_{\omega_c}\left((f_j * s)^2\right)\right)^2 (t)\, dt}. \tag{8.6}$$

The process leading to the definition of $\mathcal{I}$ is illustrated in Fig. 8.1.

It is important to find an effective optimization method which is able to take into account the constraint (8.5). As it is difficult to obtain an analytical solution, a gradient method is indicated but it suffers from the fact that the updated filterbank is not guaranteed to stay on the Stiefel manifold. A first solution to this problem could be to project back to the Stiefel manifold after each update, using the fact that the closest matrix to M on the Stiefel manifold is given by [127]

$$\hat{M} = M(M^T M)^{-\frac{1}{2}}. \tag{8.7}$$

However, it seems slightly risky and ineffective to leave the Stiefel manifold, and an opti-

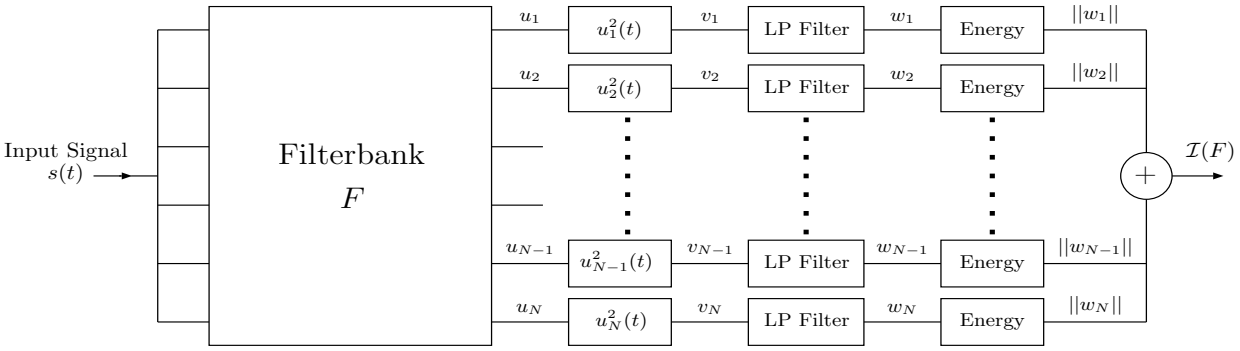


Figure 8.1: *Diagram of the model for the computation of the modulation energy.*

mization method which would take into account the particular geometrical structure of the constraint space is desirable. The natural gradient method is the natural tool for this kind of tasks [10], and in the particular case of the Stiefel manifold, the update goes as follows [66]. While the classical gradient method update is

$$F_{(n+1)} = F_{(n)} + G_{(n)}, \tag{8.8}$$

where

$$G_{(n)} = \mu(n)\frac{\partial \mathcal{I}}{\partial F}(F_{(n)}) \tag{8.9}$$

is the scaled (Euclidean) gradient of the cost function with respect to F evaluated at $F_{(n)}$, and $\mu(n)$ is a chosen step size sequence, the natural gradient method update can be written

$$F_{(n+1)} = F_{(n)} + G_{(n)}F_{(n)}^T F_{(n)} - F_{(n)}G_{(n)}^T F_{(n)}. \tag{8.10}$$

Although the natural gradient update can be proven [66] to stay in the constraint space for continuous flows, the discrete-time version presented above is numerically unstable and slowly diverges from the Stiefel manifold (making it impossible to simplify $F_{(n)}^T F_{(n)}$ in (8.10)). We thus still use the projection (8.7) every few steps to correct this tendency.

The derivative of the objective function with respect to F can be obtained as follows:

$$\frac{\partial \mathcal{I}}{\partial F_{i_0 j_0}} = \frac{1}{||w_{j_0}||}\int (\mathcal{L}_{\omega_c}(u_{j_0}^2))\big(\mathcal{L}_{\omega_c}(2\mathcal{T}_{i_0}(s)u_{j_0})\big)(t)\,dt, \tag{8.11}$$

where $\mathcal{T}_{i_0}$ is a shift operator, i.e., $\forall t, \mathcal{T}_{i_0}(s)(t) = s(t - i_0)$.

8.4 Simulations and results

The method is computationally expensive. We present preliminary results on a small sample (12 s) of speech uttered by both male and female speakers.

8.4.1 Experimental Procedure

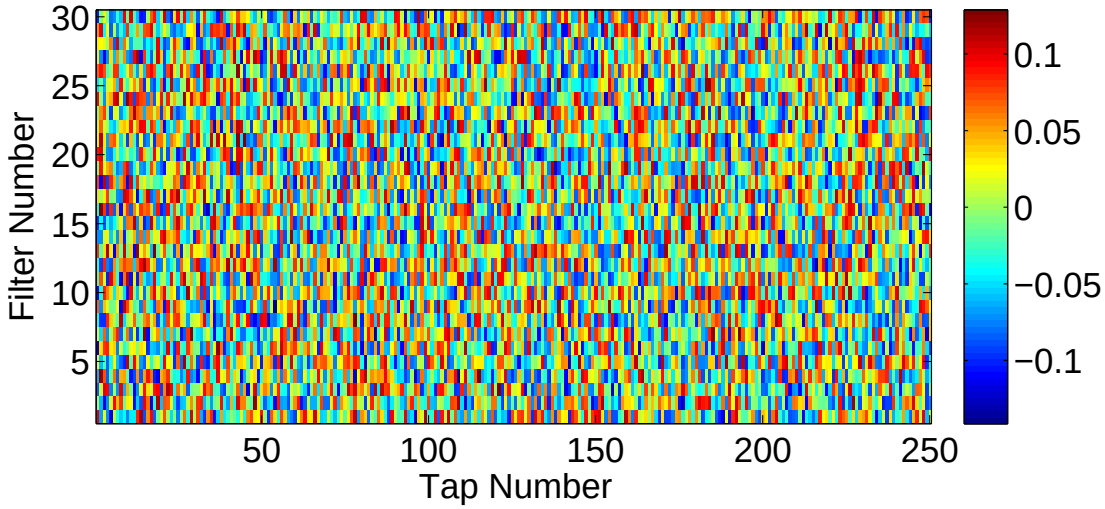
The sampling rate was 16 kHz. The initial low-pass filter applied to the envelope was implemented by convolution with a triangular window, the cutoff frequency being set to 20 Hz based on classical speech perception considerations [88]. We chose a low-pass filter with a non-negative impulse response to guarantee that the filtered envelope be non-negative as well. The filterbank consisted of 30 FIR filters with 250 taps, and was initialized by generating a random matrix with coefficients uniformly distributed on $[-0.5; 0.5)$ and then projecting it back to the Stiefel manifold (shown in Fig. 8.2 (a)). The initial value of $\mu(n)$ was set to 0.1, divided by 2 if an update yielded an energy decrease and multiplied by 1.3 after three steps without decrease.

8.4.2 Results

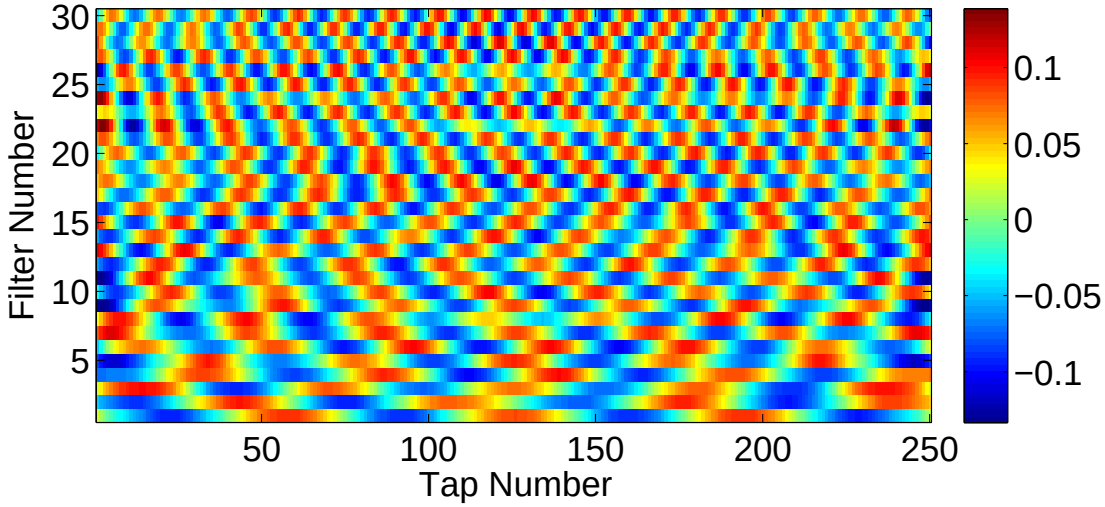
The optimized filterbank is shown in Fig. 8.2 (b). Each horizontal line represents the temporal envelope of one of the filters, which are ordered by decreasing center frequency from top to bottom. We notice that the filters are narrowband, with center frequencies ranging from 230 Hz to 1140 Hz, and usually go in pairs (in quadrature). Given the small training set, we should not assign too much significance to these values.

As a comparison, we show in Fig. 8.3 a modulation curve obtained at the output of one of the filters in response to the waveform of Fig. 8.4 ("I'd like to leave this in your safe" uttered by a female speaker, taken from the Bagshaw database [15], which we already used in Chapter 4), along with the curves obtained with Gammatone and DCT filters with the same center frequency 287 Hz. One can notice that the temporal envelopes extracted are very close. This is confirmed globally by computing the correlations between each modulation curve obtained with an optimized filter and the one obtained with Gammatone and DCT filters, as shown in Fig. 8.5. Thus our data-driven approach gives results that are quite similar, in particular, to the properties of cochlear filters. The power spectrogram obtained from the optimized filterbank is shown in Fig. 8.6. We can see by comparing it to a classical FFT-based spectrogram, shown in Fig. 8.7, that our result is pertinent.

It is too big a step to conclude from this study that the human ear has been developed under such a modulation-based criterion. However we see it as "proof of concept" that such



(a) Initial orthogonalized random filterbank.



(b) Optimized filterbank.

Figure 8.2: *Example of optimization result.*

criteria can be formulated to optimize information processing stages, including initial feature extraction, in a data-driven manner. We have developed a methodology that can be applied to a wider range of optimization criteria and targets to be optimized. An aim for future work will be to derive the number of filter channels and the low-pass cut-off frequency from the data, rather than imposing them.

8.5 Summary of Chapter 8

In this chapter, we introduced a framework for data-driven modulation analysis in which the initial filterbank analysis is optimized based on a criterion of maximum modulation energy within the low-frequency band, subject to orthogonality constraints between filters.

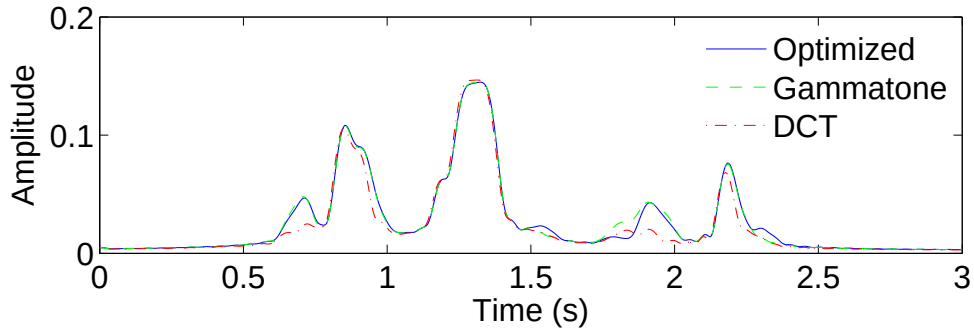


Figure 8.3: *Modulation curves for one of the optimized filters and for the Gammatone and DCT filters with the same center frequency* 287 *Hz.*

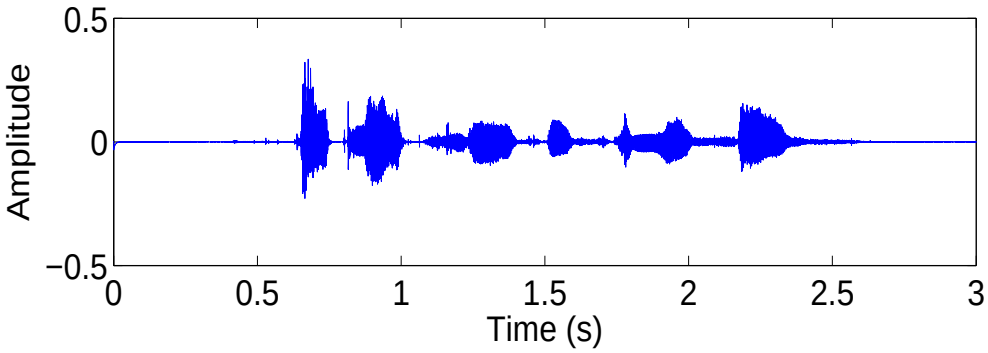


Figure 8.4: *Waveform of the input speech file.*

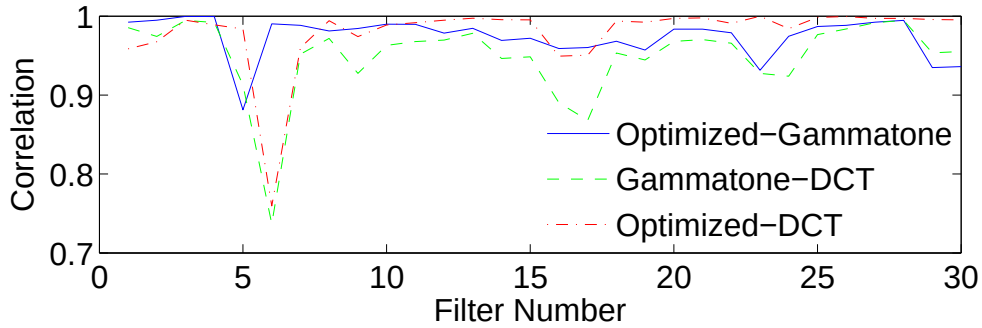


Figure 8.5: *Correlations between the modulation curves obtained with three types of filter having the same center frequencies.*

The idea was tested using speech data, and the results obtained were close to classical filterbanks, showing that the hypothesis of a tuning of an auditory system on such a criterion is pertinent.

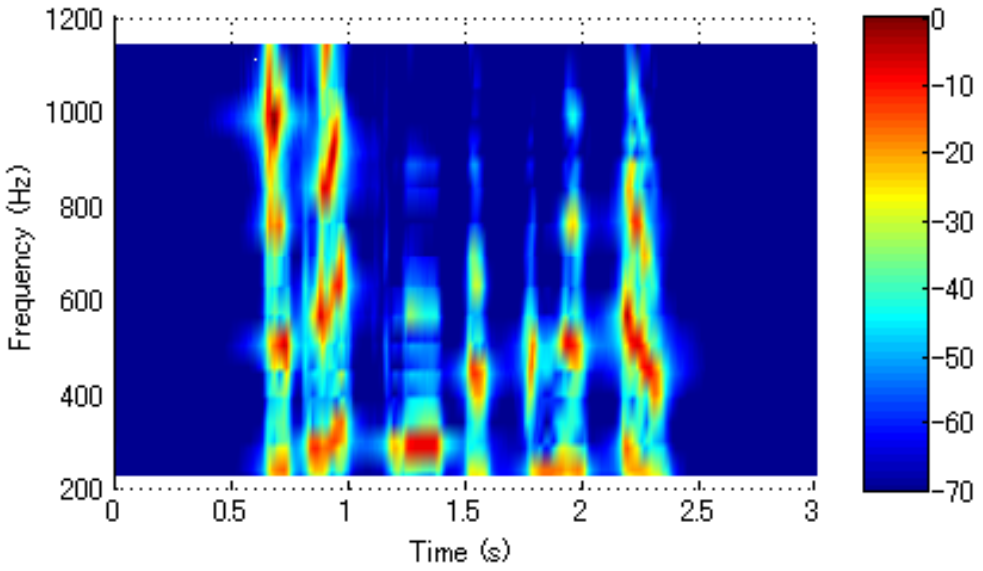


Figure 8.6: *Spectrogram obtained using the optimized filterbank.*

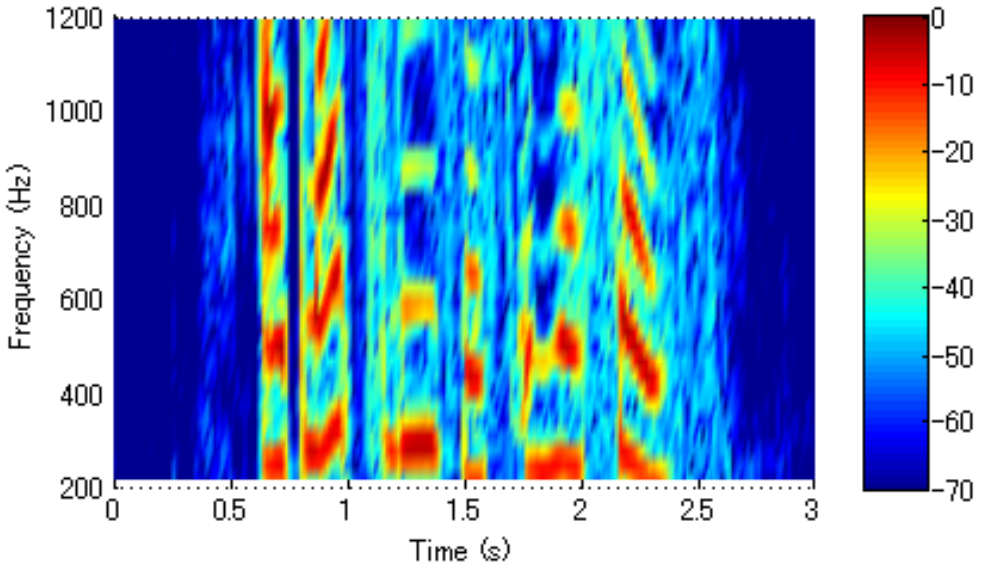


Figure 8.7: *FFT spectrogram (window size: 512, window shift: 256).*

Chapter 9

Conclusion

9.1 Summary of the thesis

The goal of this thesis was to propose a statistical approach to the analysis of natural acoustical scenes, based on models of the regularities of the acoustical environment. Our main strategy was to systematically focus on a general mathematical formulation of the problem based on an objective function, so that the various subtasks could be effectively solved as well-posed constrained optimization problems, and that our work could be easily extended or imported into other signal processing algorithms involving a statistical framework. Such a statistical approach involves solving mainly three subproblems: inference of what is happening in an acoustical scene as the best explanation of the distorted, mixed, and incomplete observations given models of the environment; reconstruction of incomplete observations based on these models; acquisition of these models from the data. We tackled all of these problems, following a common procedure: design of appropriate models and constraints; formulation of the task as an optimization problem; derivation of an effective optimization method.

We started our work by introducing in Chapter 3 a statistical model for voiced speech signals in the time-frequency power domain called Harmonic-Temporal Clustering (HTC). The time-frequency domain formulation enabled us to explicitly make use of grouping principles inspired from humans' auditory organization to derive a completely parametric model of voiced speech signals as constrained Gaussian mixture models with a smoothly evolving F_0 contour. We also introduced a broadband noise model, based on Gaussian mixture models as well, to deal with noisy environments. We explained how to formulate scene analysis tasks as the fitting of a mixture of such models to the observed spectrogram, and derived an effective method to estimate the optimal parameters, based on the EM algorithm. We showed

in Chapter 4 through experimental evaluation that our method outperforms state-of-the-art algorithms in classical scene analysis tasks such as F_0 estimation in noisy or concurrent environments, denoising, and source separation.

In Chapter 5, we explained how scene analysis based on statistical models can be extended to deal with incomplete stimuli through an auxiliary function method. Meanwhile, we studied the theoretical relation of this auxiliary function method with the EM algorithm in the particular case of Bregman divergences. We showed through experimental evaluation that the proposed method enabled to simultaneously perform the analysis of an underlying acoustical scene such as a polyphonic music signal from incomplete data, and to reconstruct its missing parts.

We then noted that discarding the phase part when working in the time-frequency magnitude domain raises several issues. First, if resynthesis is necessary, the absence of phase information needs to be dealt with by estimating the phase from the available information, i.e., the magnitude spectrogram. The estimation of a phase which corresponds well to the magnitude spectrogram is crucial to avoid very disturbing perceptual artifacts in the resynthesized signal. Second, we lose the additivity of signals, as cross-terms in the square of a sum are in general not equal to zero. Third, phase may actually be, for some classes of sounds, a relevant cue which is worth being exploited. In all cases, working in either the complex time-frequency domain or the time domain is a natural answer to deal with the problem. We presented two frameworks to do so.

The first one, which we presented in Chapter 6, is based on a careful study of the particular structure of complex STFT spectrograms. Due to the redundancy of the STFT representation, an arbitrary set of complex numbers in the complex time-frequency domain is not guaranteed to be what we call a "consistent" spectrogram, i.e., the STFT spectrogram of an actual time-domain signal. We derived a mathematical characterization of consistent spectrograms, and from it a cost function to measure the consistency of an arbitrary set of complex numbers in the complex time-frequency domain. We used this cost function to build an algorithm for phase reconstruction from magnitude spectrograms, and showed that it was both more flexible and more efficient than the state-of-the-art method. Moreover, we noted that the cost function we derived is a natural candidate to define a prior distribution on complex spectrograms, and as such likely to be used in a wide range of signal processing algorithms in the future.

The second framework, shift-invariant semi-NMF (SSNMF), was presented in Chapter 7. It is based on a direct modeling of the signal waveform, in the time domain, simply assuming

that the observed waveform is the superposition of a limited number of elementary waveforms, added with variables latencies and variable but positive amplitudes. The model is more general than the HTC model presented in earlier chapters, in the sense that it is less constrained: it can represent any kind of sound, and is not limited to harmonic ones. A sparseness prior was used on the amplitudes to ensure that the elementary waveforms capture meaningful information recurring at various time instants. We derived an optimization algorithm for this model, and showed that it could be used to effectively recover recurring templates together with their activation times from the waveform of a mixture, even in the difficult case where the various templates overlap, with examples in audio signals and extracellular recordings.

Finally, we investigated the unsupervised acquisition of models based on the data, observing that, although much can be obtained using tailored models based on prior knowledge, what we can get from them will be limited by the quality and appropriateness of that prior knowledge. We first explained how the SSNMF framework presented in Chapter 7 performs a sort of data-driven learning. Then, noting that an often overlooked but important issue when performing time-frequency analysis was to determine the analysis parameters, we considered in Chapter 8 the unsupervised learning of time-frequency analysis filterbanks. Motivated by the central role which seems to be played by modulation in auditory perception, we designed a mathematical framework to investigate the hypothesis that the human ear and brain, and in particular the peripheral system, adapted for modulation analysis through a data-driven learning process. Optimizing a filterbank on speech data under a modulation energy criterion, we showed that the optimized filterbank was close to classical ones, and the hypothesis pertinent.

9.2 Future perspectives

We tried in this thesis to tackle a wide range of aspects of the problem of acoustical scene analysis using a common statistical approach. It would however be over-ambitious to try to solve all of them completely, and, in particular due to time constraints, we had to restrain ourselves to what we believe are representative examples of each of the sub-problems involved: model design, inference, reconstruction, and learning. This implied in particular solving the subsequent questions of the design of the corresponding objective functions and of the derivation of algorithms to optimize them. We of course do not claim to have given a definitive answer to the subject, and would rather like to see our work as a set of milestones

in this young and growing field: one of our ambitions when starting this work was for instance to develop general tools which could be both used as is and later borrowed in other algorithms and frameworks, or could inspire related works. There is still much to investigate and understand. First, a few direct extensions of our work could be considered, some of which being the topic of ongoing research.

For HTC, for example, promising directions include the design of a wider range of models for other types of signals in HTC, the use of different kinds of time-frequency analysis front-ends, the development of models explicitly taking timbre into account [139], or the introduction of model selection and sparseness constraints [70], which become particularly important when dealing with polyphonic music. Experimenting with a more accurate modeling of dynamics would also be interesting, for example by looking at more complex (and more constrained) models of pitch dynamics such as the Fujisaki model for speech.

As we already noted in Chapter 6, the consistency constraints we introduced on complex STFT spectrograms can be naturally used as cost functions in the complex time-frequency domain or on phase parameters, and as such are likely to be used in a wide range of algorithms. Methods which attempt to perform some sort of separation directly in the complex time-frequency domain, such as the harmonic percussive source separation (HPSS) framework [149] or the recently introduced complex NMF framework [109], are natural candidates for its application, which we shall look into in the near future.

One of the potential developments for SSNMF is its extension to multi-channel recordings of mixed sources, introducing for example further independence assumptions on the sources. We could imagine using such a model to perform source separation and signal decomposition at the same time. Apart from audio applications, such a model could also be useful for the analysis of multi-channel extracellular recordings in which the observed signals are mixtures of several local sources composed by the same few types of spikes. It is however important to keep in mind that the non-negativity hypothesis on the amplitudes may be relevant for some signals or mixing situations, giving our algorithm an advantage over sparse coding, while for some other situations the opposite may be true.

Altogether, we believe that, along with the ever growing computational power and the now well-established theoretical grounds in both learning theory and auditory perception, the time has come for models which are both more detailed, and encompassing a broader range of phenomena. One point would be to move into the direction of a modeling which is closer and more adaptive to the data, for example by modeling directly in the complex time-frequency domain or in the time domain without relying on a wrong assumption on

the additivity of powers, or by adapting the time-frequency analysis parameters to the data and the environment. Another point would be to try to combine several tasks in a global framework, aiming at a cooperative and simultaneous solution to the various issues at stake: this is what we tried to do with the missing-data framework in Chapter 5, where the goal was to simultaneously analyze the underlying acoustical scene and reconstruct the missing data. In the same way, we could imagine using models such as SSNMF to simultaneously separate several sources and decompose each source in a few elementary signals, and even use such a model to reconstruct missing data. One could also imagine combining dereverberation and scene analysis. A particularly ambitious and, so we believe, promising program in this direction would be, in the particular case of speech, to extend the modeling scope to take into account the articulators or more generally the actual physical production process on one side, and close the loop by going as up as the recognition processes on the other side, in order to develop, through the imposition of a consistency constraint between production and perception/recognition processes, a sensory-motor approach to the acquisition of language.

Appendix A

On the relation between Poisson distribution and $\mathcal{I}$-divergence

A.1 Introduction

The possibility to interpret some distribution-fitting problems as maximum-likelihood (ML) problems, which we exploited in Chapter 3 and Chapter 5, is an important concept in signal processing and machine learning algorithms. As we briefly reviewed in Section 5.3.1, Banerjee et al. [16] showed that there is a general one-to-one correspondence between a certain type of Bregman divergences, of which the least-squares error or the $\mathcal{I}$-divergence are particular cases, and a certain type of exponential families, which are families of probability distributions including many common probability distribution families such as the gamma, Gaussian, binomial or Poisson distributions. It is well-known for example that the fitting on a domain D of a model $Q(x, \mathbf{\Theta})$ with parameters $\mathbf{\Theta}$ to observed data $W(x)$ based on the least-squares error can be interpreted as the ML estimation of the parameters $\mathbf{\Theta}$ of the model, assuming that the observation data points $W(x)$ are independently generated from a Gaussian distribution with mean $Q(x, \mathbf{\Theta})$ and fixed variance. The parameters $\mathbf{\Theta}$ minimizing the least-squares error

$$\int_D ||Q(x, \mathbf{\Theta}) - W(x)||^2 dx$$

and maximizing the log-likelihood

$$\int_D \log \Big(\frac{1}{\sqrt{2\pi}c} e^{-\frac{||Q(x,\mathbf{\Theta}) - W(x)||^2}{2c^2}} \Big) dx$$

for any positive constant c are indeed easily seen to be the same. Banerjee et al.'s result justifies a general link between distribution-fitting using a Bregman divergence and ML estimation using the corresponding exponential family. This bridge between the two theories is particularly interesting as it justifies the use in distribution-fitting problems of penalty functions on the parameters as prior functions in a MAP framework, or more generally enables the use of Bayesian techniques in such parameter estimation problems.

In the audio signal processing methods in the time-frequency magnitude domain we introduced in Chapter 3, Chapter 4, and Chapter 5, and more generally in a wide range of signal processing problems such as linear inverse problems [45], deblurring [189] or problems relying on Non-negative Matrix Factorization (NMF) techniques [118], the distributions considered are intrinsically non-negative. Csiszár showed in [53] that in such situations, the only choice of discrepancy measure consistent with certain fundamental axioms such as locality, regularity and composition-consistency is the so-called $\mathcal{I}$-divergence [52]. The question of the possibility to interpret distribution-fitting problems based on the $\mathcal{I}$-divergence as ML estimation problems is thus very important.

However, as we noted in Section 5.3.1, the relation between a given Bregman divergence and its corresponding exponential family is only useful inside the support of the carrier of the exponential family, and this support may be a strict subset of the domain of definition of the Bregman divergence. In other words, for some divergence measures, the interpretation of the distribution-fitting problem as an ML estimation one may not be straightforward. This is actually the case for the $\mathcal{I}$-divergence, which is well-defined for real values of the distributions considered, but can be shown to correspond to ML estimation based on the Poisson distribution, which is only defined on the integers. Surprisingly, to our knowledge this has never been clearly stated in the literature, the usual workaround being to quantize and scale the data to get back to the discrete Poisson distribution [45, 189].

We investigate here this mismatch between the distribution-fitting and the ML estimation problems in the particular case of the $\mathcal{I}$-divergence. We explain why the bridge between the two cannot be perfectly filled, i.e., why it is sometimes impossible to interpret in full generality a distribution-fitting problem as an ML estimation one, and we derive a theoretical workaround which gives another insight on the relation between the two theories. The goal of this chapter is simultaneously to clarify and attract the attention of the community on this easily over-looked but nonetheless important problem.

We first introduce the framework in Section A.2, then give in Section A.3 an insight on the reason for the non-existence of a normalization term giving the desired result. We finally

show in Section A.4 that the Gamma function can be used as a normalization term which asymptotically leads to the interpretation we are looking for.

A.2 Presentation of the framework

We consider a non-negative distribution W and a non-negative model $Q(\cdot, \boldsymbol{\Theta})$ parameterized by $\boldsymbol{\Theta}$, defined on a domain D. The $\mathcal{I}$-divergence, already introduced in Chapter 3, is a classical way to measure the "distance" between two such non-negative distributions:

$$\mathcal{I}(W|Q(\boldsymbol{\Theta})) \triangleq \int_D \left(W(x) \log \frac{W(x)}{Q(x;\boldsymbol{\Theta})} - \Big(W(x) - Q(x;\boldsymbol{\Theta}) \Big) \right) dx. \tag{A.1}$$

Distribution-fitting based on the $\mathcal{I}$-divergence amounts to looking for

$$\boldsymbol{\Theta}_{\text{opt}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} \mathcal{I}(W|Q(\boldsymbol{\Theta})).$$

Keeping only the terms depending on $\boldsymbol{\Theta}$ and reversing the sign of this expression, one defines the following function to maximize w.r.t. $\boldsymbol{\Theta}$:

$$\mathcal{J}(W, \boldsymbol{\Theta}) = \int_D \Big(W(x) \log Q(x;\boldsymbol{\Theta}) - Q(x;\boldsymbol{\Theta}) \Big) dx. \tag{A.2}$$

One would like to find a family of probability distributions with parameter $Q(x, \boldsymbol{\Theta})$, such that the corresponding likelihood for $\boldsymbol{\Theta}$, defined as the joint probability of all the variables $W(x)$ independently following the distribution of parameter $Q(x, \boldsymbol{\Theta})$, depends on $\boldsymbol{\Theta}$ only through $\mathcal{J}(W, \boldsymbol{\Theta})$. Remembering the case of least-squares estimation and Gaussian distributions, we would like to define the log-likelihood of $\boldsymbol{\Theta}$ as $\mathcal{J}(W, \boldsymbol{\Theta})$ itself, up to a constant which only depends on the data:

$$\log P(W|\boldsymbol{\Theta}) \triangleq \mathcal{J}(W, \boldsymbol{\Theta}) + \int_D \log f(W(x)) dx. \tag{A.3}$$

Maximization of the log-likelihood of $\boldsymbol{\Theta}$ and maximization of $\mathcal{J}(W, \boldsymbol{\Theta})$ would then be equivalent. We thus need to look for a function f such that this indeed defines a probability measure with respect to W. The point here is that, for the equality (A.3) to be useful, the function f needs to be positive on the values taken by the data, as both terms of the equality would otherwise be equal to $-\infty$, thus hiding the contribution of the $\mathcal{I}$-divergence. The corresponding distribution density on $[0, +\infty)$, with parameter θ, is then

$$\mu_{f,\theta}(z) = e^{z \log \theta - \theta} f(z) = \theta^z e^{-\theta} f(z), \forall z \in [0, +\infty), \tag{A.4}$$

which needs to be a probability distribution density for any θ:

$$\forall\theta, \int_0^{+\infty} \theta^x e^{-\theta} f(x)dx = 1. \tag{A.5}$$

A.3 Non-existence of a continuous normalization

A.3.1 Relation with the Laplace transform

We show here that there is no solution to this problem with real-valued data, in the sense that there indeed exists a unique non-negative measure ν on $\mathbb{R}^+$ such that

$$\forall\theta, \int_0^{+\infty} \theta^x e^{-\theta} \nu(dx) = 1, \tag{A.6}$$

but it is supported by $\mathbb{N} = \{0, 1, 2, \dots\}$. This measure leads to none other than the discrete Poisson distribution. In the following, we thus look for a non-negative measure ν satisfying Eq. (A.6).

If we rewrite Eq. (A.6) with $\mu = \log\theta$, our problem amounts to looking for ν such that

$$\forall\mu, \int_0^{+\infty} e^{x\mu} d\nu(x) = e^{e^\mu}, \tag{A.7}$$

i.e., to looking for a measure ν whose Laplace transform is $\mu \mapsto e^{e^\mu}$.

A direct computation gives the Laplace transform of the Poisson distribution $p(\cdot,\theta) = \sum_{k\in\mathbb{N}} \frac{\theta^k e^{-\theta}}{k!}\delta_k(\cdot)$ of parameter θ,

$$\int_0^{+\infty} e^{x\mu} p(x,\theta)dx = e^{-\theta}\sum_{k\in\mathbb{N}} \frac{e^{kt}\theta^k}{k!} = e^{-\theta}e^{\theta e^t}. \tag{A.8}$$

Up to the constant factor e^{-1}, the Poisson distribution with mean parameter 1 is thus a solution to (A.6), and conversely any solution to (A.6) must have (up to a constant factor) the same Laplace transform as the Poisson distribution of mean 1. But this Laplace transform is holomorphic in a neighborhood of 0 (it is in fact an entire function, holomorphic on the whole complex plane), and thus, as shown in [30] (Chap. 30), determines the measure it is associated with. In other words, a measure with such a Laplace transform is unique, which shows that the unique probability distribution satisfying Eq. (A.6) is $\nu = e\,p(\cdot,1)$, leading for $\mu_{\nu,\theta}$ to none other than the classical Poisson distribution family $p(\cdot,\theta)$, which is supported by $\mathbb{N}$.

A.3.2 Consequences on the interpretation of $\mathcal{I}$-divergence-based fitting as an ML estimation problem

As there is no function taking positive values on $[0; +\infty)$ which verifies (A.5), we cannot directly interpret $\mathcal{I}$-divergence-based distribution-fitting problems with real-valued data as ML estimation problems. If nothing is done, this means that all non-integer data points will have zero likelihood, even if the model fits them perfectly. The usual workaround [45, 189] is to quantize and perform a suitable scaling of the data and the model so as to act as if the data were integer, justified by the fact that computers quantize the data anyway. Quantization is a practical justification, but it may seem rather disappointing and inelegant, as it looses the continuous quality of the problem. We derive in the following section a theoretical justification which retains more of the continuity of the problem on real-valued data.

A.4 Asymptotically satisfactory normalization using the Gamma function

By analogy with the discrete case, for which, for any $\theta \in \mathbb{R}^+$, $\frac{\theta^n e^{-\theta}}{\Gamma(1+n)}$ is a probability density distribution on $n \in \mathbb{N}$ called the Poisson distribution, we consider the distribution of the variable $x \in [0, +\infty)$ with parameter θ,

$$f_\theta(x) = \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}, \tag{A.9}$$

where Γ is the Gamma function. Note that if we reverse the roles of x and θ, it is nothing else than the Gamma distribution.

This distribution is unfortunately not a probability distribution (it could not be, as shown in Section A.3), and needs a normalizing constant. Let us consider the evolution of this normalizing constant with respect to the parameter θ. We denote by

$$g(\theta) = \int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx \tag{A.10}$$

the mass of the distribution f_θ and by

$$h(\theta) = \int_0^{+\infty} \frac{x\theta^x e^{-\theta}}{\Gamma(1+x)} dx \tag{A.11}$$

its first-order moment.

A.4.1 Limit of g at the origin

We notice that $\forall\theta < 1, \forall\eta > 0$,

$$\int_0^\eta \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}dx \leq \int_0^\eta \frac{1}{\Gamma(1+x)}dx \underset{\eta\to 0}{\longrightarrow} 0,$$

and $\forall\theta < 1, \forall\eta > 0$,

$$\int_\eta^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}dx \leq \theta^\eta \int_0^{+\infty} \frac{e^{-\theta}}{\Gamma(1+x)}dx \leq C\theta^\eta \underset{\theta\to 0}{\longrightarrow} 0.$$

Thus, for any $\epsilon > 0$, by choosing η_0 such that

$$\forall\eta < \eta_0, \int_0^\eta \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}dx \leq \epsilon$$

and then θ_0 such that

$$\forall\theta < \theta_0, \int_{\eta_0}^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)}dx \leq \epsilon,$$

we show that

$$g(\theta) \underset{\theta\to 0}{\longrightarrow} 0. \tag{A.12}$$

A.4.2 Rewriting $g(\theta)$

As we can write

$$\forall\zeta > 0, \forall\theta > 0,\ g(\theta) = g(\zeta) + \int_\zeta^\theta g'(t)dt, \tag{A.13}$$

we look more closely at the derivative of g:

$$\begin{aligned} g'(t) &= \int_0^{+\infty} \frac{xt^{x-1}e^{-t} - t^x e^{-t}}{\Gamma(1+x)}dx \\ &= \int_0^{+\infty} \frac{xt^{x-1}e^{-t}}{\Gamma(1+x)}dx - \int_0^{+\infty} \frac{t^x e^{-t}}{\Gamma(1+x)}dx \\ &= \int_{-1}^{+\infty} \frac{t^u e^{-t}}{\Gamma(1+u)}du - g(t) \\ &= \int_{-1}^{0} \frac{t^u e^{-t}}{\Gamma(1+u)}du. \end{aligned} \tag{A.14}$$

By using the definition of the Gamma function

$$\Gamma(z) = \int_0^{+\infty} e^{-t}t^{z-1}dt$$

and Euler's reflection formula

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin\pi z},$$

we can perform the following derivation:

$$\begin{aligned} g'(t) &= \int_0^1 \frac{t^{-u}e^{-t}}{\Gamma(1-u)}du \\ &= \int_0^1 \frac{\sin \pi u}{\pi} t^{-u}e^{-t}\Gamma(u)du \\ &= \int_0^1 \frac{\sin \pi u}{\pi} t^{-u}e^{-t}\int_0^{+\infty} e^{-s}s^{u-1}ds\,du \\ &= \int_0^{+\infty} \frac{e^{-s}e^{-t}}{\pi s}\left(\int_0^1 \left(\frac{s}{t}\right)^u \sin(\pi u)du\right)ds. \end{aligned} \tag{A.15}$$

The inside integral can be analytically computed. If we note $\alpha = \log \frac{s}{t}$, then

$$\begin{aligned} \int_0^1 \left(\frac{s}{t}\right)^u \sin(\pi u)du &= \int_0^1 \frac{e^{(\alpha+i\pi)u} - e^{(\alpha-i\pi)u}}{2i}du \\ &= -\frac{1}{2i}\left(\frac{1}{\alpha+i\pi} - \frac{1}{\alpha-i\pi}\right)(1+e^{\alpha}) \\ &= \pi\frac{1+e^{\alpha}}{\pi^2+\alpha^2} \\ &= \pi\frac{1+\frac{s}{t}}{\pi^2+(\log\frac{s}{t})^2}. \end{aligned} \tag{A.16}$$

We get

$$g'(t) = \int_0^{+\infty} e^{-s}e^{-t}\frac{\frac{1}{s}+\frac{1}{t}}{\pi^2+(\log t - \log s)^2}ds. \tag{A.17}$$

Altogether, for $\theta > 0$ and $\zeta > 0$,

$$g(\theta) = g(\zeta) + \int_\zeta^\theta \int_0^{+\infty} e^{-t}e^{-s}\frac{\frac{1}{t}+\frac{1}{s}}{\pi^2+(\log t - \log s)^2}ds\,dt. \tag{A.18}$$

By letting ζ go to 0 in this last expression, we deduce from Eq. (A.12) another expression for $g(\theta)$:

$$g(\theta) = \int_0^\theta \int_0^{+\infty} e^{-t}e^{-s}\frac{\frac{1}{t}+\frac{1}{s}}{\pi^2+(\log t - \log s)^2}ds\,dt. \tag{A.19}$$

We can further simplify Eq. (A.19). We perform a change of variables $u = \log s/t$ for t fixed,

$$\begin{aligned} g(\theta) &= \int_0^\theta \int_{-\infty}^{+\infty} e^{-te^u}e^{-t}\frac{1+e^u}{\pi^2+u^2}du\,dt \\ &= \int_{-\infty}^{+\infty} \frac{1}{\pi^2+u^2}\int_0^\theta (1+e^u)e^{-(1+e^u)t}dt\,du \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} \frac{1 - e^{-(1+e^u)\theta}}{\pi^2 + u^2} du \\
&= \int_{-\infty}^{+\infty} \frac{1}{\pi^2 + u^2} du - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du \\
&= \frac{1}{\pi} \left[\arctan\left(\frac{x}{\pi}\right) - \arctan\left(\frac{x}{\pi}\right) \right]_{-\infty}^{+\infty} - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du \\
&= 1 - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du. \qquad \text{(A.20)}
\end{aligned}
$$

We have thus proven the following

Proposition A.1. *For all $\theta \in \mathbb{R}^+$, the following equality holds true:*

$$\int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx = 1 - e^{-\theta} \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du. \qquad \text{(A.21)}$$

Equivalently, the Laplace transform of the function $x \mapsto \frac{1}{\Gamma(1+x)}$ can be alternatively expressed as

$$\int_0^{+\infty} e^{xs} \frac{1}{\Gamma(1+x)} dx = e^{e^s} - \int_{-\infty}^{+\infty} \frac{e^{-e^{s+u}}}{\pi^2 + u^2} du. \qquad \text{(A.22)}$$

The second part of the proposition is obtained from Equation (A.21) by writing $s = \log \theta$.

A.4.3 Asymptotic behavior

We can easily see that $g(\theta)$ is concave and increasing by looking at its derivatives. We see from Eq. (A.21) that g is bounded by 1. It thus converges to a finite value.

By noticing that

$$\forall \theta > 0, \forall u \in \mathbb{R},\ 0 \leq \int_{-\infty}^{+\infty} \frac{e^{-\theta e^u}}{\pi^2 + u^2} du \leq 1,$$

we can conclude from Eq. (A.21) that

$$\lim_{\theta \to +\infty} \int_0^{+\infty} \frac{\theta^x e^{-\theta}}{\Gamma(1+x)} dx = 1. \qquad \text{(A.23)}$$

The normalization factor of f_θ thus converges to 1.

We also notice that $h(\theta) = \theta(g(\theta) + g'(\theta))$, from which we deduce by a similar analysis that

$$\lim_{\theta \to +\infty} h(\theta) - \theta = 0. \qquad \text{(A.24)}$$

So, asymptotically, f_θ behaves in the same way as the Poisson distribution, i.e., its mass converges to 1 and its first moment is asymptotically equal to its mean parameter.

A.4.4 Justifying again the cross-interpretation

As evoked in Section A.3.2, several authors present their work with the $\mathcal{I}$-divergence directly on discretized data, enabling them to fall back to the discrete Poisson distribution after proper scaling of the data and the model. The above results on the normalization factor and the mean can be considered as a different way to justify the bridge between $\mathcal{I}$-divergence-based fitting and ML estimation for real-valued data without quantization.

For sufficiently large values of the model, the distribution is indeed almost a probability distribution which behaves like the discrete Poisson distribution. If one can ensure that the values taken by the model will be bounded from below by a positive value, then by rescaling both the model and the data by multiplying them by a large constant, the continuous distribution $f_{Q(x,\boldsymbol{\Theta})}$ parameterized by the model $Q(x, \boldsymbol{\Theta})$ can be made as close to a probability distribution as desired for all the values taken by the model. Meanwhile, the optimal parameters $\boldsymbol{\Theta}$ are not changed by the scaling operation, as the log-likelihoods before and after scaling are equal up to scaling and addition of a constant which does not depend on the parameters:

$$\int (\alpha W \log \alpha Q(\boldsymbol{\Theta}) - \alpha Q(\boldsymbol{\Theta})) = \alpha \int (W \log Q(\boldsymbol{\Theta}) - Q(\boldsymbol{\Theta})) + C \tag{A.25}$$

where $\alpha > 0$ is the scaling parameter.

One can ensure that the model is bounded from below by a positive value for example if the data are themselves bounded from below by such a value and if the model is well-designed, such that it should not take infinitely small values if the data to fit is large enough. One can also add to both the model and the data a small value $\epsilon > 0$. The optimal parameters for this shifted problem can be made as close as desired to the optimal parameters for the original problem by choosing ϵ small enough, while, for ϵ fixed, the shifted problem can be made as close to a righteous ML problem as desired through scaling.

A.5 Conclusion

We presented the inherent mismatch occurring in the interpretation of some distribution-fitting problems as ML estimation ones, focusing on the example of the $\mathcal{I}$-divergence. We gave insights on the reason why distribution-fitting based on the $\mathcal{I}$-divergence on real-valued data cannot be seen directly as an ML estimation problem, and derived a theoretical workaround to this issue using the Gamma function, justifying in another way the MAP estimation as performed in Chapter 3 and Chapter 5.

Appendix B

List of Publications

Journal Papers

[J1] Erik McDermott, Timothy J. Hazen, Jonathan Le Roux, Atsushi Nakamura and Shigeru Katagiri, "Discriminative Training for Large Vocabulary Speech Recognition Using Minimum Classification Error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203–223, Jan. 2007.

[J2] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Single and Multiple F0 Contour Estimation Through Parametric Spectrogram Modeling of Speech in Noisy Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1135–1145, May 2007.

Peer-Reviewed International Conferences

[I1] Jonathan Le Roux and Erik McDermott, "Optimization Methods for Discriminative Training," in *Proceedings of the Interspeech 2005 ISCA European Conference on Speech Communication and Technology (Eurospeech)*, pp. 3341–3344, Sep. 2005.

[I2] Hirokazu Kameoka, Jonathan Le Roux, Nobutaka Ono and Shigeki Sagayama, "Speech Analyzer Using a Joint Estimation Model of Spectral Envelope and Fine Structure," in *Proceedings of the Interspeech 2006 ISCA International Conference on Spoken Language Processing (ICSLP)*, pp. 2502–2505, Sep. 2006.

[I3] Alain de Cheveigné, Jonathan Le Roux and Jonathan Z. Simon, "MEG Signal Denoising Based on Time-Shift PCA," in *Proceedings of the ICASSP 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, pp. 317–320, Apr. 2007.

[I4] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Harmonic-Temporal Clustering of Speech for Single and Multiple F0 Contour Estimation in Noisy Environments," in *Proceedings of the ICASSP 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. IV, pp. 1053–1056, Apr. 2007.

[I5] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Single Channel Speech and Background Segregation Through Harmonic-Temporal Clustering," in *Proceedings of the WASPAA 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 279-282, Oct. 2007.

[I6] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama and Alain de Cheveigné, "Modulation Analysis of Speech Through Orthogonal FIR Filterbank Optimization," in *Proceedings of the ICASSP 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4189–4192, Apr. 2008.

[I7] Nobutaka Ono, Ken-ichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka and Shigeki Sagayama, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," in *Proceedings of the EUSIPCO 2008 European Signal Processing Conference*, Aug. 2008.

[I8] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Computational Auditory Induction by Missing-Data Non-Negative Matrix Factorization," in *Proceedings of the SAPA 2008 ISCA Workshop on Statistical and Perceptual Audition*, Sep. 2008.

[I9] Jonathan Le Roux, Nobutaka Ono and Shigeki Sagayama, "Explicit Consistency Constraints for STFT Spectrograms and Their Application to Phase Reconstruction," in *Proceedings of the SAPA 2008 ISCA Workshop on Statistical and Perceptual Audition*, Sep. 2008.

[I10] Jonathan Le Roux, Alain de Cheveigné and Lucas C. Parra, "Adaptive Template Matching with Shift-Invariant Semi-NMF," in *Advances in Neural Information Processing Systems 21 (Proc. NIPS*2008)*, D. Koller, Y. Bengio, D. Schuurmans, and L. Bottou, Eds. Cambridge, MA: The MIT Press, 2009.

Other International Conferences

[C1] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono and Shigeki Sagayama, "Parametric Spectrogram Modeling of Single and Concurrent Speech with Spline Pitch Contour," in *Proceedings of the 4th Joint meeting of the Acoustical Society of America and the Acoustical Society of Japan*, Dec. 2006.

[C2] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Music and Speech Signal Processing Using Harmonic-Temporal Clustering," Invited paper at *Acoustics'08 Paris*, Jul. 2008.

Domestic (Japanese) Conferences

[D1] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono and Shigeki Sagayama, "Harmonic Temporal Clustering of Speech Spectrum," in *Proceedings of the Acoustical Society of Japan Spring Meeting*, 2-11-3, pp. 307–308, Mar. 2006.

[D2] Hirokazu Kameoka, Jonathan Le Roux, Nobutaka Ono and Shigeki Sagayama, "Harmonic Temporal Structured Clustering: A New Approach to CASA," in *Proceedings of the Technical Committee of Psychological and Physiological Acoustics of the Acoustical Society of Japan*, vol. 36, no. 7, H-2006-103, pp. 575–580, Oct. 2006. *(in Japanese)*

[D3] Ken-ichi Miyamoto, Mari Tatezono, Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono and Shigeki Sagayama, "Separation of Harmonic and Non-Harmonic Sounds Based on 2D-Filtering of the Spectrogram," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, 1-1-7, Sep. 2007. *(in Japanese)*

[D4] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Monaural Speech Separation Through Harmonic-Temporal Clustering of the Power Spectrum," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, 3-4-3, Sep. 2007.

[D5] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama and Alain de Cheveigné, "Filterbank Optimization for Amplitude Modulation Analysis of Audio Signals," in *Proceedings of the Acoustical Society of Japan Autumn Meeting*, 1-2-1, Sep. 2007.

[D6] Yu Mizuno, Jonathan Le Roux, Nobutaka Ono and Shigeki Sagayama, "Real-Time Time-Scale/Pitch Modification of Music Signal by Stretching Power Spectrogram and

Consistent Phase Reconstruction", in *Proceedings of the Acoustical Society of Japan Spring Meeting*, 2-8-4, Mar. 2009. *(in Japanese)*

Technical Reports

[T1] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono and Shigeki Sagayama, "On the Interpretation of $\mathcal{I}$-Divergence-Based Distribution Fitting as a Maximum-Likelihood Estimation Problem," Technical Report of the University of Tokyo, METR 2008-11, Mar. 2008.

Patents

[P1] Shigeki Sagayama, Nobutaka Ono, Hirokazu Kameoka, Ken-Ichi Miyamoto and Jonathan Le Roux, "Method for Harmonic/Percussive Signal Separation," Japanese Patent Application No. 2008-054826.

Bibliography

[1] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 179–196, Jan. 2006.

[2] M. Abe and S. Ando, "Nonlinear time-frequency domain operators for decomposing sounds into loudness, pitch and timbre," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1995, pp. 1368–1371.

[3] ——, "Application of loudness/pitch/timbre decomposition operators to auditory scene analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1996, pp. 2646–2649.

[4] ——, "Computational auditory scene analysis based on loudness/pitch/timbre decomposition," in *Proceedings of the IJCAI Workshop on Computational Auditory Scene Analysis (CASA97)*, 1997, pp. 47–54.

[5] ——, "Auditory scene analysis based on time-frequency integration of shared FM and AM (I): Lagrange differential features and frequency-axis integration," *Systems and Computers in Japan*, vol. 33, no. 11, pp. 95–106, Oct. 2002.

[6] ——, "Auditory scene analysis based on time-frequency integration of shared FM and AM (II): Optimum time-domain integration and stream sound reconstruction," *Systems and Computers in Japan*, vol. 33, no. 10, pp. 83–94, Sep. 2002.

[7] K. Achan, S. Roweis, and B. Frey, "Probabilistic inference of speech signals from phaseless spectrograms," in *Advances in Neural Information Processing Systems 16 (Proc. NIPS*2003)*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: The MIT Press, 2004, pp. 1393–1400.

[8] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segment-based probabilistic generative model of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Mar. 2005, pp. 221–224.

[9] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[10] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[11] S. Ando, "Consistent gradient operators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 252–265, Mar. 2000.

[12] S. Arkadiusz, M. Sayles, and I. M. Winter, "Spike waveforms in the anteroventral cochlear nucleus revisited," in *ARO midwinter meeting*, no. Abstract #678, 2008.

[13] H. Asari, B. A. Pearlmutter, and A. M. Zador, "Sparse representations for the cocktail party problem," *The Journal of Neuroscience*, vol. 26, no. 28, pp. 7477–7490, Jul. 2006.

[14] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proceedings of the 6th International Congress on Acoustics*, 1968.

[15] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Sep. 1993, pp. 1003–1006.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[17] J. P. Barker, "Robust automatic speech recognition," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D.-L. Wang and G. J. Brown, Eds. IEEE Press/Wiley, 2006.

[18] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, Jan. 2005.

[19] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed. Cambridge, MA: The MIT Press, 1961, pp. 217–234.

[20] ——, "Redundancy reduction revisited," *Network: Computation in Neural Systems*, vol. 12, no. 3, pp. 241–253, Mar. 2001.

[21] ——, "The exploitation of regularities in the environment by the brain," *Behavioral and Brain Sciences*, vol. 24, no. 4, pp. 602–607, Aug. 2001.

[22] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[23] J. R. Bellegarda, *Latent Semantic Mapping: Principles and Applications*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, 2007.

[24] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[25] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[26] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *SIGGRAPH 2000, Computer Graphics Proc.*, 2000, pp. 417–424.

[27] A. Bertrand, K. Demuynck, V. Stouten, and H. V. hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorisation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 4713–4716.

[28] A. Biem, S. Katagiri, and B.-H. Juang, "Discriminative feature extraction for speech recognition," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, vol. 2, Apr. 1993, pp. 275–278.

[29] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An application of discriminative feature extraction to filter-bank-based speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, p. 2001, Feb. 2001.

[30] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[31] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, 1993, pp. 97–110.

[32] P. Boersma and D. Weenin, "Praat system," http://www.fon.hum.uva.nl/praat/.

[33] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[34] A. S. Bregman, *Auditory Scene Analysis*. The MIT Press, 1990.

[35] G. J. Brown, "Computational auditory scene analysis: A representational approach." Ph.D. dissertation, University of Sheffield, 1992.

[36] G. J. Brown and D.-L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York, NY: Springer, 2005, pp. 371–.402.

[37] A. T. Cemgil, "Bayesian inference in non-negative matrix factorization models," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR.609, Jul. 2008.

[38] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2005.

[39] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2008, ch. 43, pp. 843–871.

[40] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[41] A. de Cheveigné, "The auditory system as a separation machine," in *Proceedings of the International Symposium on Hearing*, 2000.

[42] ——, "The cancellation principle in acoustic scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer Academic, 2004, pp. 245–259.

[43] ——, "Multiple F_0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D.-L. Wang and G. J. Brown, Eds. IEEE Press/Wiley, 2006.

[44] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, 2002.

[45] K. Choi, "Minimum I-divergence methods for inverse problems," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[46] A. Christophe, A. Gout, S. Peperkamp, and J. Morgan, "Discovering words in the continuous speech stream: The role of prosody," *Journal of Phonetics*, vol. 31, pp. 585–598, 2003.

[47] P. Clark and L. Atlas, "Modulation decompositions for the interpolation of long gaps in acoustic signals," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 3741–3744.

[48] M. P. Cooke, "Modeling auditory processing and organisation," Ph.D. dissertation, University of Sheffield, 1993.

[49] M. P. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[50] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[51] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[52] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[53] ——, "Why least squares and maximum entropy? - an axiomatix approach to inverse problems," *The Annals of Statistics*, vol. 19, pp. 2033–2066, 1991.

[54] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, vol. 102, pp. 2892–2905, 1997.

[55] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1808–1816, Sep. 2006.

[56] ——, "Représentations parcimonieuses des sons musicaux: modèles, algorithmes et applications," Mémoire d'Habilitation à Diriger des Recherches, Université Paris VI – Pierre et Marie Curie, 2008, in French.

[57] L. Daudet and B. Torrésani, "Sparse adaptive representations for musical signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. Springer, 2006, ch. 3, pp. 65–98.

[58] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive Approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[59] P. B. Denes and E. N. Pinson, *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd ed. New York, NY: W. H. Freeman & Co., 1993.

[60] L. Deng, *Dynamic Speech Models, Theory, Algorithms, and Applications*. Morgan & Claypool, 2006.

[61] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in Neural Information Processing Systems 18 (Proc. NIPS*2005)*, Y. Weissa, B. Schölkopf, , and J. Platt, Eds. Cambridge, MA: The MIT Press, 2006.

[62] U. Dicke, S. Ewert, T. Dau, and B. Kollmeier, "A neural circuit transforming temporal periodicity information into a rate-based representation in the mammalian auditory system," *Journal of the Acoustical Society of America*, vol. 121, pp. 310–326, 2007.

[63] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorization for clustering and low-dimension representation," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-60428, 2006.

[64] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[65] S. C. Douglas, S.-I. Amari, and S.-Y. Kung, "Gradient adaptation with unit-norm constraints," Southern Methodist University, Tech. Rep. EE-99-003, 1999.

[66] ——, "Gradient adaptive paraunitary filter banks for spatio-temporal subspace analysis and multichannel blind deconvolution," *Journal of VLSI Signal Processing Systems*, vol. 37, no. 2-3, pp. 247–261, 2004.

[67] B. Doval, "Estimation de la fréquence fondamentale des signaux sonores," Ph.D. dissertation, Université Pierre et Marie Curie, 1994, in French.

[68] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: Wiley, 2001.

[69] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 169–172.

[70] K. Egashira, K. Miyamoto, N. Ono, and S. Sagayama, "Unsupervised adjustment of the number of clusters in harmonic-temporal clustering for multipitch estimation," in *Proceedings of the Acoustical Society of Japan Spring Meeting*, Mar. 2008, pp. 899–900, in Japanese.

[71] M. Elhilali, J. B. Fritz, J. Z. Simon, and S. A. Shamma, "Dynamics of precise spike timing in primary auditory cortex," *The Journal of Neuroscience*, vol. 24, no. 5, pp. 1159–1172, 2004.

[72] D. P. W. Ellis, "Hierarchic models of sound for separation and restoration," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1993.

[73] ——, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Jun. 1996.

[74] ——, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D.-L. Wang and G. J. Brown, Eds. IEEE Press/Wiley, 2006.

[75] J. S. Erkelens, "Autoregressive modelling for speech coding: Estimation, interpolation and quantisation," Ph.D. dissertation, Delft University of Technology, 1996.

[76] P. A. A. Esquef and L. W. P. Biscainho, "An efficient model-based multirate method for reconstruction of audio signals across long gaps," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1391–1400, july 2006.

[77] K. Fang, S. Kotz, and K. Ng, *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall, 1990.

[78] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[79] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," TELECOM ParisTech, Tech. Rep. 2008D006, May 2008.

[80] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 174–185, Jan. 2008.

[81] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 438–449, Feb. 2002.

[82] H. Fujisaki and S. Nagashima, "A model for synthesis of pitch contours of connected speech," *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[83] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed. Marcel Dekker, 2001.

[84] S. J. Godsill, A. Doucet, and M. West, "Maximum a posteriori sequence estimation using Monte Carlo particle filters," *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 1, pp. 82–96, 2001.

[85] ——, "Monte Carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 156–168, Mar. 2004.

[86] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1998.

[87] M. Goto, "Development of the RWC Music Database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, vol. I, 2004, pp. 553–556.

[88] S. Greenberg and E. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Apr. 1997, pp. 1647–1650.

[89] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[90] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, 2007.

[91] Y. H. Gu and W. M. G. van Bokhoven, "Co-channel speech separation using frequency bin nonlinear adaptive filter," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1991, pp. 949–952.

[92] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1990, pp. 361–364.

[93] H. von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. (Second English edition, translated by A. J. Ellis, 1885. First German edition: 1863). Reprinted by Dover Publications, 1954.

[94] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," IDIAP, Tech. Rep. IDIAP-RR 03-50, 2003.

[95] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proceedings of the European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Sep. 2005.

[96] H. Hermansky and N. Malayath, "Spectral basis functions from discrimant analysis," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Dec. 1998.

[97] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, pp. 257–264, 1988.

[98] W. J. Hess, *Pitch Determination of Speech Signals*. New York: Springer, 1983.

[99] G. Hu and D.-L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[100] ——, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[101] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[102] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY: Wiley, 2001.

[103] F. Itakura and S. Saito, "Analysis-synthesis telephony based on the maximum-likelihood method," in *Proceedings of the 6th International Congress on Acoustics*, 1968.

[104] A. J. E. M. Janssen, R. Veldhuis, and L. B. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as AR processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, Apr. 1986.

[105] P. X. Joris, L. H. Carney, P. H. Smith, and T. C. T. Yin, "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency," *Journal of Neurophysiology*, vol. 71, pp. 1022–1036, 1994.

[106] H. Kameoka, "Statistical approach to multipitch analysis," Ph.D. dissertation, The University of Tokyo, 2007.

[107] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2004, pp. 297–300.

[108] ——, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[109] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2009.

[110] T. Karrer, E. Lee, and J. Borchers, "PhaVoRIT: A phase vocoder for real-time interactive time-stretching," in *Proceedings of the International Computer Music Conference (ICMC)*, Nov. 2006, pp. 708–715.

[111] M. Kashino, "Phonemic restoration: The brain creates missing speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 318–321, 2006.

[112] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and

periodicity," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, vol. 6, Sep. 1999, pp. 2781–2784.

[113] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, Feb. 2007.

[114] A. P. Klapuri and M. Davy, *Signal processing methods for music transcription.* New York: Springer, 2005.

[115] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets," *Nonlinear Processes in Geophysics*, vol. 13, pp. 151–159, 2006.

[116] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[117] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech'2004 - ICSLP)*, Oct. 2004.

[118] D. D. Lee and H. S. Seung, "Learning of the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[119] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13 (Proc. NIPS*2000).* Cambridge, MA: The MIT Press, 2001, pp. 556–562.

[120] P. Leveau, "Décompositions parcimonieuses structurées: Application à la représentation objet de la musique," Ph.D. dissertation, Université Pierre et Marie Curie - GET-Télécom Paris, 2007, in French.

[121] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[122] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2006, pp. 362–371.

[123] J. C. R. Licklider, "A duplex theory of pitch perception," *Experimentia*, vol. 7, pp. 128–133, 1951.

[124] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[125] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18 866–18 869, 2006.

[126] L. Lu, Y. Mao, W. Liu, and H.-J. Zhang, "Audio restoration by constrained audio texture synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Apr. 2003, pp. 636–639.

[127] D. Luenberger, *Optimization by Vector Space Methods*. Wiley, 1969.

[128] R. F. Lyon, "A computational model of binaural localization and separation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1983, pp. 1148–1151.

[129] ——, "Computational models of neural auditory processing," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 1984, pp. 41–44.

[130] R. C. Maher, "A method for extrapolation of missing digital audio data," in *95th AES Convention*, New York, 1993.

[131] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 21, no. 3, pp. 140–148, 1973.

[132] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, Sep. 2007.

[133] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionnaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[134] D. Marr, *Vision*. San Francisco, CA: W. H. Freeman, 1982.

[135] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[136] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[137] E. McDermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE Transactions, Special Issue on Statistical Modeling for Speech Processing*, vol. E89D, no. 3, pp. 1006–1014, Mar. 2006.

[138] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[139] K. Miyamoto, H. Kameoka, T. Nishimoto, N. Ono, and S. Sagayama, "Harmonic-temporal-timbral clustering (HTTC) for the analysis of multi-instrument polyphonic music signals," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 113–116.

[140] M. Mørup, K. H. Madsen, and L. K. Hansen, "Shifted non-negative matrix factorization," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Aug. 2007, pp. 139–144.

[141] M. Mørup and M. N. Schmidt, "Sparse non-negative matrix factor 2-D deconvolution," Technical University of Denmark, Tech. Rep., 2006.

[142] M. Mørup, M. N. Schmidt, and L. K. Hansen, "Shift invariant sparse coding of image and music data," Technical University of Denmark, Tech. Rep. IMM2008-04659, 2008.

[143] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175–206, 1995.

[144] T. Nakatani, M. Goto, and H. G. Okuno, "Localization by harmonic structure and its application to harmonic sound segregation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1996, pp. 653–656.

[145] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[146] ——, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, pp. 481–487, 2004.

[147] N. Ono and S. Ando, "A theory of filter banks which represent signals as holomorphic functions on the time-frequency domain," in *Proceedings of the SICE annual conference*, Aug. 2002, pp. 1622–1625.

[148] ——, "Signal analysis/synthesis based on zeros in holomorphic time-frequency plane," in *Proceedings of the SICE annual conference*, Aug. 2003, pp. 2543–2546.

[149] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Sep. 2008, pp. 139–144.

[150] J. K. O'Reagan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 883–917, 2001.

[151] J. J. K. Ó Ruanaidh and W. J. Fitzgerald, "Interpolation of missing samples for audio restoration," *Electronics Letters*, vol. 30, no. 8, pp. 622–623, 1994.

[152] P. Paches-Leal, R. C. Rose, and C. Nadeu, "Optimization algorithms for estimating modulation spectrum domain filters," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Sep. 1999.

[153] S. E. Palmer, *Vision Science*. Cambridge, MA: The MIT Press, 1999.

[154] D. Philipona, J. K. O'Reagan, and J.-P. Nadal, "Is there something out there? inferring space from sensorimotor dependencies," *Neural Computation*, vol. 15, pp. 2029–2049, 2003.

[155] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Processing*, vol. 86, no. 3, pp. 417–431, Mar. 2006.

[156] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 2002.

[157] M. S. Puckette, "Phase-locked vocoder," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1995.

[158] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[159] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, Sep. 2005.

[160] J. J. Rajan, P. J. W. Rayner, and S. J. Godsill, "A Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *IEE Proceedings on Vision, Image, and Signal Processing*, vol. 144, no. 4, pp. 249–256, Aug. 1997.

[161] M. Rangaswamy, D. Weiner, and A. Oeztuerk, "Non-Gaussian random vector identification using spherically invariant random processes," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 29, no. 1, pp. 111–123, Jan. 1993.

[162] C. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, Jan. 2006.

[163] E. Ravelli, G. Richard, and L. Daudet, "Matching pursuit in adaptive dictionaries for scalable audio coding," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2008.

[164] P. J. W. Rayner and S. J. Godsill, "The detection and correction of artefacts in archived grammophone recordings," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1991.

[165] B. H. Repp, "Perceptual restoration of a "missing" speech sound: Auditory induction or illusion?" *Perception & Psychophysics*, vol. 51, no. 1, pp. 14–32, 1992.

[166] M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis, "Towards single-channel unsupervised source separation of speech mixtures: the layered harmonics/formants separation-tracking model," in *Proceedings of the ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Oct. 2004, pp. 25–30.

[167] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 2000.

[168] S. Saito, "Music signal processing by specmurt analysis," Master's thesis, University of Tokyo, Jan. 2007, in Japanese.

[169] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[170] P. Sajda, S. Du, and L. C. Parra, "Recovery of constituent spectra using non-negative matrix factorization," in *Proceedings of SPIE Wavelets X*, Aug. 2003, pp. 321–331.

[171] J. Särelä and H. Valpola, "Denoising source separation," *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.

[172] L. K. Saul, F. Sha, and D. D. Lee, "Statistical signal processing with nonnegativity constraints," in *Proceedings of the European Conference on Speech Communication and Technology (Interspeech'2003 - Eurospeech)*, Sep. 2003, pp. 1001–1004.

[173] S. M. Schimmel, "Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices," Ph.D. dissertation, University of Washington, 2007.

[174] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. IV, Apr. 2007, pp. 605–608.

[175] M. N. Schmidt and M. Mørup, "Sparse non-negative matrix factor 2-D deconvolution for automatic transcription of polyphonic music," Technical University of Denmark, Tech. Rep., 2005.

[176] ——, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, Apr. 2006, pp. 700–707.

[177] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 1985, pp. 937–940.

[178] F. Sha and L. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Advances in Neural Information Processing Systems 17 (Proc. NIPS*2004)*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: The MIT Press, 2005, pp. 1233–1240.

[179] R. Shannon, F.-G. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *Journal of the Acoustical Society of America*, vol. 104, pp. 2467–2476, 1998.

[180] M. V. S. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. II, Apr. 2007, pp. 641–644.

[181] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Apr. 1990, pp. 357–360.

[182] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1994, pp. 77–80.

[183] P. Smaragdis, "Redundancy reduction for computational audition, a unifying approach," Ph.D. dissertation, Massachusetts Institute of Technology, Jun. 2001.

[184] ——, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Sep. 2004, pp. 494–499.

[185] ——, "Discovering auditory objects through non-negativity constraints," in *Proceedings of the ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Oct. 2004.

[186] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[187] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 2069–2072.

[188] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, 2006.

[189] D. L. Snyder, T. Schulz, and J. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Transactions on Signal Processing*, vol. 40, pp. 1143–1150, 1992.

[190] S. Sra and I. S. Dhillon, "Nonnegative matrix approximation: Algorithms and applications," University of Texas at Austin, Tech. Rep. TR-06-27, Jun. 2006.

[191] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[192] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 708–716, Nov. 2000.

[193] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech'2006 - ICSLP)*, Sep. 2006.

[194] H. Valpola, "Behaviourally representations from normalisation and context-guided denoising," AI Lab, University of Zurich, Tech. Rep., 2004.

[195] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Sep. 1997.

[196] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *Communications, Speech and Vision, IEE Proceedings I*, vol. 137, no. 1, pp. 38–46, Feb. 1990.

[197] R. Vautard and M. Ghil, "Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time series," *Physica D*, vol. 35, pp. 395–424, 1989.

[198] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals," *Physica D*, vol. 58, pp. 95–126, 1992.

[199] R. Veldhuis, *Restoration of Lost Samples in Digital Audio Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[200] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 173–185, Mar. 2002.

[201] E. Vincent, "Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux," Ph.D. dissertation, Université Paris VI – Pierre et Marie Curie, 2004, in French.

[202] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 109–112.

[203] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[204] T. Virtanen, A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008, pp. 1825–1828.

[205] D.-L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[206] D.-L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley, 2006.

[207] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 1970.

[208] ——, *Auditory Perception: A New Synthesis*. New York, NY: Pergamon, 1982.

[209] M. Weintraub, "A theory and computational model of monaural auditory sound separation," Ph.D. dissertation, Stanford University, 1985.

[210] L. Wiskott and L. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, pp. 715–770, 2002.

[211] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. V, Mar. 2005, pp. 517–520.

[212] M. Wu, D.-L. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, May 2003.

[213] R. K. Young, *Wavelet Theory and its Applications*. New York, NY: Springer, 1992.

[214] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice Hall, 1969.

[215] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, Jul. 2007.

[216] "Edinburgh Speech Tools Library," http://www.cstr.ed.ac.uk/.

[217] "Speech Filing System," http://www.phon.ucl.ac.uk/resource/sfs/.