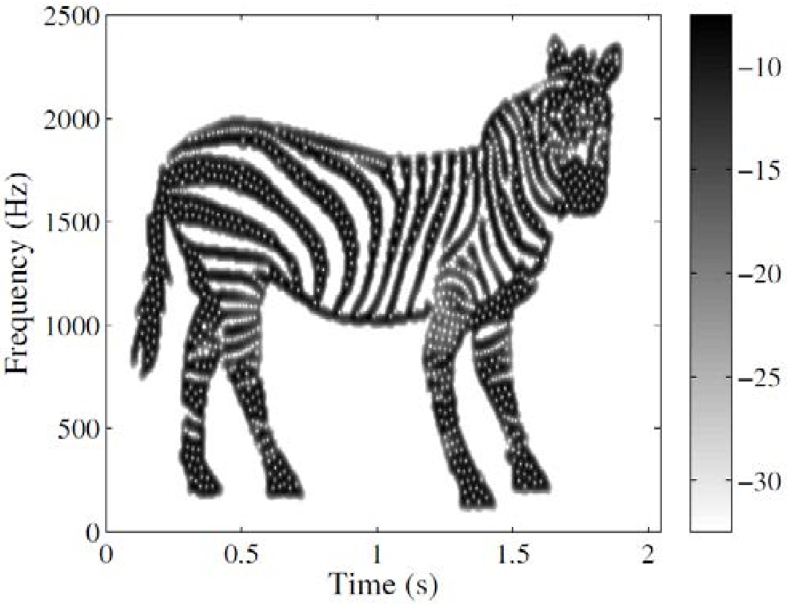
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 14

Rémi Decorsière

Spectrogram inversion and potential applications for hearing research



Centre for Applied Hearing Research
Department of Electrical Engineering

Spectrogram inversion and potential applications for hearing research

PhD thesis by
Rémi Decorsière

CAHR
Centre for Applied Hearing Research

Technical University of Denmark
2013

Preprint version for the assessment committee.
Pagination will differ in the final published version.

This PhD-dissertation is the result of a research project at the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark).

The project was financed by a stipend from the Oticon foundation.

Supervisors

Main supervisor

Prof. Torsten Dau
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Co-supervisors

Peter L. Søndergaard
Mathematics and Signal Processing in Acoustics
Acoustic research institute
Austrian academy of science
Vienna, Austria

Ass. Prof. Ewen N. MacDonald
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

A common way of analyzing signals in a joint time-frequency domain is found in the spectrogram, which can be interpreted as a multi-channel envelope representation of the signal. The envelope cannot fully represent a signal because it only reflects slow changes in the amplitude of a signal and lacks information regarding its fast variations, the temporal fine structure (TFS). However, the main hypothesis explored in this thesis is that a spectrogram could be a faithful representation of a signal, that is, TFS information could be recovered by across-channel comparison of envelopes. Based on this consideration, an approach for spectrogram inversion was proposed: time-domain signals were recovered from spectrograms computed using both inner hair-cell envelope (i.e., traditional half-wave rectification followed by low-pass filtering) and Hilbert envelope definitions. The high accuracy of the inversion scheme (as measured by root mean square error and spectral convergence) implies that the main hypothesis holds true for the designs chosen. Two practical applications of this result were then presented. (1) Spectrograms that are computed using the inner hair-cell (IHC) envelope definition are a reasonable model of the signal processing performed by the human cochlea. The robustness of the reconstruction from such spectrograms with regards to the properties of the cochlear model showed that, for previously documented IHC models as well as for more restrictive conditions, the TFS-related information is retained by the (modeled) cochlear processing even at high audio frequencies. (2) Using the inversion framework, it is possible to manipulate signals in the modulation domain, while preserving their long-term power spectra. Thus, this enabled the creation of mixtures of speech and noise where the signal-to-noise ratio in the envelope domain (SNR_{env}) was directly controlled. Behavioral measures of the intelligibility for such mixtures were compared to predictions from a model of speech intelligibility. Conditions where noise was processed led to modest intelligibility improvements for increased SNR_{env}, providing direct validation of the intelligibility model. Processing speech proved to be challenging and did not result in improved intelligibility, in contrast to the model predictions. The challenges encountered when processing speech were further explored, but could not be completely circumvented, and accurate modulation filtering of speech signals remains challenging.

Resumé

Spektrogrammer anvendes ofte til at analysere signaler i et samlet tid-frekvens domæne. Et spektrogram kan tolkes som en multi-kanal indhyldningskurve af et signal. Indhyldningskurven repræsenterer langsomme ændringer i signalets amplitude, men indeholder ikke information om signalets hurtige fluktuationer, den såkaldte temporale fin-struktur (TFS). Derfor er indhyldningskurven ikke en fuldstændig repræsentation af et signal. Hovedhypotesen i denne afhandling er at fordi et spektrogram består af flere kanaler kan et spektrogram potentielt set være en fuldstændig repræsention af det oprindelige signal. Baseret på denne antagelse foreslås en fremgangsmåde til at invertere et spektrogram: Et signal i tids-domænet genskabes ud fra et spektrogram hvor indhyldningskurven er baseret på enten en model af indre hårceller (dvs. en ensretter efterfulgt af et lavpas-filter), eller ud fra definitionen givet ved Hilbert indhyldningskurven. Nøjagtigheden af rekonstruktionen indikerer at hovedhypotesen er sand for de konfigurationer der er anvendt i dette studie. To praktiske anvendelser af metoden præsenteres efterfølgende: (1) Spektrogrammer, beregnet ved hjælp af indhyldningskurver defineret ud fra en model af indre hårceller (IHC), er en rimelig repræsentation af den menneskelige cochleas signal-behandling. Signaler, rekonstrueret ud fra spektrogrammer dannet ved hjælp af tidligere foreslåede IHC modeller, såvel som mere restriktive implementationer af modellerne, viser en høj grad af robusthed, hvilket indikerer at TFS informationen bevares i den modellerede cochleare signal-behandling; selv ved høje audio-frekvenser. (2) Rekonstruktion af modificerede spektrogrammer muliggør at temporale modulationer (dvs. variationer i indhyldningskurven over tid) kan ændres. Dette gør at man kan skabe signaler bestående af tale og støj hvor signal-støjforholdet i indhyldningskurve-domænet (SNR_{env}) kan styres direkte. Målinger af taleforståeligheden af sådanne signaler blev sammenlignet med beregninger fra en model for taleforståelighed. Forsøgsbetingelser hvor støjen var modificeret gav beskedne forbedringer af taleforståeligheden ved en øget SNR_{env}, hvilket understøtter modellen for taleforståelighed. Forsøgsbetingelser med modificeret tale viste sig at være en udfordring og gav ikke en øget taleforståelighed, i modsætning til modellens beregninger. Udfordringerne der opstod i forbindelse med at modificere tale blev yderligere undersøgt, men det lykkedes ikke at overkomme dem fuldstændigt. En nøjagtig modulations-filtrering af tale er derfor stadig et åbent problem.

Résumé

Le spectrogramme, qui est une méthode communément utilisée pour l'analyse de signaux dans un domaine conjoint en temps-fréquence, peut être interprété comme une représentation multicanale de l'enveloppe du signal. L'enveloppe, en tant que telle, reflète les lents changements d'amplitude d'un signal et ne renseigne donc pas sur ses variations rapides, i.e., sur sa structure temporelle fine (STF). Par conséquent, l'enveloppe seule ne peut représenter entièrement un signal. L'hypothèse principale développée dans cette thèse est que le spectrogramme, cependant, parce qu'il implique plusieurs canaux, pourrait être une représentation fidèle du signal. Partant de cette hypothèse, une méthode d'inversion du spectrogramme est proposée permettant de reconstruire des signaux temporels à partir de spectrogrammes basés soit sur un modèle d'enveloppe des cellules ciliés internes (CCI) consistant en un redressement demi-onde suivi d'un filtre passe-bas, soit sur l'enveloppe de Hilbert. La fidélité des signaux reconstruits suggère que notre hypothèse est vérifiée pour les deux cas étudiés. Deux applications pratiques de ces résultats sont ensuite proposées. (1) Les spectrogrammes calculés à partir de l'enveloppe des CCI sont un modèle raisonnable de la représentation obtenue en sortie de la cochlée. La robustesse de la méthode de reconstruction face à la modifications des paramètres du modèle cochléaire montre que, pour les modèles de CCI présentés dans la littérature ainsi que pour des conditions plus sévères, l'information sur la STF est toujours présente dans la représentation cochléaire, même en hautes fréquences. (2) La reconstruction de spectrogrammes modifiés rend possible la manipulation des modulations temporelles (i.e., la variation de l'enveloppe au cours du temps), permettant ainsi d'imposer un certain rapport signal-bruit dans le domaine de l'enveloppe (SNR_{env}) pour des mélanges de parole et de bruit. L'intelligibilité de ces signaux est mesurée et comparée aux prédictions d'un modèle. Lorsque le bruit est manipulé, il est démontré qu'un SNR_{env} accru génère une amélioration modérée de l'intelligibilité. En revanche, l'intelligibilité est dégradée lorsque le signal de parole est manipulé. Les limitations rencontrées dans ce cas précis sont étudiées en détail, mais n'ont pu être entièrement contournées. Ainsi, le développement d'une méthode suffisamment précise pour filtrer les modulations dans la parole reste encore un défi.

Preface

Some say "time flies when you are having fun". Well, that must be it then. I remember my first steps on Danish land, about six years ago. It was different from what I imagined. Instead of axe-wielding furious vikings riding lego-made bicycles and chewing on pickled herring, I found myself surrounded by very calm, very discreet, though fairly pleasant adepts of the Danish "hygge". Not quite the same, even though I happened to be right for the bike-and-herring part.

What struck me the most was my first day in building 352. A couple hours were more than enough to convince me that I had made the right choice in traveling there to pursue my studies. As a student, receiving such a warm welcome from the staff felt surreal, almost suspicious some might say. But it was indeed honest, and today I am very thankful that I could become a part of it, a bit more than three years ago when this PhD project started.

Forty months is a long time for one project. And it is thanks to the support I got, both professionally and personally, that this manuscript could "come to life". My deepest acknowledgments therefore go to my supervisors. Torsten, with his enthusiasm in the project and his overwhelming continuous flow of brilliant ideas, truly put a literal meaning to his function of "vejleder". He was always there when needed, smoothing my motivational ups and downs; he brought me down to earth when I got too excited over details and more importantly, cheered me up when I was down. To this I am extremely thankful, and I challenge anyone to try and combine such perfectly tailored supervision with that busy an agenda.

Peter, as a much needed technical adviser, really was the cornerstone of this project. He managed to communicate his enthusiasm and his impressive knowledge of the topic to me, despite the few thousand kilometers separating us in the last year. Thank you for being there, it would have been impossible (with a probability of 1) to conduct this project without your non-measurable help.

I am very thankful to Ewen too. He has played a fundamental role in the last milestones of the project. Although I pushed his native English skills (well, probably more his patience) to the limit with all the reviewing I asked him to do, I rarely saw someone as receptive and cheerful.

Additionally, I would like to thank fellow researchers from CAHR. You are what makes this group a lovely place to work. You gave me the support I needed from colleagues, as well as from friends.

On a more personal note, deep thanks go to all my friends here in Denmark. They are the reason why I call this country "home" now. But I have not forgotten what used to be home, and I thank my family and friends back there for their support and for the quality time they spent with me.

Rémi Decorsière, June 2013.

Contents

Abstract **v**

Resumé på dansk **vii**

Résumé en français **ix**

Preface **xi**

Table of contents **xiii**

List of abbreviations **xvii**

List of symbols **xix**

1 Introduction **1**

1.1 The envelope: a conceptual attribute of a signal 1

1.1.1 An intuitive concept... 1

1.1.2 ... that needs mathematical formalizing 2

1.1.3 Interaction between envelope and carrier wave in a narrowband signal . . 4

1.1.4 The spectrogram as an envelope representation of broadband signals . . . 4

1.1.5 Redundancies in the spectrogram playing a role in its representation of the signal 5

1.2 Temporal modulation processing and spectrogram reconstruction 7

1.3 Outline of the thesis 7

2 On the relationship between envelope and temporal fine structure **11**

2.1 Formalization of the concept of envelope 11

2.1.1 Historical approach to envelope detection 11

2.1.2 Inner hair-cell envelope detection models 12

2.1.3 Mathematical and signal processing approaches 14

2.2 The spectrogram as a faithful signal representation 18

2.2.1 Single channel envelope 18

2.2.2 Short-time Fourier transform with Gaussian windows 19

2.2.3 General redundant linear system with Hilbert envelope 21

2.2.4 Simple auditory models 23

2.3 The role of envelope in psychoacoustics 23

3 Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations 29

3.1 Introduction 29
3.2 General framework 31
3.3 Reconstruction from IHC inspired envelope extraction 34
3.3.1 Gammatone filterbank 34
3.3.2 IHC inspired envelope extraction 34
3.3.3 Direct reconstruction 35
3.3.4 Two-step reconstruction 36
3.4 Reconstruction from STFT magnitude 37
3.5 Evaluation and comparison of techniques 39
3.5.1 Implementation, testing material and evaluation of convergence 39
3.5.2 Results from IHC inspired envelope representations 39
3.5.3 Results for STFT magnitude 41
3.6 Discussion 42
3.6.1 Direct vs. two-step approach in reconstruction from IHC envelope 43
3.6.2 Loudness-based objective function for reconstruction from STFT magnitude 43
3.6.3 Reducing intrinsic limitations of reconstruction from STFT magnitude . . 44
3.6.4 Implications for current IHC models 46
3.7 Conclusion 46

4 Retrieval of temporal fine structure from inner hair-cell envelopes of unresolved complex tones 47

4.1 Introduction 47
4.2 Method 48
4.2.1 Stimulus 48
4.2.2 Envelope extraction 48
4.2.3 Reconstruction from the envelope 49
4.2.4 Evaluating the retrieved TFS 49
4.3 Results 50
4.4 Discussion 51

5 Effects of manipulating the envelope signal-to-noise ratio on speech intelligibility 53

5.1 Introduction 53
5.2 Method 55
5.2.1 Speech material, apparatus, and procedure 55
5.2.2 Stimulus conditions 56
5.2.3 Modulation processing framework 57
5.2.4 Speech intelligibility prediction 59
5.3 Results 60
5.4 Discussion 62

5.4.1 Modulation processing of the noise interferer 62
5.4.2 Modulation processing of clean speech 63
5.4.3 Limitations of modulation enhancement in speech processing 63
5.4.4 Usefulness of modulation processing for speech intelligibility enhancement 64
5.5 Summary and conclusions 65

6 Insights in spectrogram filtering: directions for improvements 67
6.1 Short-time based spectrogram filtering 67
6.1.1 Method 68
6.1.2 Results 71
6.1.3 Conclusion 72
6.2 Non-negative realizations of filtered spectrograms 73
6.2.1 Problem formulation 74
6.2.2 Feasibility of a non-negative impulse response modulation filter 74
6.2.3 Channel dependent non-negative filtering 76
6.3 Towards a target-less approach to modulation filtering 82

7 General discussion 85
7.1 Results overview 85
7.2 Additional discussions 88

Bibliography 93

Collection volumes 101

List of abbreviations

AC	Alternating current
AI	Articulation index
AM	Amplitude modulation
CLUE	Conversational language understanding evaluation
DC	Direct current
DFT	Discrete Fourier transform
EPSM	Envelope power spectrum model
ERB	Equivalent rectangular bandwidth
ESII	Extended speech intelligibility index
FFT	Fast Fourier transform
FIR	Finite impulse response
HINT	Hearing in noise test
IHC	Inner hair-cell
l-BFGS	Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm
LTFAT	Linear time-frequency analysis toolbox
MTF	Modulation transfer function
OLA	Overlap-add
PESQ	Perceptual evaluation of speech quality
RMS	Root mean square
RMS_n	Normalized root mean square error
SII	Speech intelligibility index
sEPSM	Speech-based envelope power spectrum model
SNR	Signal-to-noise (energy) ratio
SNR_{env}	Signal-to-noise ratio in the envelope domain
SPL	Sound pressure level
SRT	Speech reception threshold
SSN	Speech-shaped noise
STFT	Short-time Fourier transform
STI	Speech transmission index
stMTF	Short-time modulation transfer function average
TFS	Temporal fine structure
WOLA	Weighted overlap-add
ΔSRT	Change in speech reception threshold

List of symbols

$\forall$	For all
$\in$	In
$\mathbb{Z}$	Set of integers
$\mathbb{N}$	Set of natural numbers (non-negative integers)
$[a..b]$	Subset of integers n with $a \leq n \leq b$
$\mathbb{R}$	Field of real-valued numbers
$\mathbb{C}$	Field of complex-valued numbers
$\mathbb{C}^{M \times N}$	Set of M-lines, N-columns matrices of complex coefficients
$\mathbf{A}_{m,n}$	(m,n)-th coefficient of matrix $\mathbf{A}$ in $\mathbb{C}^{M \times N}$
$s^{\mathbf{T}}$	Transpose of vector s
f'	Derivative of univariate function f
$\frac{\partial f}{\partial x}$	Partial derivative of multivariate function f with regard to variable x
∇f	Gradient of function f
$\lVert s \rVert = \lVert s \rVert_2$	Euclidean norm of finite-length signal (or vector) s
$\lVert \mathbf{A} \rVert_{fro}$	Frobenius norm of matrix $\mathbf{A}$

i	Imaginary unit
$\lvert c \rvert$	Magnitude (if $c \in \mathbb{C}$), or absolute value (if $c \in \mathbb{R}$) of c
$\angle c$	Phase of coefficient c
$\Re(c)$	Real part of complex coefficient c
$\Im(c)$	Imaginary part of complex coefficient c
$\overline{c}$	Complex conjugate of c
$\log(\cdot)$	Natural logarithm
$\log_{10}(\cdot)$	Base 10 logarithm
e^x	Exponential of variable x

$s_1 * s_2$	Convolution between signals s_1 and s_2
$\mathscr{F}\{\cdot\}$	Fourier transform
$\mathfrak{H}(s) = \widehat{s}$	Hilbert transform of signal s
s_a	Analytic signal associated to s: $s_a = s + j\widehat{s}$
$\mathscr{T}$	Translation operator: $(\mathscr{T}\{s\})[k] = s[k-1]$

$\mathscr{H}\left\{\mathbf{A}\right\}$	Heaviside function applied to matrix **A**

$\left\{g_m\right\}_{1\leq m\leq M}$	Set of M filters' impulse responses used for filterbank analysis
$V_g s$	General filterbank analysis operator applied to signal s using the filters $\left\{g_m\right\}_{1\leq m\leq M}$
$U_g s$	General filterbank synthesis operator applied to signal s using the filters $\left\{g_m\right\}_{1\leq m\leq M}$
$\left\{g_m^{(d)}\right\}_{1\leq m\leq M}$	Dual windows associated to $\left\{g_m\right\}_{1\leq m\leq M}$: $U_{g^{(d)}}\left(V_g s\right) = s$

$E\left(s\right)$	General denomination for envelope extraction of signal s
$\mathbf{E}_s$	Multi-channel envelope of signal s
$\mathscr{G}$	General denomination of the objective function in an optimization procedure
s^+, $\mathbf{A}^+$	Half-wave rectification of signal s or matrix **A**
$\mathscr{C}$	Spectral convergence

1

Introduction

This chapter presents a conceptual introduction to the problems posed in this thesis. It describes the intuitive approach to defining the envelope of a signal, as well as the limitations this concept faces in real-life scenarios. A simple example then leads to the formulation of the main hypothesis behind this work: that multi-channel envelope representations (i.e., spectrograms) could, unlike single-channel envelopes, be a faithful representation of a signal, i.e., a representation that retains all the information related to the signal. An application of this concept to spectrogram inversion and later to modulation filtering is suggested. An overview of the content of this thesis is then finally presented.

1.1 The envelope: a conceptual attribute of a signal

1.1.1 An intuitive concept...

The concept of an envelope is easily introduced by a common perceptual experiment. When a musical instrument produces two notes that are close in frequency, though slightly different (e.g., if the instrument is not properly tuned), a beating phenomenon can be heard: the resulting signal presents slow periodic changes in its "volume", as if it was "beating". It can, for example, be very clearly heard when striking two tuning forks tuned at slightly different frequencies, as illustrated in fig.1.1. Letting the two forks resonate simultaneously will result in a sound with two audible attributes:

- The resulting sound's pitch is "in between" the two pitches of individual forks.
- The resulting sound is pulsating at a fixed rate. Experimenting with different tuning forks shows that this rate increases for increasing mistuning of the forks.

This experiment reveals an interesting phenomenon: although the resulting signal is composed of two tones with very close fundamental frequencies, we perceive a pulsating tone where the pulse rate and the pitch of the tone involve two frequencies far apart. This perceptual dichotomy (a given pitch and a given pulsating rate) can be accounted for by introducing the decomposition of the signal into the product of an envelope and a carrier wave. The envelope would account for the slow variations of the signal, in this case the beating, while the carrier wave would "carry" this envelope and provide a fundamental frequency, in this case giving the pitch information.

An intuitive definition of the envelope of a signal could be given with the two following points:

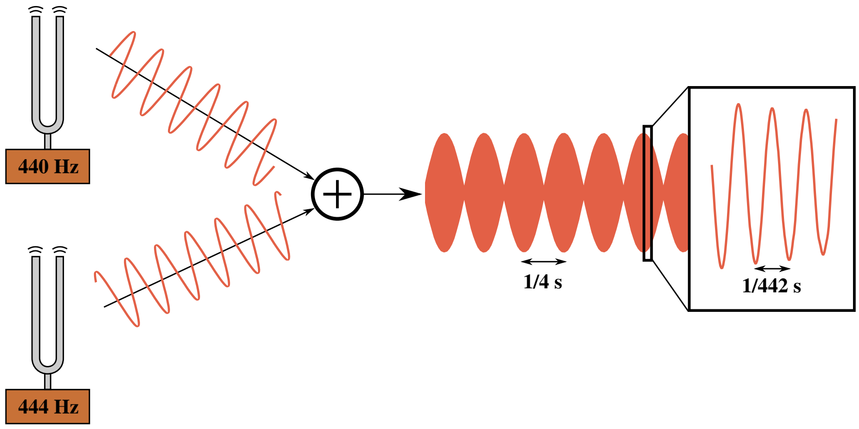


Figure 1.1: The tuning forks experiment: two tuning forks of slightly different tuning (e.g., 440 and 444 Hz) will produce a resulting signal with a clearly audible 4 Hz modulation and a fundamental frequency at 442 Hz

(i) The envelope is tangent to the signal at, or in a close vicinity to each of its local maxima.

(ii) The envelope env surrounds the signal s, i.e., $-env \leq s \leq env$ at every point in the signal.

These two points in the definition imply that the envelope is a non-negative signal. This is an arbitrary choice, as point (i) above could be replaced with "tangent to local minima" and the inequality in (ii) reversed, resulting also in a viable, though slightly less practical, non-positive envelope definition. Under this consideration, a more specific definition can be formulated:

> The envelope is tangent to the absolute value of the signal at, or in a close vicinity to its local extrema.

1.1.2 ... that needs mathematical formalizing

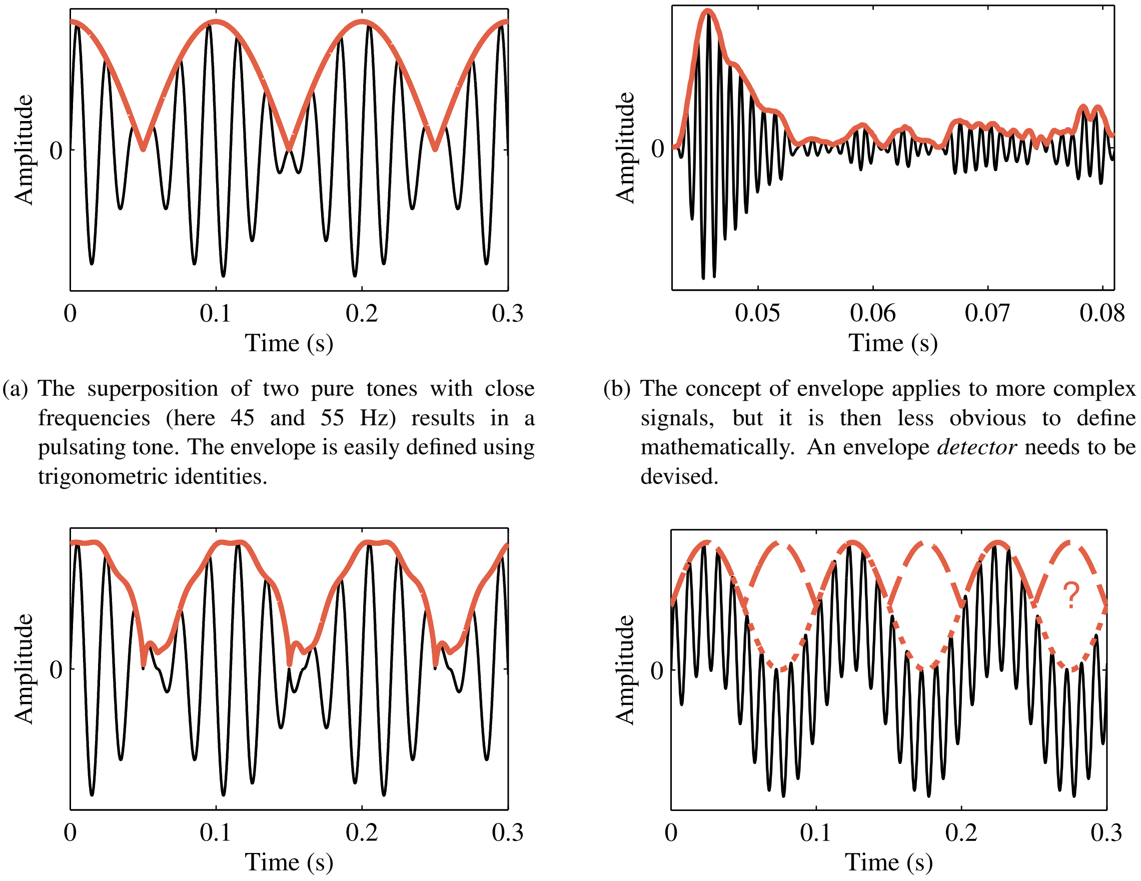
The tuning fork example is convenient to expose a mathematical decomposition into envelope and carrier wave, as the signal it generates is very close to a pure sine wave. When both forks are struck, the resulting signal is a sum of two sine waves, with frequencies f_1 and f_2, with $f_1 \approx f_2$. Without loss of generality, it can be assumed that the two sinusoidal components have the same amplitude and starting phase. The resulting signal is then given as follows:

$$s = \sin(2\pi f_1 t) + \sin(2\pi f_2 t) \tag{1.1}$$

where t denotes time. Using a standard trigonometric formula, we can reformulate this sum into a product:

$$s = 2\cos\left(2\pi \frac{f_1 - f_2}{2} t\right) \sin\left(2\pi \frac{f_1 + f_2}{2} t\right) \tag{1.2}$$

Keeping in mind that $f_1 \approx f_2$, one can relate the product form of (1.2) to the perceptual decomposition evoked earlier, of a product of a slow varying envelope (as $f_1 - f_2 \approx 0$) with a fast varying carrier wave (as $\frac{f_1+f_2}{2} \approx f_1$ or f_2). The envelope for this signal can therefore



(a) The superposition of two pure tones with close frequencies (here 45 and 55 Hz) results in a pulsating tone. The envelope is easily defined using trigonometric identities.

(b) The concept of envelope applies to more complex signals, but it is then less obvious to define mathematically. An envelope *detector* needs to be devised.

(c) Envelope and carrier wave cannot be manipulated independently. Here, the envelope of (a) was time shifted and recombined with the original carrier. The envelope of the new signal differs from the original time-shifted envelope.

(d) The concept of envelope holds for narrowband signals. Here, two pure tones of 10 and 100 Hz are superimposed. The intuitive definition of envelope becomes ambiguous.

Figure 1.2: The envelope (orange lines) for various signals (black lines) illustrating basic envelope properties: (a) the envelope of two superimposed pure tones with similar frequencies is easily defined, (b) the envelope exists also for "complex" signals, (c) the envelope cannot be manipulated independently of the carrier wave, (d) the concept of envelope becomes ambiguous for signals of wider bandwidth.

be defined as the absolute value of the low-frequency cosine component (as it is by definition non-negative), together with the amplitude factor 2:

$$\mathrm{env}\left(s\right) = 2\left|\cos\left(2\pi\frac{f_1-f_2}{2}t\right)\right| \tag{1.3}$$

This decomposition is illustrated in fig.1.2(a), though with lower frequencies than in the tuning fork experiment for better visibility.

The identification of an envelope is possible in the case of pure tones due to particular properties of the sine waves. However, the intuitive concept of envelope holds as well for more complex signals. For example, it is clear that the signal plotted in fig.1.2(b) has a distinct envelope (orange line), but it cannot be defined mathematically as easily as in the previous case. For general cases, a so-called *envelope detector* needs to be devised: a mathematical tool that extracts the envelope from a given signal. Various mathematical definitions of envelope are suggested in the literature, with each fitting our intuitive definition of envelope to different extents. Some of these definitions and their resulting properties will be presented later in chapter 2. Two common envelope definitions are then used throughout the thesis: the Hilbert envelope (in chapters 3, 5 and 6) and inner hair-cell envelope models (in chapters 3 and 4).

1.1.3 Interaction between envelope and carrier wave in a narrowband signal

Both the envelope and its associated carrier wave are needed to entirely describe a signal. Simple examples can be found in sine waves. The envelope of a pure sine wave is usually considered to be a flat line, at a value corresponding to the amplitude of the sine wave. The associated carrier is then the corresponding sine wave, with a normalized amplitude. Hence, given a flat envelope, it is impossible to figure out the original signal associated to it, as it could be a sine wave of any frequency. Conversely, given a sine wave carrier, no information regarding the amplitude of the original signal is given and it is impossible to deduce, from the carrier wave only, what exactly the original signal was.

For single sine waves, envelope and carrier are actually independent, i.e., the envelope (flat line) contains no "information" regarding the carrier wave, and vice-versa. They can be manipulated individually to change the amplitude or frequency of the sine. However this is not the case in general, and a manipulated envelope might not be "compatible" with its previously associated carrier wave. A simple example, illustrated in fig.1.2(c), is obtained by slightly time shifting the envelope obtained in fig.1.2(a), and recombining it (i.e., multiplying it) with the original carrier wave from fig.1.2(a). The resulting signal does not present the expected time-shifted envelope, but its new envelope is significantly different, suggesting that in that case there was an interaction between both envelope and carrier wave, and they cannot be considered either to be independent or manipulated independently.

1.1.4 The spectrogram as an envelope representation of broadband signals

So far, the notion of envelope was introduced for narrow-band signals (sine wave, beating tone and the "complex" signal in fig.1.2(b)). Real-life acoustic signals such as speech, music, or noise, are generally broadband. The concept of envelope as introduced in 1.1.1 becomes ambiguous for such broadband signals. Figure 1.2(d) illustrates this ambiguity, once again with the waveform of the sum of two sine waves, but this time with frequencies further apart (here 10 and 100 Hz). Defining an envelope that complies to the two points given in 1.1.1 becomes challenging. One envelope candidate, plotted as a dotted line, follows point (i), being tangent to the local maxima of the signal. However it goes against (ii), as the signal would not be included in between plus/minus the envelope. Another candidate, plotted as a dashed line, fits better to the final definition given in 1.1.1, but it goes against intuition and point (i), as it is not tangent to local maxima in a wide

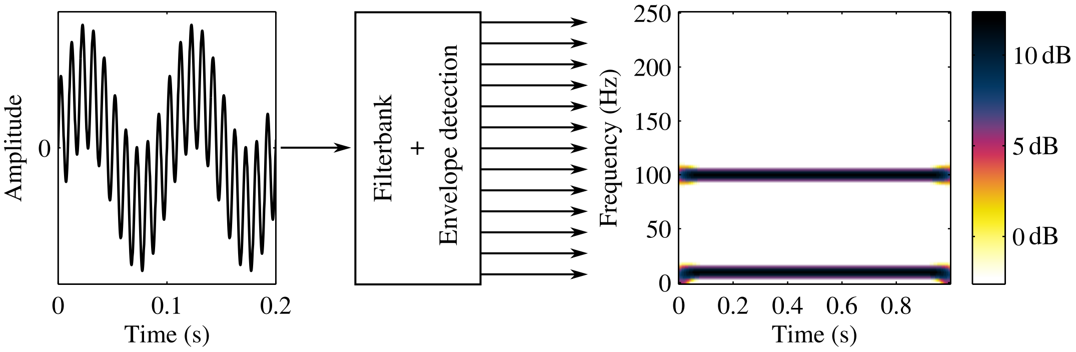


Figure 1.3: The envelope of broadband signals can be obtained by decomposing them into narrowband channels by use of a filterbank. The envelopes for every channels forms the spectrogram, expressed in dB. With this representation, the 10 Hz and 100 Hz components from the signal in 1.2(d) are separated and the ambiguity of the envelope definition is removed.

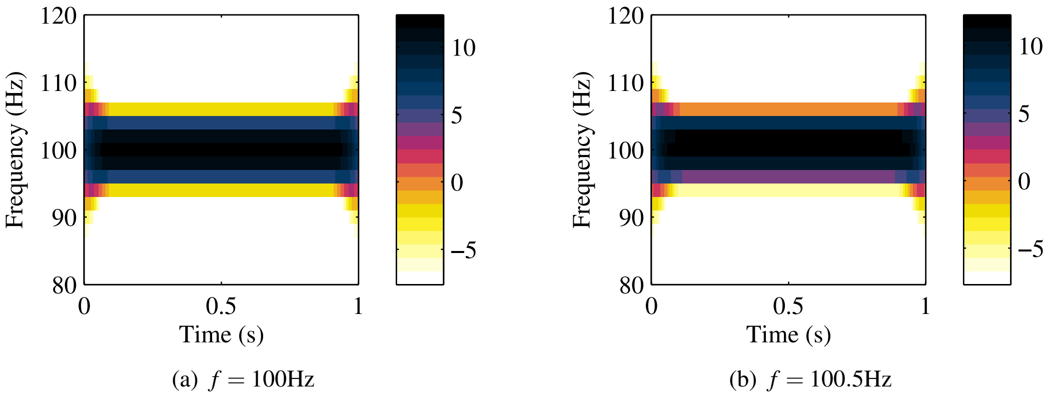
portion of the signal. Because of this ambiguity, there is no definition of envelope that is applicable to all broadband signals.

The common way to avoid this problem is to decompose broadband signals into narrowband components, by means of a bank of bandpass filters. For each of these components, or *channels*, the ambiguity is removed and an envelope can be identified. This decomposition process is illustrated in fig.1.3 for the signal in fig.1.2(d). This representation, with two horizontal lines for the channels centered at 10 and 100 Hz, reveals that the signal is composed of two sine components at these frequencies.

In this thesis, the multichannel envelope representation (i.e., the collection of all the channels' envelopes) will be referred to as a *spectrogram* of the signal. Note that the spectrogram is not unique, as it depends on the choice of a filterbank and how the envelope is defined. Although the term "spectrogram" does not have a strict definition, it is often used in the literature for a particular case of the multichannel envelope representation, that obtained from the squared magnitude of the short-time Fourier transform (STFT) coefficients. To avoid any confusion, this particular representation will be referred to as "traditional spectrogram" in sections where it is introduced (i.e., mainly in chapter 3).

1.1.5 Redundancies in the spectrogram playing a role in its representation of the signal

Individual filters from the filterbank involved in the spectrogram derivation usually present a finite slope in their frequency responses. Hence, there is some degree of overlap between filters in the frequency domain. This implies that even if a signal is very narrowband, its spectrogram will not be concentrated in one channel. For example, the spectrogram of a 100 Hz sine wave presented in fig.1.4(a) is clearly centered around the 100 Hz channel, but its six closest neighbor channels also present a significant activity. Hence the spectrogram is usually a redundant envelope representation, as a given channel contains some amount of information related to the content of its neighbors.



(a) $f = 100$Hz

(b) $f = 100.5$Hz

Figure 1.4: Spectrograms (detail) of a sine wave (a) of frequency 100 Hz and (b) of frequency 100.5 Hz

These redundancies may result in the spectrogram being a faithful representation of the signal. In section 1.1.3, an example was provided to illustrate the fact that for a narrowband signal, neither the envelope nor the carrier wave could, in general, truly represent the signal. In other words, taking the envelope of a signal results in a loss of information. However in the case of a spectrogram, each channel in the vicinity of the frequency region of this narrowband signal will contain a redundant but different representation of this signal. When the envelopes of the channels are taken to form the spectrogram, many "versions" of the envelope will be described. Hence the spectrogram contains more information regarding the signal than only a single envelope. Extrapolating this idea leads to the hypothesis around which all the work conducted in this thesis revolves:

> "Given a sufficiently redundant filterbank (i.e., sufficiently many channels and overlap between channels), the spectrogram alone is a faithful representation of the signal, i.e., it contains the same information as the original signal."

An illustration of this is given in fig.1.4 which presents the spectrogram of two sine waves with very close frequencies (100 Hz for (a) and 100.5 Hz for (b)). As mentioned earlier, the envelope of these two sine waves are the same constant signals, so their envelope only would not allow to differentiate between them. However their spectrograms differ. The proximity of the two frequencies causes the same channels to be activated in the two spectrograms. Although the center channel (at 100 Hz) present the same amplitude for both spectrograms, the neighboring channels present different amplitude such that it is possible, from their spectrograms, to differentiate the two signals. Moreover, with knowledge of the filterbank that was used, and particularly the amount of overlap between channels, the distribution in amplitude of the neighboring channels could allow one to derive very precisely the original frequencies of the two tones, i.e. 100 and 100.5 Hz.

In this thesis, a method will be presented that recovers a time-domain signal corresponding to a given spectrogram, i.e., a method for spectrogram inversion, or spectrogram reconstruction. It will not imply a direct comparison between channels, as could be done for the sine wave examples that were just illustrated. However, it will rely on the redundancy that the spectrogram offers as a representation of the signal. It will be seen that time-domain signals can be recreated with a large degree of accuracy in terms of spectral convergence and for some cases root mean square error (two metrics that will be introduced in chapter 3), suggesting that the main hypothesis proposed above holds true for common designs of filterbanks.

1.2 Temporal modulation processing and spectrogram reconstruction

In this thesis, the main motivation in designing a mathematical tool to retrieve time-domain signals associated to a given spectrogram has to do with temporal modulation processing. "Temporal modulation" is a term used to refer to the variations of the envelope representation of a signal over time. For general broadband signals, that is the fluctuation of the channels in the spectrogram across time. There is a close relationship between temporal modulations and speech intelligibility that has been found in previous studies, and an overview of this will be presented in the following chapter. Hence for some applications there is a need for control or manipulation of the (temporal) modulation content of signals, either the speech or the noise present in the listening environment. Some manipulations could potentially lead to a speech intelligibility improvement, or provide a temporal modulation manipulation framework that would allow for further investigations of the interaction that exists between temporal modulation and speech intelligibility.

However, the computation of the spectrogram involves envelope detection in several channels, and as it will be shown while introducing common mathematical definitions of the envelope, envelope extraction is always a non-linear operation. Manipulating the envelope in the temporal domain of any signal is therefore non-trivial. A good alternative is to manipulate it in the spectrogram domain. Any type of processing can be applied to the spectrogram of a signal, resulting in a processed spectrogram. As was illustrated in fig.1.2(c), the processed spectrogram cannot be recombined with the original carrier waves obtained for each channel. That is where a spectrogram reconstruction is needed, allowing for the recovery of a time-domain signal whose spectrogram is similar to the processed spectrogram obtained earlier. If these steps are followed successfully, then effective *modulation filtering* of the signal can be carried out.

1.3 Outline of the thesis

This first chapter aimed at presenting, through simple examples and illustrations, a conceptual background for the main content that will follow. The concept of envelope was introduced, as well as some of its basic properties. In general, for broadband signals, the concept of envelope is only clearly defined through spectrograms: a collection of the envelopes taken for each output channel of a redundant filterbank. Although the envelope in one channel is not sufficient to characterize this channel, the collection of envelopes in a spectrogram might provide sufficient information to fully recover the original signal.

Chapter 2 follows up on the conceptual introduction of this first chapter, and presents previous elements of literature related to the concepts involved in this thesis. Historical overviews of the concepts relating to envelope, spectrogram, and their relationship with psycho-acoustical findings are developed.

In **Chapter 3**, an optimization approach to spectrogram reconstruction is investigated. If a spectrogram is sufficiently redundant, then it should be possible to recover the time-domain signal associated to a given, *target* spectrogram. In this chapter, the time-domain signal is considered a variable in an optimization procedure, where the distance between the signal's spectrogram and the target is minimized iteratively. If convergence is reached, the time-domain signal created will have a spectrogram as close as possible (i.e., shortest distance in the envelope domain) to the target spectrogram. This framework is applied to spectrograms obtained with a Gammatone filterbank, which models human peripheral auditory filters, and for two different envelope definitions: an auditory motivated inner hair-cell (IHC) envelope model. The "traditional" spectrogram, obtained from the magnitude of the short-time Fourier transform coefficients, is also investigated.

In **Chapter 4**, the robustness of the reconstruction from an IHC envelope based spectrogram with respect to the parameters used in the IHC model is investigated. The spectrogram reconstruction framework of chapter 3 is applied to an auditory-based spectrogram of a stimulus used in a previous psycho-acoustic experiment regarding pitch. This stimulus contains only frequencies above 5 kHz and the repetition rate of its envelope and of its temporal fine structure (TFS) differ. Results of a previous behavioral study with this stimulus suggested that humans can make use of its TFS information, even though this information is assumed to be lost at such high frequencies due to cochlear processing. If reconstruction of the waveform of the signal from its multi-channel IHC envelope is successful using the suggested framework, this would indicate that this auditory representation preserves the TFS information even at high frequencies.

Chapter 5 develops the concept of modulation filtering using spectrogram reconstruction. The study from Jørgensen and Dau (2011) provides a model that estimates speech intelligibility solely based on the concept of signal-to-noise ratio in the envelope domain (SNR_{env}). This model predicts better intelligibility for increased SNR_{env}. In this chapter, modulation filtering is performed either on the speech or the noise component of a mixture in order to manipulate its SNR_{env}. Intelligibility is measured behaviorally and compared to predictions from the model. The results provide additional validation of the model from Jørgensen and Dau (2011), and explore the possibility of speech intelligibility enhancement through modulation filtering.

Chapter 6 addresses some limitations of the approach used in chapter 5 when attempting to enhance speech. The first step in the modulation filtering scheme suggested here involves filtering the channels of a spectrogram. Issues that are not usually encountered when processing signals arise when processing envelopes. On the one hand the frequency range of interest is much lower than traditional audio frequencies, which can be a limitation in some scenarios. On the other hand an envelope is a non-negative signal, which is shown to be a problematic constraint when trying to manipulate it. These issues are investigated and potential methods to solve them are presented and evaluated.

Finally, **Chapter 7** proposes an overview of the results obtained in this study, with a discussion on the future and viability of spectrogram reconstruction as a method to perform modulation filtering.

2

On the relationship between envelope and temporal fine structure

This chapter presents former work in fields related to the present thesis. It provides a description of the evolution of envelope detection, historically needed for practical applications in physics, and later formalized mathematically for its use in modern signal processing. Former studies on the concept of multi-channel envelope, or spectrogram, and its relationship to the signal it represents are then described. It will be shown how it was proven that in particular scenarios the spectrogram faithfully represents the signal. Finally, previous mentions of the role of envelope in psycho-acoustical studies are listed. In particular, studies that involved manipulations of the envelope of speech and their limitations will be reviewed.

2.1 Formalization of the concept of envelope

2.1.1 Historical approach to envelope detection

The first occurrence of envelope detection in history is probably connected to the early development of radio communication and the conception of the first radio receivers in the very early 20th century. Amplitude modulation (AM) radio broadcasting relies on conveying sound information using electromagnetic waves, as the latter can travel over far longer distances while still being detectable. It uses a high-frequency electromagnetic carrier wave (ranging from a few hundred to several thousand kilo-Hertz in modern AM broadcasting) that is modulated in amplitude by the audio signal, which is lower in frequency. The signal of interest, therefore, formed the *envelope* of the broadcasted signal, as described in section 1.1. Hence some form of physical envelope detection was needed to extract the audio signal back from the high-frequency electrical signal picked up by an antenna.

Early simple AM radio receivers, crystal radios, were composed of an antenna, a crystal detector, and earphones. The central and main component, the crystal detector, was an early version of a semiconductor (i.e., a diode). It let the current flow in one direction while blocking it in the opposite direction. In signal processing, this operation is referred to as *half-wave rectification*: the negative-valued segments of the signal are set to zero but positive-valued segments are left unchanged. The half-wave rectified signal at the output of the crystal still contains mostly very high frequencies and does not reflect the acoustic signal. However, due to their electromechanical properties, the

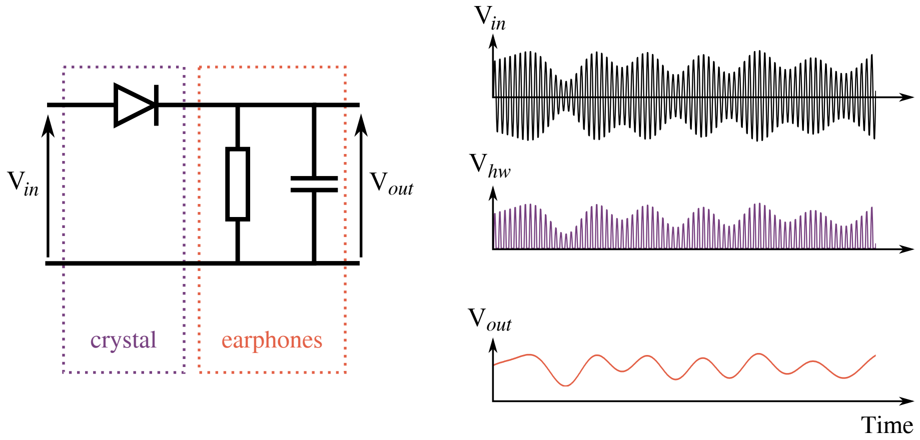


Figure 2.1: Electrical circuit of a basic envelope detector (left) with corresponding input V$_{in}$ and output V$_{out}$ voltages (right). The intermediate half-wave rectified voltage V$_{hw}$ would be measured at the output if the capacitor was removed. This circuit also models the behavior of the crystal and the earphones in early AM radio receivers

earphones could not respond to such high frequencies and produced instead a *low-pass filtered* version of the signal. This mechanical low-pass filtering, later implemented electrically for better rendition, allowed the recovery of the audio signal, i.e., of the *envelope* of the transmitted signal. Though implemented with modern and more accurate electronics and involving an amplification of the electrical signal before delivering it to a loudspeaker, today's versions of AM radios are still based on a similar concept for envelope detection:

- **half-wave rectification** using a diode, assembled in series with a
- **low-pass filter** electric circuit, i.e., a resistor and a capacitor assembled in parallel.

This basic envelope detector is illustrated in fig.2.1, with an example of input and corresponding output voltages (respectively, V$_{in}$ and V$_{out}$). The voltage obtained with the half-wave rectification step alone could be measured on a load resistor at the output of the diode (e.g., at the resistor if the capacitor was removed). It is pictured here as V$_{hw}$ and it is clear that the diode itself is not sufficient and that an additional low-pass filter is needed to recover the envelope. For this AM radio example, the low-pass filter characteristics, given by the resistance and capacitance values of the elements in the circuit, are designed so that the frequencies of the acoustic signal fall into the bandpass section of the filter and the carrier wave frequencies are strongly attenuated. This kind of envelope detector is also commonly used in electronics, where for each specific application, properties of the low-pass filter are adjusted depending on where in frequency the envelope lies.

2.1.2 Inner hair-cell envelope detection models

The previous example showed how envelope detectors were introduced historically in the conversion from an electric to acoustic signal. Conversely, envelope detection is also involved in what is maybe the most common acoustic to electric transducer: the (mammalian) ear. It is a long succession of physiological operations from acoustical pressure waves at our ears to us perceiving sounds.

But in that chain, one element is specifically responsible for the transduction from mechanical vibrations to electrical signals. The inner hair-cells (IHC), situated in the organ of Corti, respond to the vibration of the basilar membrane by sending electrical pulses to the auditory nerve. It is the collection of pulses from all of the IHCs traveling up the auditory pathway that will then stimulate our perception of the sound.

A simplified description of the mechanisms responsible for this transduction can be given as follows. Vibrations of the basilar membrane will cause deflections of the stereocilia, hair-like projections situated at one extremity of the IHC. Deflection in one way (but not the other way) will cause the IHC to depolarize, and send an electric pulse to the afferent nerve connections, at the opposite extremity of the IHC. This unidirectional response results in half-wave rectification of the input basilar membrane vibration. The cell however needs to repolarize before firing a new electrical pulse; there is an upper limit at which it can respond to oscillating vibrations. As a result, it presents low-pass characteristics. The output for one cell is a series of electrical pulses, with a probability of firing a pulse that is related to the rate at which the basilar membrane vibrates, though limited by repolarization time. However, a given input will cause many IHCs to respond. In a bundle of IHCs, the responses of individual cells will add up. Due to their probabilistic behavior, the sum of many individual responses will no longer be a series of pulses, but will resemble the input signal (the displacement of the basilar membrane with time). The half-wave rectification property as well as the low-pass characteristics of individual cells remain, and the afferent electrical signal at the output of a bundle of IHC can be modeled as the *envelope* of the input mechanical vibration.

This description is elementary, and an accurate model of all the underlying mechanisms would be much more complex. However, basic models of IHC behavior as a traditional envelope detector have been accepted in the literature. These models differ in the characteristics of the low-pass filter involved. As described earlier, the low-pass filter is designed to remove the carrier wave information and keep the envelope. Its design is easy and unambiguous when the frequency ranges of envelope and carrier wave are very distant, as is the case for AM radio. However in this case there is no clear definition of the envelope, hence the low-pass filter properties for the different IHC envelope models were chosen empirically by their authors. In chapter 3 and 4 of this thesis, three IHC envelope models will be used, with the following properties:

- Second order low-pass filter with cutoff frequency at 1000 Hz (Dau *et al.*, 1996a)
- First order low-pass filter with cutoff frequency at 800 Hz (Lindemann, 1986)
- Fifth order low-pass filter with cutoff frequency at 770 Hz (Breebaart *et al.*, 2001)

2.1.3 Mathematical and signal processing approaches

The Hilbert envelope

Envelope detection can be performed by half-wave rectification followed by low-pass filtering, but this definition is not unique as it involves a particular design for the low-pass filter. When there is a clear separation between envelope and carrier, as is the case for AM radio broadcasting, there is no ambiguity nor technical limitations faced when implementing envelope detection. However, this is not the case for audio signals where typical envelope frequency range (or modulation frequency range) and carrier wave frequency range are not clearly separated, and might even be overlapping (e.g., the IHC models presented above will extract envelopes with frequency contents that are far above the lowest audible frequencies). In that case, an ambiguity remains as to where to place the low-pass filter's cutoff frequency.

Gabor (1946) introduced the use of the Hilbert transform to extract the envelope of a signal. Unlike the half-wave rectification followed by low-pass filtering definition, the *Hilbert envelope* is a non-parametric definition, and therefore holds independently of the signal or the application. Hence, it was widely accepted as the canonical way of defining the envelope of a signal in mathematics and signal processing applications.

The Hilbert transform $\mathfrak{H}$ of a given real-valued signal s provides the signal $\widehat{s}$ in *quadrature* with s: the negative-frequency components of s are rotated around the complex plane by $\pi/2$ and the positive-frequency components of s by $-\pi/2$. Introducing the Fourier transform $\mathscr{F}$, the frequency f, the sign function sgn and the imaginary unit i, this formalizes as:

$$\mathscr{F}\{\mathfrak{H}(s)\}(f) = \mathscr{F}\{\widehat{s}\}(f) = -i \cdot \operatorname{sgn}(f) \cdot \mathscr{F}\{s\}(f) \tag{2.1}$$

This specific Fourier domain approach to the definition was employed in (Gabor, 1946), but the Hilbert transform is also commonly defined in the time domain by[1]:

$$\widehat{s}(t) = s(t) * \frac{1}{\pi t} = \int_{-\infty}^{+\infty} \frac{s(\tau)}{\pi(t-\tau)} d\tau \tag{2.2}$$

where $*$ denotes convolution. Gabor (1946) suggested to use the *analytic* signal defined as the complex-valued signal having s as real part and $\widehat{s}$ as imaginary part:

$$s_a = s + i\widehat{s} \tag{2.3}$$

Because of the quadrature relationship between real and imaginary part, the analytic signal "transforms an oscillating into a rotating vector" (Gabor, 1946). Using (2.1), the analytic signal is

[1] Note that the function $\frac{1}{\pi t}$ is not integrable. An accurate formulation involves taking the Cauchy principal value of the integral in (2.2).

easily defined in the Fourier domain:

$$\mathscr{F}\{s_a\}(f) = \begin{cases} 2\mathscr{F}\{s\} & \text{if } f > 0 \\ \mathscr{F}\{s\} & \text{if } f = 0 \\ 0 & \text{if } f < 0 \end{cases} \tag{2.4}$$

The Hilbert envelope is then taken as the magnitude of this "rotating vector":

$$\operatorname{env}(s) = |s_a| = |s + i\widehat{s}| \tag{2.5}$$

The Hilbert envelope was later extensively studied in (Dugundji, 1958). Although in many scenarios, it fits to the conceptual approach of the envelope given in 1.1.1 (as examples, the envelopes in fig.1.2(a-c) were not "hand-drawn" but were actually Hilbert envelopes), it exhibits some counter-intuitive limitations:

- **The Hilbert envelope of a bounded signal may be unbounded**. Consider for example the simple step function s, $s(t) = 0$ if $t < 0$ and $s(t) = 1$ if $t \geq 0$. According to (2.2) and the definition of convolution, the Hilbert transform at $t = 0$ is given by the integral

  $$\widehat{s}(t=0) = \int_0^\infty -\frac{1}{\pi\tau} d\tau = -\infty$$

  Hence the envelope of $s(t)$ is infinite at $t = 0$. In general, continuous-time signals that presents a discontinuity will have an infinite Hilbert envelope at the discontinuity point. In practice, signals are discrete-time so the concept of discontinuity does not exist. Their envelope will not be unbounded, but will present an unexpected large peak if there is a large amplitude difference between two consecutive samples. Note also that a discontinuity in the signal implies a broad-band signal, and the issue presented here could be related to the envelope not being adequately defined for broad-band signals (see section 1.1.4).

- **The Hilbert envelope of a bandlimited signal may not be bandlimited**. This was mentioned in (Dugundji, 1958), along with the proof that the *squared* envelope of a bandlimited signal is, however, bandlimited. An example is given in fig.1.2(a) where the signal is bandlimited (two sine wave of 45 and 55 Hz) but the envelope, as the absolute value of a cosine (see (1.3)), presents discontinuities in its derivative at every period. Such discontinuities implies that it is not bandlimited. The squared envelope however is a squared cosine which is bandlimited.

Although they presented an ambiguity in their definitions, envelopes based on IHC models do not present such limitations, due to the low-pass filter they involve. There is a compromise in the choice of one of these two methods, as the Hilbert envelope has an unambiguous definition but IHC-based envelopes behave more naturally.

Further developments on envelope definition

Further studies have been conducted since the work from (Gabor, 1946) and (Dugundji, 1958) aiming at developing new definitions of the envelope that would not exhibit the limitations of the Hilbert envelope.

In a study on a generalized concept of the analytic signal, Vakman (1996) investigated the class of complex-valued signals whose real part is the input signal and imaginary part is given by any arbitrary operator **H** applied to the signal. Vakman (1996) stated three "reasonable physical conditions" that the envelope (i.e., the magnitude of such complex-valued signals) should satisfy:

1. A small perturbation in the input signal should cause only a small perturbation of the envelope (**H** must be continuous).
2. Scaling the signal by a positive value should not alter the carrier (**H** must be homogeneous).
3. A sinusoid should have a constant envelope.

He then proved that these three conditions were fulfilled if and only if **H** was the Hilbert transform, i.e., for the Hilbert envelope. Hence, alternative envelope definitions based on such a complex-valued signal will violate one of the three conditions above.

Loughlin and Tacer (1996) followed the approach taken in Vakman (1996) and defined four reasonable assumptions to be satisfied by the envelope. The two last conditions are the same as Vakman's conditions 2 and 3, but he replaced condition 1 by the following two:

- Boundedness of the magnitude: If the signal is bounded in magnitude then the envelope should also be bounded (but not necessarily by the same bound).
- Bandlimitedness: If the signal is bandlimited then the carrier signal should be bandlimited within the same band as the signal.

As was pointed out earlier and further explained in (Loughlin and Tacer, 1996), the Hilbert envelope violates these two conditions. They proposed a new method based on estimating the instantaneous frequency of the carrier wave in a signal as the first moment of its time-frequency distributions along the frequency axis. The envelope is then obtained by deconvolution of the signal by the carrier wave, and was shown to fulfill the four conditions the authors suggested.

In (Cohen *et al.*, 1999), the authors acknowledged the three conditions given by Vakman (1996) but noted that given the analytic signal, it was the non-negativity constraint of the envelope that was responsible for the limitations stated in (Loughlin and Tacer, 1996). They proposed to still use the Hilbert transform but to allow the envelope of a signal to become negative; adding to condition 1 in (Vakman, 1996) that not only the envelope be continuous, but also the phase of the carrier wave. For example, for the combination of two tones given in (1.1), the envelope would be defined as the cosine component and not the cosine's absolute value as in (1.3). Li and Atlas (2004) proposed an implementation of envelope detection based on these considerations, yielding an envelope that could take negative values. They showed how such a method allowed smoother and more realistic estimations of the instantaneous frequency of a signal (the derivative of the phase of the carrier wave) in over-modulated signals (i.e., signals where the envelope crosses the zero-line, as the two-tones combination in (1.1)).

In a following study, Atlas *et al.* (2004) described how a real-valued envelope, even when allowed to be negative, was "too restrictive" and introduced the need of a complex-valued envelope. In later work (Li and Atlas, 2005; Schimmel and Atlas, 2005) this concept was further developed in the scope of modulation filtering, presenting how such a complex-valued envelope could be manipulated before being recombined with the original carrier wave. As introduced in (Ghitza, 2001), combining a manipulated envelope with the original carrier wave is prone to yield signals that do not present the expected manipulations in their envelope. This is due to unwanted interactions between the manipulated envelope and the carrier wave in the recombination. To assess this effect, Ghitza (2001) proposed the *projection test*: a given manipulation framework would have a “valid” design if the envelope re-extracted from the recombined signal (i.e., the output signal) would equal the manipulated envelope (i.e., prior to recombination). Inspired by this approach, Clark and Atlas (2009) proposed two conditions for modulation filtering to be successful: the processed signal should have the same bandwidth as the original signal, and the carrier must be redetectable, i.e., the carrier extracted from the processed signal should be the same as the original signal’s carrier (which amounts to the projection test from (Ghitza, 2001)). They present a framework to perform modulation filtering for both stationary and non-stationary signal, showing that it approximately follows their first condition, but fails for the second. This second condition is then circumvented by retaining information from the original carrier wave.

A more recent study by Sell and Slaney (2010) proposed an optimization framework to perform demodulation, i.e., to estimate from a time-domain signal s a suitable decomposition into the product of a modulator m and a carrier c. They propose two approaches, each with their own advantages and limitations. One approach is in the logarithmic domain, where the product $m \cdot c$ is conveniently translated to a sum of logarithms. They suggest to minimize the sum of individual cost functions of the logarithms of m^2 and c^2, where the squaring is intended to avoid logarithms of negative numbers. The product of m and c resulting in the signal to be demodulated, s, is expressed as a constraint in the minimization process. One advantage of this logarithmic domain approach is that this constraint is convex. They then define cost functions for both the modulator and the carrier that are convex, and that respectively (i) penalizes abrupt changes in the derivative of the modulator (i.e., minimizes inflexion points, or the second-order derivative) and (ii) encourages sparsity of the carrier in the frequency domain (that is following their assumption that the carrier is harmonic). Similarly, they propose another approach, in the linear domain, where formulating a convex optimization problem is more challenging, but where cost functions can be devised more easily, and can be based directly on the spectrum of the modulator (e.g, by penalizing high frequency content in the modulator). Among other interesting results, Sell and Slaney (2010) show how their logarithmic domain approach is very good at estimating modulators that contain zero-crossings (i.e., phase inverting modulators), for which the linear domain approach does not perform as well since it only accounts for non-negative modulators. Conversely, their linear domain approach is very robust and accurate in conditions where the carrier is stochastic (uniformly distributed noise), as well as for speech at the output of a filterbank. Importantly, Sell and Slaney (2010) acknowledge that complex-valued modulators, such as proposed by Atlas *et al.* (2004), might be *mathematically* appropriate for solving the demodulation problem, but question their *perceptual* relevance, as they do not track the real amplitude of the signal. This is a point that we will get return to in section 7.2.

2.2 The spectrogram as a faithful signal representation

As was mentioned in chapter 1, a given envelope does not contain a sufficient amount of information to recover the original signal it is associated with. In other terms, the envelope extraction is not an injective operation. However, it was introduced how it is reasonable to consider whether the spectrogram (i.e., a multi-channel envelope representation) could be an injective representation of the signal, and hence that a signal could be recovered from its spectrogram only. This section presents previous results that are in favor of such an hypothesis. They will not be presented in a chronological order, but rather organized in subsections of increasing complexity in the definition of the spectrogram, as they assess problems that are different in nature.

2.2.1 Single channel envelope

Although simple examples were sufficient to prove that in general an envelope does not represent a unique signal, some interesting results were achieved in the past. In an early study, Licklider and Pollack (1948) constructed a test signal from an input signal by *infinite peak clipping*, i.e., the signal is infinitely amplified, and clipped at a value of ± 1. The amplitude of the test signal jumps from -1 to 1 at the zero crossings of the input signal. Although such signals would seem to contain no information in the envelope but only in the TFS, they remain intelligible. It is known today that this method does not in fact remove envelope information, as a regular band-pass filtering like the one performed by the auditory system will reconstruct almost all of the original envelope (e.g., Zeng *et al.*, 2004).

In connection with the field of optics, Gerchberg and Saxton (1972) constructed the first iterative algorithm to reconstruct a complex-valued signal from its magnitude and the magnitude of its Fourier coefficients. The algorithm works by alternating projections back and forth between time and frequency domain to estimate the missing phase signals in both the time and frequency domains. At each iteration, the phase is kept but the magnitude is reset to its known value. The authors also provided in a proof of concept how the reconstruction error had to be monotonically decreasing with increasing number of iteration. The algorithm has been widely used to reconstruct images from diffraction patterns.

2.2.2 Short-time Fourier transform with Gaussian windows

The short-time Fourier transform (STFT) of a continuous time signal $s(t)$ is defined mathematically as:

$$V_g s(\tau, \omega) = \int_{-\infty}^{\infty} s(t)\, g(t-\tau)\, e^{-i\omega t} dt, \quad \tau, \omega \in \mathbb{R}, \tag{2.6}$$

where g is the window function that determines the resolution in time and in frequency.

Magnitude and phase interdependency for Gaussian windows

In this section we shall only study STFTs computed using the Gaussian window $\varphi(t) = e^{-\pi t^2}$. The STFT with a Gaussian window has very special properties. It has been known since Bargmann (1961) that the STFT with the Gaussian window φ multiplied by a fixed function is an *entire* function[2] no matter what the input signal is. As an entire function, the Cauchy-Riemann equations for the complex logarithm hold, and provide an explicit relationship between magnitude and phase of the STFT, shown in Chassande-Mottin *et al.* (1997):

$$-\frac{\partial}{\partial \tau} \angle V_\varphi s(\tau, \omega) = \frac{\partial}{\partial \omega} \log \left| V_\varphi s(\tau, \omega) \right|, \tag{2.7}$$

$$\frac{\partial}{\partial \omega} \angle V_\varphi s(\tau, \omega) - 2\pi\tau = \frac{\partial}{\partial \tau} \log \left| V_\varphi s(\tau, \omega) \right|. \tag{2.8}$$

The terms on the left hand side are the derivatives of the phase of the STFT of the signal. The first term is commonly known as the *instantaneous frequency*. The second term is sometimes known as the *local group delay*. In Flanagan and Golden (1966) it was shown that the instantaneous frequency provides a suitable representation for manipulating the signal in various ways with a minimum of distortion.

The equation (2.7) shows that for a Gaussian window, there are two possible ways of calculating the instantaneous frequency: by computing the time derivative of the phase of the STFT (as done for the original phase vocoder by Flanagan and Golden (1966)) or by computing the frequency derivative of the logarithm of the absolute value of the STFT (as proposed in Chassande-Mottin *et al.* (1997)). Since we have two different methods for computing the instantaneous frequency, the following procedure should allow the recovery of the phase from the magnitude of the STFT:

1. Compute the (real valued) log of the magnitude of the STFT.
2. Compute the partial derivative with respect to frequency of the result.
3. Integrate the result with respect to time.

As the boundary conditions in the integration (step 3) are unknown, the phase can be restored up to a global phase shift. This is no surprise, as the absolute value of the STFT of a signal will not change if the signal is multiplied by a complex number with magnitude of 1.

[2] An entire function is a function that is complex differentiable over the whole complex plane.

These considerations hold for continuous-time signals, but approximations of the derivatives involved in (2.7) and (2.8) by finite differences between samples or between channels can be done for discrete-time signals, given that the sampling rate and the number of channels are sufficiently large. These results provide evidence that the phase of the STFT computed with Gaussian windows can be recovered from the magnitude and vice-versa, hence that the original signal can be recovered from any of them, up to a global phase shift factor.

Magnitude of the STFT and Hilbert envelope

Although we are concerned with the relationship between multi-channel envelope and its associated time-domain signal, the previous result that relates the magnitude of the STFT to the signal is of high relevance. There is indeed a close relationship between the Hilbert envelope and the magnitude of the STFT. The definition of the STFT in (2.6) takes a form very similar to a convolution. Assuming that the window g is symmetric (if it is not, then it can be replaced with its time-reversed version in the following), it can be rewritten as follows:

$$V_g s(\tau,\omega) = e^{-i\omega\tau} \cdot (s * \widetilde{g_\omega})(\tau) \tag{2.9}$$

where $\widetilde{g_\omega}(\tau) = g(\tau)\, e^{i\omega\tau}$ corresponds to the original window modulated by a complex exponential and $*$ denotes convolution. The magnitude of the STFT therefore yields

$$|V_g s(\tau,\omega)| = |(s * \widetilde{g_\omega})(\tau)| \tag{2.10}$$

Assuming that the signal s is not band-limited, the bandwidth of $(s * \widetilde{g_\omega})(\tau)$ is given by the bandwidth of the modulated window. As multiplication by a complex exponential in the time domain translates to a frequency shift in the Fourier domain, the support[3] of the modulated window in the Fourier domain is the interval $[\omega - \Omega, \omega + \Omega]$, given that the support of the original window g is $[-\Omega, \Omega]$. Hence, for channels with sufficiently high center frequency ω (i.e., for $\omega \geq \Omega$), $(s * \widetilde{g_\omega})(\tau)$ will have no negative frequency content, and according to (2.4) will be an analytic signal.

If the window used for the computation of the STFT is band-limited by Ω, then the magnitude of the STFT in channels with center frequency larger than Ω will correspond to the Hilbert envelope of the corresponding channels prior to their demodulation (i.e., env $\left(e^{-i\omega\tau} \cdot V_g s(\tau,\omega)\right)$). It is common that filterbank implementations of the STFT provide such non-demodulated outputs. In that case, there is an equivalence between Hilbert envelope and magnitude in higher frequency channels.

The Gaussian window used in the previous example is not, strictly speaking, band-limited. But its energy is very concentrated around the origin and for STFTs computed with a Gaussian window, it is reasonable to state that the results given in Chassande-Mottin *et al.* (1997) mean that, for

[3] The support of a function or signal is the domain on which it is non-zero. In the Fourier domain, concepts of support and bandwidth are similar.

practical purposes, the signal can be recovered from the collection of the Hilbert envelopes of the channels of the STFT.

2.2.3 General redundant linear system with Hilbert envelope

Similarly to the reformulation of the STFT in section 2.2.2, a general filterbank operation can be seen for each channel as a convolution of the original signal with the impulse response of the filter in this channel. For discrete-time signals, it is formulated as follows:

$$(V_g s)_{m,n} = \sum_{k=1}^{L} s[k]\, g_m[n-k] \tag{2.11}$$

If the filters g_m are modulated gaussian windows then this representation is the STFT "without demodulation" described earlier. This formulation is therefore a generalization to any type of window. In the general case, the equivalence between the magnitude of such coefficients and the Hilbert envelope of each channel holds as well for channels with high enough center frequency. Moreover, if the impulse responses g_m are designed to be analytic signals (which will be the case for the filterbank implementations in the following chapters), then this equivalence will always hold, and the multi-channel Hilbert envelope of the signal s will be the magnitude of the coefficients $(V_g s)_{m,n}$. The magnitude of the coefficients $(V_g s)_{m,n}$ is what we refer to as spectrogram of s in the following.

Evidence of an injective representation

A very general result for finite, discrete systems has been shown in (Balan *et al.*, 2006). Consider a linear system given by a complex matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$:

$$c[j] = \sum_k \mathbf{A}_{j,k} s[k], \tag{2.12}$$

where $s \in \mathbb{R}^N$ is the input signal and $c \in \mathbb{C}^M$ are the output coefficients. The filterbank given by (2.11) is such a system. Balan *et al.* (2006) showed that if $M \geq 4N-2$ (i.e., if the system produces 4 times or more as many output coefficients as it takes input coefficients) then the mapping $s \longmapsto |c|$ was injective for a generic[4] system $\mathbf{A}$, up to a global phase factor. This means that given a matrix $\mathbf{A}$ there exists a non-linear reconstruction method $\text{reconstruct}_{\mathbf{A}}$ such that

$$s_r = \text{reconstruct}_{\mathbf{A}}(|c|), \tag{2.13}$$

and

$$s_r = e^{i\phi} s, \tag{2.14}$$

[4] Here, *generic* means that it holds *almost everywhere*, i.e., the mapping will be injective with a probability of 1 for an arbitrary $\mathbf{A}$, though some systems can be constructed for which the mapping is not injective.

for some constant $\phi \in [0;2\pi]$. Applied to filterbanks this result proves that if the filterbank has more than 4 times as many filters than its decimation rate, then the spectrogram representation is injective, i.e., that a given spectrogram is associated to a unique signal (and globally phase-shifted versions of it). Conversely, it should be possible to find a unique signal that is associated to a given spectrogram.

Spectrogram reconstruction methods

The conclusions from (Balan *et al.*, 2006) proved the existence of a solution to the problem of spectrogram inversion but did not provide a reconstruction method. Although this result is relatively recent, the interest in reconstructing a signal from its spectrogram originated much earlier.

Griffin and Lim (1984) adapted the single channel algorithm from (Gerchberg and Saxton, 1972) to a multiple channel representation, providing the first method to reconstruct a signal from its spectrogram (i.e., the magnitude of its STFT). Like the algorithm from (Gerchberg and Saxton, 1972), the method iteratively projects the signal back and forth between time and time-frequency domains. The STFT of a first estimate of the signal is computed, yielding a magnitude and a phase. The magnitude is discarded, but the phase estimate is combined with the provided magnitude (i.e., the given spectrogram) and inverted back to the time domain by inverse STFT, yielding a new signal estimate. By iterating this procedure, Griffin and Lim (1984) showed that the reconstruction error, measured as the distance between the STFT magnitude at a given step and the spectrogram to reconstruct, was monotonically decreasing.

Many following studies revolved around the algorithm proposed in (Griffin and Lim, 1984), increasing its accuracy in some ways while still being based on the same approach of repeated projections between domains. Sturmel and Daudet (2011) provided a review of such methods and expanded on their common limitations. Some spectrogram reconstruction methods however based on new approaches have also been suggested. Achan *et al.* (2004) proposed reconstruction of time-domain speech signals from their spectrograms using a probabilistic model of speech and searching for the signal maximizing the likelihood of the spectrogram. Bouvrie and Ezzat (2006) suggested a root-finding algorithm for reconstructing the signal at each position of the STFT window in time, using smoothness between neighboring segments as an additional constraint.

Recent methods complemented the existence result from (Balan *et al.*, 2006): the PhaseLift (Candes *et al.*, 2011) and PhaseCut (Waldspurger *et al.*, 2012) algorithms allowed for perfect recovery of a signal from the magnitude of its projections through a finite, linear system, up to a global phase factor. Sun and Smith (2012) applied the PhaseLift approach to the recovery of a time-domain signal from its spectrogram. Their method relies on estimating a matrix S obtained by the outer product of the sought signal s with itself: $S = ss^{\mathbf{T}}$, $s^{\mathbf{T}}$ being the transpose of s. This reformulation, although drastically increasing its dimensionality, results in a convex problem, mathematically simpler to solve. This method allow for very accurate recovery, but does not scale well to real world signals, as it attempts to estimate a matrix of size $L \times L$, where L is the length of the signal, making it impractical for signals with more than 100 samples.

2.2.4 Simple auditory models

The peripheral auditory system performs an operation that is to some extent similar to a spectrogram extraction. The mechanical properties of the basilar membrane vary along its length such that only localized sections will respond to a narrow-band excitation. In effects, it acts as a bank of bandpass filters. As introduced in 2.1.2, the transduction of the mechanical vibrations of local sections of the basilar membrane to electrical signals is performed by the inner hair-cells (IHCs), whose behavior can be modeled as an envelope extraction. A holistic, simplified model of the transduction from mechanical signal to electrical signal in the cochlea can be found in spectrogram extraction.

The STFT however is not a good model of the peripheral bandpass filtering of the cochlea, as it produces channels with equally spaced center frequencies. Moreover, the filters in the STFT are all derived from a base window which is modulated by the center frequency of the channel and therefore all channels present the same bandwidth. In contrast, Glasberg and Moore (1990) suggested that the spacing in frequency between auditory filters increases exponentially for increasing center frequency and that the filters bandwidth increases proportionally to this center frequency: at low frequency they are narrow and close to each other but broader and spaced further apart at higher frequencies. A well accepted model of the human auditory filterbank is given by Gammatone filterbanks (Patterson *et al.*, 1988). Such filterbanks can be implemented as a linear system described in (2.11).

As mentioned in 2.1.2, the IHCs extract the envelope of the mechanical signal, but not the Hilbert envelope. Discrepancies between IHC and Hilbert envelopes are significant enough that the previous results in section 2.2.3 do not to apply. Further, applying the Hilbert envelope at the output of a filterbank is not a good auditory model. However, some results of reconstruction from spectrogram obtained from IHC envelope at the output of a filterbank have been obtained.

Slaney *et al.* (1994) proposed an algorithm to recover a time-domain signal from a cochleogram, i.e., a time-frequency representation of the sound that mimics the analysis done in the cochlea. His cochleogram is the equivalent of the half-wave rectified output from a filterbank. He showed how the half-wave rectification, when carried out in a multiple channels representation, was invertible by alternating projections, in the fashion of (Griffin and Lim, 1984). As half-wave rectification amounts mostly to adding harmonics in the frequency domain, he also suggested band-pass filtering as a method to invert it. Along with half-wave rectification, common models of IHC envelope extraction usually involve low-pass filtering, which is not considered in (Slaney *et al.*, 1994).

2.3 The role of envelope in psychoacoustics

As introduced in the tuning forks example from fig.1.1, the decomposition of a signal into the product of an envelope and a carrier was mainly motivated by human perception. Hence it is quite logical that the concept of the envelope played an important role in the history of psycho-acoustics. Many studies have attempted to relate perceptual properties to attributes of either the envelope or the fine structure (i.e., the carrier). A thorough listing of the contributions to the field of psycho-acoustics that involved the concept of envelope or fine structure would be extensive, and well beyond the scope of this thesis. Instead, some of the main studies that involved manipulation of the envelope of speech, and how such manipulations were limited by attributes of the cochlea will be presented.

Recreating speech signals from envelope-based information originated with the work of Dudley (1939). He devised a system that could "code" the voice using a few low-frequency control signals and recreate artificial speech from it. He named the apparatus a *vocoder*, a voice coder. Although not mentioned under these terms, the main operation of Dudley's vocoder was to extract the envelope by rectification and low-pass filtering in ten narrow-band channels. These envelopes, or "control" signals, could then be used to modulate narrowband noise generators, recreating artificial though intelligible speech. Two types of noise generators, a harmonic "buzz" and an inharmonic "hiss", were used and switched on the fly to respectively recreate voiced or unvoiced speech. Dudley (1939) noted that using his apparatus, "not only can the speech be remade to simulate the original but it can be changed from the original in a variety of ways" suggesting that it allowed for basic envelope-based sound processing.

Developments on the vocoder technique were later brought by Flanagan *et al.* (1965) who proposed the use of the instantaneous frequency to derive sub-band modulators. The instantaneous frequency is given by the derivative of the phase with respect to time in a given sub-band, giving its name, *phase vocoder*, to the method. For each sub-band, the envelope and the changes in instantaneous frequency along time are computed, and both are low-pass filtered. Signals are then generated from the sum of the synthesized sub-bands, where each sub-band is obtained from a sinusoid at the band's center frequency which is amplitude modulated and frequency modulated respectively by the low-passed envelope and smoothed instantaneous frequency. Flanagan showed that speech could be conveyed this way, with cutoff frequencies for the low-pass filters as low as 20 cycles per second, suggesting a significant reduction in terms of data bandwidth. Additionally, simple applications of the phase vocoder are found in time-stretching the signal without influencing its frequency content, or conversely pitch-shifting without affecting its temporal structure. The concept of instantaneous frequency was also used in studies involving refinement on the definition of Hilbert envelope presented in the previous section 2.1.3.

Drullman *et al.* (1994) investigated the effect of temporal modulation smearing (i.e., low-pass filtering of the envelope) on speech intelligibility. The Hilbert envelopes, extracted at the output of a filterbank, were low-pass filtered and recombined with their corresponding original carrier. Using their processed stimuli, the authors determined that the intelligibility was improving when increasing cutoff frequency of the low-pass filter, but only up to 16 Hz. Conditions with higher cutoff frequency did not show further improvement. For very low cutoff frequencies (2 Hz and below), poor intelligibility in quiet prohibited a standard speech-in-noise intelligibility test. However, intelligibility scores in quiet showed an improvement with increasing cutoff frequency that was significant only from 2 Hz. Overall, Drullman *et al.* (1994) concluded that the modulation frequency range of influence for speech intelligibility is in the interval between 2 and 16 Hz.

In a critique of the work of Drullman, Ghitza (2001) noted that the auditory system was partially recovering the original envelope of manipulated signals. This phenomenon is generally referred to as *envelope recovery*: such signals, when re-analyzed through a basic auditory model composed of a Gammatone filterbank and IHC envelope detectors, would yield envelopes which do not exhibit the low-pass behavior expected, but rather an unexpectedly close similarity with the envelopes of the original signal, suggesting that the modulation filtering approach has a very limited efficiency. Ghitza (2001) relates the envelope recovery in this context to two theorems stating respectively that (i) in a band-limited signal, the envelope and phase will be related and that (ii) if the cosine of a band-limited phase signal is used as input of a bandpass filter (hence, a flat, or constant-envelope input), the output will have a non-constant envelope that will be related to the phase signal. As a corollary, Ghitza (2001) points out to the band-widening properties of the Hilbert transform. As mentioned in section 2.1.3, the Hilbert envelope of a bandlimited signal is not necessarily bandlimited. In the processed signal, recombined channels will overlap in frequency and interact, such that re-analyzed envelopes will not be band-limited as would be expected from the low-pass modulation filtering. Ghitza (2001) presents an alternative processing scheme to circumvent this issue using dichotic presentation, where every second channel is presented to one ear, and the remaining channels to the other ear. This doubles the distance in frequency between neighboring channels, hence limiting the negative impact of band-widening properties of the Hilbert envelope. Higher processing stages in the auditory pathway then integrates information over the two ears but without yielding unwanted interaction between neighbor channels.

Smith *et al.* (2002) constructed *audiological chimaeras*, test signals obtained by combining the envelope of a given class of signals (either speech, music, or noise) with the TFS of another class. The signal processing is done by splitting the input signal into various numbers of frequency bands (1 to 64) ranging from 80 Hz to 8820 Hz. The frequency bands are equally wide on the cochlear frequency map with a constant overlap (in terms of Hz) determined by the smallest filter. The Hilbert transform is then used to extract the envelope. They came to the main conclusion that "the perceptual importance of the envelope increases with the number of frequency bands" in tasks such as speech reception or melody recognition. It is worthwhile to note that the mere concept of "audiological chimaeras" conflicts with the results from (Balan *et al.*, 2006): for sufficiently many channels (i.e., more than 4), it should not be possible to impose both an arbitrary envelope and an arbitrary TFS to a signal.

Zeng *et al.* (2004) later noted that the work in (Smith *et al.*, 2002) was influenced by the envelope recovery phenomenon introduced in (Ghitza, 2001). In (Zeng *et al.*, 2004), the authors used the recovered envelope to modulate noise in the same way as (Shannon *et al.*, 1995). They demonstrated that the high intelligibility measured in (Smith *et al.*, 2002) for stimuli built from noise envelope and speech TFS for few channels (1 and 2) was largely influenced by envelope recovery. This study, together with (Ghitza, 2001), popularized the concept of envelope recovery: follow-up studies on the role of TFS/envelope in speech intelligibility which involved similar manipulations (e.g.,Gilbert and Lorenzi (2006); Lorenzi *et al.* (2006); Sheft *et al.* (2008)) took precautions to limit envelope recovery and quantify its influence. For example, Gilbert and Lorenzi (2006) extended the study of (Zeng *et al.*, 2004) and concluded that envelope recovery played a major role in consonant identification only when the bandwidth of analysis filters in the processing framework was wider than 4 times the bandwidth of an auditory filter (i.e., if there were less than 8 channels in the 80 Hz to 8820 Hz frequency band).

Envelope recovery provides, for processing schemes involving only a low number of channels, a direct illustration that imposing a given envelope *and* TFS to a signal is problematic. In this thesis, it will be shown indirectly how this problem also extends to representations involving more channels: if a signal can be recovered to a high degree of accuracy from its spectrogram only (in terms of the envelope domain or time domain metrics that will be introduced in the following chapter), then any combination of the spectrogram of signal A with the TFS of signal B cannot result in a signal presenting exactly the sought spectrogram (nor TFS). The resulting signal will still contain information related to the discarded TFS of A and the envelopes of B, but a simple re-analysis (as is the case for envelope recovery) will not be sufficient to exhibit this interdependency.

3

Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations[†]

Envelope representations such as auditory or traditional spectrogram can be defined by the set of envelopes from the outputs of a filterbank. Common envelope extraction methods discard information regarding the fast fluctuations, or phase, of the signal. Thus, it is difficult to invert, or reconstruct a time-domain signal from, an arbitrary envelope representation. Here, a general approach to this problem is proposed, which iteratively minimizes the distance between a target envelope representation and that of a reconstructed time-domain signal. Two implementations of this framework are presented for auditory spectrograms, where the filterbank is based on the behavior of the basilar membrane and envelope extraction is modeled on the response of inner hair cells. One implementation is direct while the other is a two-stage approach that is computationally simpler. While both can accurately invert an auditory spectrogram, the two-stage approach performs better on time-domain metrics. The same framework is applied to traditional spectrograms based on the magnitude of the short-time Fourier transform. Inspired by human perception of loudness, a modification to the framework is proposed, which leads to more accurate inversion of traditional spectrograms.

3.1 Introduction

Constructing a time-domain signal from an arbitrary time-frequency representation is an interesting problem in mathematics and signal processing and has received significant attention (e.g., Balan *et al.*, 2006; Slaney *et al.*, 1994; Hayes *et al.*, 1980). A particular case in this class of problems is the inversion of a spectrogram (i.e., the squared short-time Fourier transform, STFT, magnitude). While this classic problem is well known, it remains a current topic of research (e.g., Sun and Smith, 2012; Sturmel and Daudet, 2011; Le Roux *et al.*, 2010; Balan, 2010; Beauregard *et al.*, 2005; Bouvrie and Ezzat, 2006). Most of these approaches invert time-frequency representations where the frequency axis is linearly sampled (e.g., STFT). However, auditory models based on human perception involve time-frequency representation where the frequency axis is logarithmically scaled.

[†] This chapter extends the study presented in (Decorsière *et al.*, 2011) and was submitted for publication to the IEEE transactions on Audio, Speech and Language processing.

Thus, for the generation of acoustic signals intended for human perception (e.g., test stimuli in auditory research) a more flexible framework to invert time-frequency representations is needed. In this study, we present an approach to this problem that is both informed and inspired by human auditory processing.

In humans, the transduction from mechanical vibrations to electrical impulses in neurons occurs in the cochlea. Mechanical vibrations are transmitted into the cochlea via the middle ear and cause the basilar membrane to vibrate. The mechanical properties of the basilar membrane vary along the length of the cochlea. At the base, the basilar membrane is narrow and stiff, resulting in a high resonant frequency. At the opposite end, the apex, the basilar membrane is wide and less stiff, resulting in a low resonant frequency. For a given input, the vibration along the basilar membrane will vary based on the tuning at each position. Conceptually, we can model this using a bank of bandpass filters with center frequencies and bandwidths that increase logarithmically. Situated atop the basilar membrane is the organ of Corti which contains inner hair cells (IHCs). These cells have stereocilia, small hair-like projections, which deflect in response to displacement of the basilar membrane. When the stereocilia are deflected in one (but not the other) direction, the IHCs become depolarized, leading to action potentials in afferent neurons. Conceptually, we can model this as a half-wave rectifier. After depolarizing, the IHC and afferent neurons must re-polarize. This imposes an upper limit to the frequency at which action potentials can be generated. This upper limit can be modeled as a low-pass filter, and, when applied after half-wave rectification, performs envelope extraction. Thus, as a first approximation, we can model the transduction in the cochlea as envelope extraction of the outputs from a filterbank.

In this study, we focus on a particular class of time-frequency representation, which we term the envelope representation of a signal and define as the set of envelopes of individual narrow-band channels at the output of a filterbank. For example, the classical spectrogram (i.e., STFT magnitude) can be considered as an envelope representation. An auditory spectrogram (e.g., Dau *et al.*, 1996a), where the filterbank and the envelope extraction is inspired by human auditory processing, is another such example. Previous work has demonstrated that the envelope representation of a signal plays an important role in human perception. For example, the envelope information from only a handful of bands can be sufficient for speech intelligibility (e.g., Shannon *et al.*, 1995; Smith *et al.*, 2002), and some models for predicting speech intelligibility rely on information derived from the envelope representations of the speech and noise signals (e.g., Steeneken and Houtgast, 1980; Jørgensen and Dau, 2011). Furthermore, faithful representation of the envelope has been shown to be crucial for the perception of complex sounds (e.g., Chi *et al.*, 2005).

This chapter, which expands the findings of (Decorsière *et al.*, 2011), describes a tool for reconstructing the signal from its envelope representation. The development of such a tool is important for two reasons. First, it allows researchers to directly modify the envelope representations of stimuli used in perceptual experiments. Second, with such a tool, it may be possible to develop novel speech-enhancement or noise-suppression algorithms based on envelope processing to improve speech intelligibility. In the following sections, a general framework for constructing a signal from a target envelope representation is presented. This general framework is developed with no specific assumptions as to which filterbank or envelope extraction methods are used. Using this general framework, two cases are developed based on two different envelope extraction methods. The first case is motivated by the human auditory system, with an envelope extraction model based on IHC activity. Two implementation approaches for this method are presented, a straightforward application of the devised framework, as well as a two-step process where the low-pass filter of the IHC envelope extraction is inverted and a signal is reconstructed from the half-wave rectified filterbank output, as suggested by Slaney (1995). The second case is based on the STFT magnitude representations, the term-by-term square root of the "traditional" spectrogram. While this representation does not model auditory perception, our approach is still applicable, and can be used to reduce the perceptual consequences of reconstruction error. Results from these cases are evaluated and compared.

3.2 General framework

This section formalizes and suggests a general solution to the problem of reconstructing a time-domain signal from a given envelope representation. Here, the term envelope representation is used to denote the set of envelopes of narrow-band channels obtained at the output of a filterbank. This representation is therefore dependent on the choice of a given filterbank and a given envelope extraction method. An approach to formalizing the role of a filterbank is to define a filterbank operator $\mathbf{V}$ which is used jointly with a set of analysis windows $\{g_m\}_{1\leq m\leq M}$, forming the analysis operator $\mathbf{V}_g$. It operates on an input signal s with a finite duration of L samples as follows:

$$(\mathbf{V}_g s)_{m,n} = \sum_{k=1}^{L} s[k]\, g_m[an-k] \tag{3.1}$$

Here, $s[k]$ denotes the kth sample of signal s. In practice, with this definition, the output $(\mathbf{V}_g s)$ is a matrix. Each row, later denoted by $(\mathbf{V}_g s)_m$, corresponds to a different frequency channel output (i.e., a subchannel) and is indexed by m, with $1 \leq m \leq M$. The columns of this matrix span time and are indexed by n, with $1 \leq n \leq N$. The decimation rate of the filterbank is controlled by the parameter a, which represents the hop-size, in terms of samples of the original signal s, between two consecutive points in any given subchannel. Given that a suitable set of synthesis (or dual) windows $\left\{g_m^{(d)}\right\}_{1\leq m\leq M}$ exists, this representation admits an inverse, the synthesis operator $\mathbf{U}$:

$$s[k] = \mathbf{U}_{g^{(d)}}(\mathbf{V}_g s)[k] = \sum_{m,n} (\mathbf{V}_g s)_{m,n}\, g^{(d)}_{m,an-k} \tag{3.2}$$

Note that in the following, two different filterbanks will be used, and will have their own notation for the windows $\{g_m\}_{1\leq m\leq M}$ to avoid confusion.

Now, consider an envelope extractor function $E(\cdot)$ that operates on band-limited signals, the envelope representation is the set of envelopes of each subchannel $\left\{E\left((\mathbf{V}_g s)_m\right)\right\}_{1\leq m\leq M}$. As for the filterbank output, it is then convenient to adopt a matrix notation for the envelope representation, where each line represents a different channel, and the columns are for different samples in time:

$$\mathbf{E}_s = \begin{pmatrix} E\left((\mathbf{V}_g s)_1\right) \\ \vdots \\ E\left((\mathbf{V}_g s)_M\right) \end{pmatrix} \tag{3.3}$$

The matrix $\mathbf{E}_s$ is the envelope representation of the signal s, and is an $M \times N$ matrix of non-negative real coefficients. As common analysis filterbanks usually provide band-limited outputs centered at different frequencies, each line in this matrix representation provides information related to the frequency content of the input signal. Hence, this matrix provides a time-frequency representation of the signal s. For example, in the case of the filterbank being the short-time Fourier transform (STFT), and the envelope extraction being the squared magnitude function, this matrix would correspond to the "traditional" spectrogram of the signal s.

Given a target envelope representation of a signal $\mathbf{T}$, an $M \times N$ matrix of non-negative real coefficients, the reconstruction problem is then stated as follows: Find a signal s such that $\mathbf{E}_s = \mathbf{T}$, or alternatively, such that $\mathbf{E}_s - \mathbf{T} = 0$. Note that, in the following, this problem will not be solved exactly, i.e., we will find a signal s such as $\mathbf{E}_s - \mathbf{T} \approx 0$. Hence the term "reconstruction" is strictly speaking slightly inaccurate as perfect reconstruction is not achieved, and the approach taken is closer to the "synthesis" of a suitable time-domain signal. However the term "reconstruct" will be used as we felt it reflects well the step-by-step, iterative build up of a solution that is described in the following.

To better account for $\mathbf{E}_s - \mathbf{T} \approx 0$, it is convenient to define the real-valued function $\mathscr{G}$ that applies on any signal s with a length of L samples as follows:

$$\mathscr{G}(s) = \|\mathbf{E}_s - \mathbf{T}\|_{fro}^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} \left((\mathbf{E}_s)_{i,j} - (\mathbf{T})_{i,j} \right)^2 \tag{3.4}$$

The Frobenius norm $\|\cdot\|_{fro}$ is a matrix norm; hence $\mathscr{G}$ is the square of a norm-induced distance measure between the envelope of signal s and the target envelope $\mathbf{T}$. Therefore, the function $\mathscr{G}$ is positive-valued and equal to zero if, and only if, the matrices $\mathbf{E}_s$ and $\mathbf{T}$ are equal (i.e., $\mathscr{G}$ is positive definite). Hence, $\mathscr{G}$ reaches a global minimum when the signal s has the required envelope $\mathbf{T}$. This suggests an optimization approach where the problem is restated as follows: find s that minimizes $\mathscr{G}(s)$. In this approach, the function $\mathscr{G}$ is now referred to as the objective function. Using an iterative optimization algorithm, a minimum of this function can be found. A block diagram of the general procedure is illustrated in Fig. 3.1. Note that solving for $\mathbf{E}_s - \mathbf{T} = 0$ amounts to solving for the criterion proposed by Ghitza (2001) which states that the reconstruction is valid if the "recovered" envelope (i.e., the envelope extracted from the final signal, in our case $\mathbf{E}_s$) is the same as the envelope prior to reconstruction (or target envelope, $\mathbf{T}$).

The optimization process begins with a random initial signal estimate. Each iteration i starts by calculating the envelope representation of the current signal estimate, $\mathbf{E}_{s_i}$. This is compared to the target $\mathbf{T}$ using the objective function $\mathscr{G}$. The value of $\mathscr{G}$ and its gradient $\nabla\mathscr{G}$ are used to update

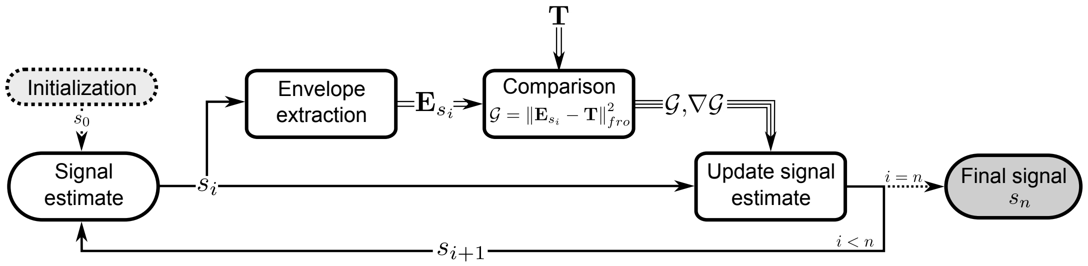


Figure 3.1: Block diagram of the processing scheme. At a given iteration i, the envelope representation of the signal estimate is extracted. Its distance to the target envelope $\mathbf{T}$ and the gradient of this distance provide information to update the signal to a new estimate. This process is then repeated for n iterations.

the signal estimate, resulting in a new estimate s_{i+1}. Thus, with each iteration, the optimization procedure generates an updated signal estimate that is "closer" (with relation to the distance $\mathscr{G}$) to the target envelope. The iteration process is terminated after n iterations.

From a practical perspective, it is important to note that numerical optimization methods often require knowledge of the first and sometimes second order derivatives of the minimized function (particularly when many dimensions are involved and "brute-force" search of the minimum becomes impractical). Although some algorithms can numerically estimate these derivatives, it is preferable to have an analytical expression to increase accuracy and reduce computational load. In the context of the present problem, the analytical expression of $\mathscr{G}$ will depend on how the filterbank is constructed and how envelope extraction is defined. The applicability of the method relies on being able to efficiently compute the gradient $\nabla\mathscr{G}$ of the function $\mathscr{G}$. In the following sections, we will present how this gradient can be efficiently computed for two particular cases: an auditory-motivated envelope representation (i.e., auditory filterbank and inner hair-cell envelope), and the traditional spectrogram (i.e., STFT squared magnitude). As an analytic expression for the derivative is needed, there may be some filterbank/envelope combinations that are not be compatible with our approach. For example, it might not be possible to analytically derive a gradient expression for envelope definitions that are based on the estimation of the instantaneous frequency of an associated carrier wave, as is done in some vocoder studies (e.g., in Flanagan *et al.*, 1965).

To speed up convergence and improve accuracy, information regarding the second-order derivative is also necessary. However, for a signal with a length of L samples, the matrix of the second-order derivative of $\mathscr{G}$, the Hessian matrix, contains L^2 elements. Thus, for typical speech signals, computing and storing all the elements of this matrix is often impractical, if not impossible. However, this can be overcome using a specific class of optimization algorithms, a limited memory Broyden-Fletcher-Goldfarb-Shanno (l-BFGS) algorithm (Liu and Nocedal, 1989). This algorithm only manipulates a sparse representation of the Hessian matrix consisting of a few vectors of size L instead of the full L^2 matrix.

By taking these practical considerations into account, implementing this optimization approach to reconstructing signals from their envelope representation becomes reasonable even on a standard computer, as will be described in section 3.5.

3.3 Reconstruction from IHC inspired envelope extraction

At the simplest functional level, two elements are required to generate envelope representations of a signal: an analysis filterbank to generate a time-frequency representation and an envelope extraction operator. For the analysis filterbank, many studies have employed the short-time Fourier transform (STFT; e.g., Griffin and Lim, 1984; Le Roux *et al.*, 2008; Sturmel and Daudet, 2011) because it produces a linear time-frequency representation of a signal. However, in our approach, we have opted to use a Gammatone filterbank, which provides a simplified model of the time-frequency analysis conducted by the human cochlea (Patterson *et al.*, 1988). Similarly, we use an envelope extraction operator that is based on IHC processing.

3.3.1 Gammatone filterbank

The Gammatone filterbank provides a simplified, linear model of basilar membrane motion. Unlike the linear spacing of frequency bins in the STFT, the center frequencies of the Gammatone filterbank are equally spaced on an Equivalent Rectangular Bandwidth (ERB) scale (see Glasberg and Moore, 1990, for further details). An individual Gammatone filter with center frequency f_c, bandwidth β, amplitude α and order n_f is given by its impulse response as follows:

$$\gamma(t) = \alpha t^{n_f-1} e^{-2\pi\beta t} \cos(2\pi f_c t) \tag{3.5}$$

Given a vector of ERB-spaced center frequencies, $\{(f_c)_m\}_{1\leq m\leq M}$, we obtain a set of filters $\{\gamma_m\}_{1\leq m\leq M}$ that forms a Gammatone filterbank and applies to a signal according to (3.1).

3.3.2 IHC inspired envelope extraction

As described in the introduction, we have based our envelope extractor on a simplified IHC model that consists of a half-wave rectifier followed by a low-pass filter. While similar envelope extractors are used in electronic circuits, the filter parameters that we have used are based on psychoacoustic data (e.g., Dau *et al.*, 1996a). Thus, to generate an envelope representation of a signal, the half-wave rectification and low-pass filtering are applied to the output from the Gammatone filterbank. Given the mth channel at the output of the Gammatone filterbank $\left(\mathbf{V}_\gamma s\right)_m$, the first step of the envelope extraction is half-wave rectification, which will be denoted by $\left(\mathbf{V}_\gamma s\right)_m^+$. This consists of setting all negative valued samples to zero while leaving positive-valued samples unchanged:

$$\left(\mathbf{V}_\gamma s\right)_{m,n}^+ = \begin{cases} \left(\mathbf{V}_\gamma s\right)_{m,n} & \text{if} \quad \left(\mathbf{V}_\gamma s\right)_{m,n} \geq 0 \\ 0 & \text{else} \end{cases} \tag{3.6}$$

Alternatively, we can introduce the Heaviside function $\mathscr{H}\{\cdot\}$. Given an input vector, this function returns a vector of the same size with values of 1 for indices where the input was positive and values of 0 for indices where the input was non-positive:

$$\left(\mathbf{V}_\gamma s\right)_m^+ = \mathscr{H}\left\{\left(\mathbf{V}_\gamma s\right)_m\right\} \cdot \left(\mathbf{V}_\gamma s\right)_m \tag{3.7}$$

In (3.7) and further equations, $\cdot$ denotes term-by-term multiplication of two vectors or matrices with the same number of elements. The second step of the envelope extraction is low-pass filtering. Assuming a low-pass filter with an impulse response h, the envelope of the mth channel is given by:

$$\left(\mathbf{E}_s\right)_m = \left(\mathbf{V}_\gamma s\right)_m^+ * h \tag{3.8}$$

Here, $*$ denotes convolution. Given a target envelope representation $\mathbf{T}$ computed according to (3.8), the reconstruction problem is to recreate the signal having $\mathbf{T}$ as envelope representation.

3.3.3 Direct reconstruction

Using the definition of the envelope from (3.8), the objective function is given by the following equation:

$$\mathscr{G}(s) = \|\mathbf{E}_s - \mathbf{T}\|_{fro}^2 \tag{3.9}$$

It can be seen from (3.1) that the derivative of the Gammatone analysis operator with relation to the kth coefficient of the input can be expressed as follows:

$$\frac{\partial}{\partial s[k]}\left(\mathbf{V}_\gamma s\right)_{m,n} = \gamma_m[n-k] \tag{3.10}$$

Combining (3.10) with (3.7), (3.8) and (3.9), and assuming the low-pass filter has a finite impulse response (i.e., an FIR filter), it is possible to express the gradient of $\mathscr{G}$ analytically. While typical IHC models do not use FIR filters, truncating the otherwise infinite impulse response is a reasonable approximation. If $h[k] = 0$ for $k > K$, then the gradient of the objective function can be expressed as follows:

$$\nabla\mathscr{G} = 2\sum_{k=0}^{K} h[k]\,\mathbf{U}_{\mathscr{T}^k\{\gamma\}}\left\{\left(\mathbf{E}_s - \mathbf{T}\right)\mathscr{T}^k\{\mathbf{H}\}\right\} \tag{3.11}$$

Here, $\mathscr{T}$ denotes the time-shift (translation) operator,

$$\left(\mathscr{T}\{s\}\right)[k] = s[k-1] \tag{3.12}$$

or similarly,

$$\left(\mathscr{T}^p\{s\}\right)[k] = s[k-p]\text{, for any integer } p \tag{3.13}$$

The matrix $\mathbf{H}$ represents the Heaviside function applied to all the channels at the output of the filterbank:

$$\mathbf{H}_{m,n} = \mathscr{H}\left\{\left(\mathbf{V}_\gamma s\right)_{m,n}\right\} \tag{3.14}$$

The gradient in (3.11) is expressed as a finite sum (of $K+1$ elements) under the assumption of the low-pass filter being an FIR filter. Each element of the sum is expressed using the filterbank synthesis operator $\mathbf{U}$ defined in (3.2) applied with a time-shifted version of the original filterbank *analysis* window. Importantly, note that direct knowledge of the *synthesis* windows $\left\{\gamma_m^{(d)}\right\}_{1\leq m\leq M}$ introduced in section 3.2 is not needed. This will also be the case for the gradient expressions in later sections where different envelope extraction schemes are used. As can be seen from (3.2), the operator $\mathbf{U}$ can be implemented using the fast Fourier transform (FFT), hence the gradient in (3.11) can be efficiently computed. Using (3.9) and (3.11), $\mathscr{G}(s)$ can be minimized with an iterative optimization procedure (l-BFGS algorithm).

3.3.4 Two-step reconstruction

The direct approach detailed above attempts to reconstruct a signal directly from a target envelope representation. However, it is also possible to process the envelope representation before applying the iterative optimization algorithm. Here, we propose a two-step reconstruction method inspired by (Slaney, 1995). Conceptually, the approach is straightforward. Recall that the IHC envelope extraction is modeled as half-wave rectification followed by low-pass filtering. Thus, if the inverse of the low-pass filter is applied to the target envelope representation, the result is the half-wave rectified output of the filterbank. The signal can then be reconstructed from this representation using the iterative optimization approach. Under the assumption that each channel has a narrow bandwidth, Slaney (1995) suggested that a bandpass filter should be used to remove harmonics introduced by the half-wave rectification. However, we propose a more global approach based on the general framework suggested in section 3.2 that takes the interactions between channels into account. This has the advantage of using the information from neighboring channels to recover the information lost in a given channel by the half-wave rectification.

Low-pass filter inversion and regularization

The low-pass filter impulse response h in (3.8) results in the filter response H in the frequency domain. The low-pass filtering is inverted by multiplying each channel in the frequency domain with the inverse filter response $1/H$. However, by definition, the response of a low-pass filter at higher frequencies is very small. Thus, a direct application of the inverse filter response would result in a very large unbounded gain being applied at high frequencies. This would introduce instability in the reconstruction, as any errors at high frequencies (e.g., rounding error) would be unreasonably amplified. Hence, it is necessary to regularize the inverse filter response by introducing an upper bound G_{max}, on the maximum gain allowed on the inverse filtering procedure. Given the mth channel of the target $(\mathbf{T})_m$ and the classic Fourier transform operator $\mathscr{F}\{\cdot\}$, the regularized inverse low-pass filtering generates the new target for this channel $(\mathbf{T}^+)_m$ as follows:

$$\left(\mathbf{T}^+\right)_m = \mathscr{F}^{-1}\left\{\mathscr{F}\{(\mathbf{T})_m\}\cdot\max\left(\frac{1}{|H|},G_{max}\right)e^{-i\angle H}\right\} \tag{3.15}$$

Here, the function $\max(\cdot)$ operates on individual coefficients of the vector $1/H$. The phase of the inverse response is maintained by multiplying with $e^{-i\angle H}$. The outcome of this step is the new target $\mathbf{T}^{+}$, where the superscript $(\cdot)^{+}$ suggests that this target corresponds to a half-wave rectified output of the filterbank. The regularization introduces inaccuracies in the representation, and in practice there is a tradeoff in the choice of the maximal gain G_{max}. Low gain results in good stability of the procedure but large inaccuracies. Alternatively, a high-gain limits the loss of information from the regularization, but at the cost of reduced stability. We have observed that, when increasing G_{max}, transients of large amplitude appear at the beginning and end (i.e., first and last few milliseconds) of the reconstructed signals. Increasing it further will eventually lead to global instability of the reconstruction scheme. This phenomenon can be used in an actual blind scenario, i.e., when the original signal is unknown, to adjust G_{max}, by increasing its value while monitoring these onset and offset transients. We have found that $G_{max} = 50$ dB is a suitable compromise for speech signals. However, G_{max} could be increased further for more stationary signals.

Half-wave rectification inversion

For the two-step approach, the reconstruction problem is to estimate a signal whose half-wave rectified output from the filterbank equals the target $\mathbf{T}^{+}$. This can be solved using the optimization approach proposed above, by defining the objective function as follows:

$$\mathscr{G}(s) = \left\| \left(\mathbf{V}_{\gamma} s\right)^{+} - \mathbf{T}^{+} \right\|_{fro}^{2} \tag{3.16}$$

With this formulation, the gradient is expressed as follows:

$$\nabla \mathscr{G} = 2\mathbf{U}_{\gamma} \left[\left(\left(\mathbf{V}_{\gamma} s\right)^{+} - \mathbf{T}^{+} \right) \cdot \left(\mathbf{V}_{\gamma} s\right)^{+} \right] \tag{3.17}$$

In comparison to (3.11), the gradient here has a simpler form and requires approximately K times fewer calculations. Thus, there is a clear advantage of this two-step approach in terms of implementation.

3.4 Reconstruction from STFT magnitude

Most of previous work concerning similar signal reconstruction has been conducted on the "traditional" spectrogram, i.e. an envelope representation given by the squared magnitude of the STFT coefficients. Because it provides a time-frequency representation with linearly spaced frequency channels, the STFT is not considered a good model for the auditory peripheral filterbank. However, our method is applicable and this case is considered here. We also propose a modification to reduce the perceptual consequences of reconstruction error.

The STFT with a window w and a hop-size a can be implemented as a filterbank, where individual filters have as impulse response the original window w modulated by a complex-valued exponential at a given channel frequency:

$$\tilde{g}_m[k] = w[k]\, e^{2\pi i f_m k} \tag{3.18}$$

The channel frequencies $\{f_m\}_{1 \leq m \leq M}$ are chosen such that they span the Nyquist domain and are linearly spaced. The number of channels, M, corresponds to the length in samples of the window w. Given these filters, the STFT is applied to an input signal using (3.1).

The envelope extraction in this case is the magnitude function:

$$\mathbf{E}_s = |\mathbf{V}_{\tilde{g}} s| \tag{3.19}$$

This definition could be used in (3.4) to form the objective function for the reconstruction of a signal from its STFT magnitude. However, to be consistent with the traditional definition of the spectrogram, we define the objective function from the squared magnitude as follows:

$$\mathscr{G}(s) = \left\| |\mathbf{V}_{\tilde{g}} s|^2 - \mathbf{T}^2 \right\|_{fro}^2 \tag{3.20}$$

With this definition, individual coefficients of the envelope contribute to the objective function with regard to their energy (i.e., squared magnitude). A convenient property of this definition is that the derivative of the squared magnitude function can be expressed as follows:

$$\left(|u|^2\right)' = 2\Re\left(\overline{u}u'\right) \tag{3.21}$$

Here, $(\cdot)'$ is the derivative, $\overline{(\cdot)}$ the complex conjugate, and $\Re(\cdot)$ the real part. Hence, by combining (3.21), (3.10) and later (3.2), the gradient of the objective function can be expressed using the Gammatone synthesis operator, but once again applied using the original analysis window:

$$\nabla\mathscr{G} = 4\Re\left(\mathbf{U}_{\tilde{g}}\left[\left(|\mathbf{V}_{\tilde{g}} s|^2 - \mathbf{T}^2\right) \cdot \mathbf{V}_{\tilde{g}} s\right]\right) \tag{3.22}$$

Optimizing the objective function $\mathscr{G}(s)$ as written in (3.20) will reduce the average error in the envelope representation of the reconstructed signal. However, for applications involving human listening, it is important to reduce the perceptual consequences of the error. A small error in a time-frequency region with otherwise little energy may be audible while an error with the same magnitude but in a region with high energy may not. To account for this compressive behavior, a modified objective function $\mathscr{G}_L(s)$ is proposed:

$$\mathscr{G}_L(s) = \left\| |\mathbf{V}_{\tilde{g}} s|^p - \mathbf{T}^p \right\|_{fro}^2 \tag{3.23}$$

If $p < 1$ in (3.23), the dynamic range of individual contributions to the objective function is reduced, which in effect increases the relative contribution of regions with lower energy. Because of this compressive behavior, we refer to $\mathscr{G}_L$ as the *compressed* objective function. In the following, a compression ratio of $p = 2/3$ was chosen, based on Stevens' power law for loudness (Stevens, 1957). Using this value, the contribution to the objective function of individual time-frequency bins is approximately proportional to their loudness (e.g., as modeled in Zwicker and Scharf, 1965; Moore and Glasberg, 1996). For an arbitrary p though, the gradient corresponding to this function is given as follows:

$$\nabla \mathscr{G}_L(s) = 2p\Re\left(\mathbf{U}_{\tilde{g}}\left[(|\mathbf{V}_{\tilde{g}}s|^p - \mathbf{T}^p)\cdot|\mathbf{V}_{\tilde{g}}s|^{\frac{p}{2}-1}\cdot\mathbf{V}_{\tilde{g}}s\right]\right) \tag{3.24}$$

which simplifies to (3.22) if $p = 2$.

3.5 Evaluation and comparison of techniques

3.5.1 Implementation, testing material and evaluation of convergence

To evaluate the proposed techniques, the general framework was implemented in Matlab. The Matlab implementation of the l-BFGS optimization algorithm was found in (Schmidt, 2005) and used with all default settings, except for one. The termination tolerance was reduced to avoid the algorithm from stopping prematurely. The reconstruction framework was tested using a speech corpus containing individual recordings of 70 English words, spoken by a female native speaker. The corpus is formed from the segmented keywords from the NU-6 WIN test (D. of Veterans Affairs, 2006). Results depend on the random initialization of the algorithm. Hence, when different methods are compared in the following, they will be initialized with the same random signal. A useful measure to compare algorithms introduced in (Sturmel and Daudet, 2011) is the spectral convergence $\mathscr{C}$ (a related measure was earlier proposed by Le Roux *et al.* (2010)). The spectral convergence measures the distance between target and reconstructed signals, in the time-frequency (STFT-magnitude) domain. It is the normalized Euclidean distance between the target envelope and the envelope of the reconstructed signal:

$$\mathscr{C} = \frac{\||\mathbf{V}_{\tilde{g}}s| - \mathbf{T}\|_{fro}}{\|\mathbf{T}\|_{fro}} \tag{3.25}$$

Thus, for reconstructions with little error (i.e., the spectrogram of the reconstructed signal is very similar to the target), $\mathscr{C}$ is small. Here, $\mathscr{C}$ is defined based on the STFT magnitude, even for evaluating reconstruction from other types of envelope. This allows for the comparison of results across different methods, and with results from other studies of reconstruction from STFT magnitude. Note that, in the case where $p = 1$ in (3.23), minimizing the objective function $\mathscr{G}$ is equivalent to minimizing $\mathscr{C}$.

3.5.2 Results from IHC inspired envelope representations

Although there are various models of the IHC envelope extraction documented in the literature, most use a similar structure and differ only with regards to the low-pass filter order and cutoff frequency (e.g., Dau *et al.*, 1996a; Lindemann, 1986; Breebaart *et al.*, 2001). Here, the IHC envelope extraction

Table 3.1: Results for IHC envelope representations.

Method	$\mathscr{C}$ (dB)	Time (s)	Iterations	RMS$_n$ (dB)
Direct	-27.7	581	62	-6.5
Two-steps	-24.1	10.1	28	-23.4

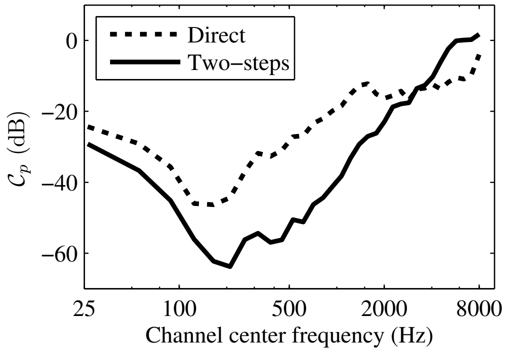


Figure 3.2: Partial spectral convergence $\mathscr{C}_p$ (in dB) plotted against center frequency of the channel for the direct and two-step approaches averaged over 70 speech signals.

model from (Dau *et al.*, 1996a) was selected. This model uses half-wave rectification followed by a second order butterworth low-pass filter with cutoff frequency at 1000 Hz. Time-domain signals were reconstructed from such envelope representations, using both the direct and two-step approaches to reconstruction described in section 3.3, and for the corpus of 70 words. Results, averaged over the whole corpus, are presented in table 3.1.

For both methods, the maximum number of iterations was set to 80, but the algorithm often stopped prematurely without being able to find a better solution. Hence, the average number of iterations and elapsed time are presented in table 3.1, along the averaged spectral convergence $\mathscr{C}$ expressed in dB. In terms of spectral convergence, the two approaches provide results of similar accuracy, with a small benefit for the direct approach. However, from a practical point of view, the two-steps approach has a clear advantage with much lower computation time and faster convergence (lower number of iterations).

In addition to the spectral convergence, and since we have knowledge of the original signal, it is possible to measure the root mean square (RMS) error. The normalized RMS error of the reconstructed signal s_r with relation to the original signal s, is expressed with the Euclidean norm $||\cdot||$ as follows:

$$\mathrm{RMS}_n = \frac{||s - s_r||}{||s||} \tag{3.26}$$

This assumes that the signals have the same number of samples, which is the case here. The RMS measures reconstruction errors in the signal domain, whereas the spectral convergence measures an error in the time-frequency domain. The RMS$_n$, averaged over the whole corpus, is presented in table 3.1. In terms of normalized RMS error, the two-steps approach performs far better than the direct approach.

To account for the discrepancies in the results between the two metrics used, we introduce a "partial" spectral convergence, $\mathscr{C}_p$, which is obtained in a similar way as (3.25) but by summing only across time, and not both time and frequency as the Frobenius norm does in (3.25). $\mathscr{C}_p$ therefore measures a normalized error for each frequency channel. Fig. 3.2 shows $\mathscr{C}_p$ as a function of channel center frequency, for the two approaches involved. Under this metric, it appears that the two-steps approach has a clear advantage over the direct approach up to 2 kHz. Its poorer performance above 3 kHz explains why this advantage is not reflected in the "total" spectral convergence $\mathscr{C}$. These results are further discussed in section 3.6.1.

3.5.3 Results for STFT magnitude

As with many studies presenting new or improved methods for spectrogram inversion (e.g., Le Roux *et al.*, 2010; Sturmel and Daudet, 2011; Beauregard *et al.*, 2005; Bouvrie and Ezzat, 2006; Le Roux *et al.*, 2008; Sun and Smith, 2012), we use the Griffin and Lim (1984) algorithm as a baseline to evaluate our method. In their algorithm, Griffin and Lim use an iterative approach that allows the reconstruction of a time-domain signal from only the magnitude of a Short-Time Fourier Transform (STFT). Therefore, this approach attempts to find the missing phase associated with this target magnitude. To do so, an initial estimate of the phase (e.g., a random phase) is combined with the target magnitude and the inverse STFT of the combination is computed. This provides an initial estimate of the desired time-domain signal. The STFT of this signal is then computed, providing a new STFT magnitude and phase. The magnitude is discarded and the new estimate of the phase is combined with the original target magnitude. The inverse STFT is computed again, leading to an updated estimate of the time-domain signal. This process is then iterated until some stop criterion is met. Griffin and Lim (1984) proved that the mean squared error of the STFT magnitude of the generated time-domain signal monotonically decreases with each iteration. In this way, for the particular case of using the STFT and Hilbert envelope, a signal whose envelope representation matches the target envelope can be generated.

As introduced earlier, the spectrogram is determined by a given window w, which determines the number of frequency points, and a hop-size a (sometimes given as a percentage of overlap between neighboring windows). As a thorough investigation of the reconstruction from STFT magnitude is beyond the scope of this paper, only one combination of these parameters was investigated: a 1024-sample Hann window, with $a = 64$ (i.e., about 94% overlap). The choice of a large amount of overlap was deliberate, to have a highly redundant representation such as the one obtained in 3.5.2 where the filterbank had no decimation (i.e., a hop-size of one).

Figure 3.3 presents the spectral convergence $\mathscr{C}$ measured in dB as a function of iteration number (left panel) and computation time (right panel) for the suggested method, with energy-based ($p = 2$) and loudness-based ($p = 2/3$) objective function, as well as for the Griffin and Lim (G&L) method for comparison. Spectral convergence and computation time were averaged over the 70 available speech signals. Note that computation times are to be compared in a relative way, as they are strongly influenced by the implementation and hardware used.

The proposed method with an objective function based on the spectrogram ($p = 2$) shows the

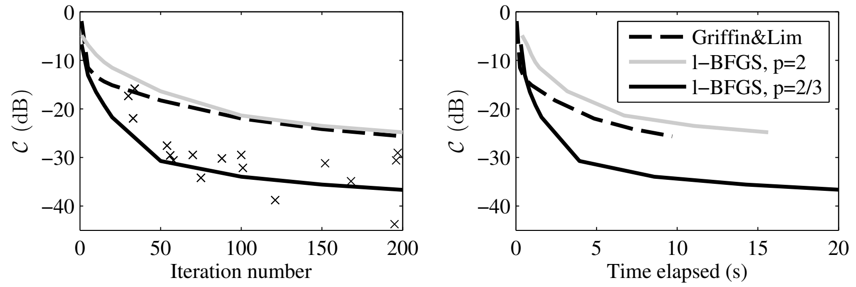


Figure 3.3: Spectral convergence $\mathscr{C}$ as a function of the number of iterations (left panel) and computation time (right panel), measured for the Griffin and Lim algorithm and the proposed algorithm (labeled l-BFGS) when $p = 2$ and $p = 2/3$. These results are averages obtained over 70 individual word tokens, with the same random initialization for the three methods. Individual crosses on the left panel represent results obtained for the l-BFGS approach ($p = 2/3$) when the algorithm failed to reach the requested number of iterations.

applicability of the method to the case of STFT magnitude: the algorithm converges to a solution similar (in terms of spectral convergence $\mathscr{C}$) to the G&L algorithm. However, the G&L approach shows a clear benefit in the first 100 iterations that is emphasized when plotted against computation time. Introducing the loudness-based objective function (i.e., when $p = 2/3$) substantially changes the results. The proposed method quickly outperforms G&L and maintains a reduction in spectral convergence of around 10 dB from 50 iterations on. Note that, despite the care taken to force the l-BFGS algorithm to stop at a requested number of iterations, the algorithm sometimes failed to find a better solution and stopped prematurely. In this case, the results are plotted as individual crosses. This occurred for 17 conditions out of 70. These points were not considered for the average in the right panel of Fig. 3.3 as they would bias the average computation time towards lower values.

In practice, these results were obtained using the `dgtreal` and `isgramreal` scripts from the linear time-frequency analysis toolbox (LTFAT) for Matlab (Søndergaard *et al.*, 2012) which implements both the G&L and the l-BFGS methods (for $p = 2$). The script was slightly modified to account for $p = 2/3$ and to allow for a user-provided initial phase, in order to provide the three methods with the same initial estimate.

3.6 Discussion

In this study, a general framework for reconstructing time-domain signals from an arbitrary multi-channel envelope representation was presented. The framework is based on minimizing the distance between the envelope of a signal and a target envelope, by means of a numerical optimization algorithm. Methods were developed for two common envelope definitions: envelope extraction based on an IHC model (i.e., half-wave rectification followed by a low-pass filter) and for the more common STFT magnitude. For both envelope definitions, it was possible to reconstruct a signal from a multi-channel envelope representation. The framework is general and can also be applied to other filterbank or envelope definitions, assuming that the filterbank output is computed according to (3.1), and that the gradient of the objective function defined in (3.4) can be efficiently computed. For envelope representations based on an IHC model of envelope extraction, two reconstruction approaches were considered. The first approach was a direct application of the general framework. It is possible to implement this under the reasonable assumption that the low-pass filters from the inner hair-cell model have a finite impulse response. The second approach used two-steps, where a regularized inverse filter was applied to the envelope representation and the filtered output was processed using a slightly modified version of the general framework. While it is possible to successfully reconstruct a signal using both approaches, the reconstruction error of the two-step approach was smaller.

3.6.1 Direct vs. two-step approach in reconstruction from IHC envelope

For both the direct and two-step approaches, the distribution of the reconstruction error across frequency (Fig. 3.2) was not uniform. The bandwidth of each channel of the Gammatone filterbank used in this method increases with center frequency. Since the half-wave rectification applied to each channel can be roughly seen as introducing harmonics in the signals, the higher the center frequency of a channel, the more the rectified subband will be attenuated by the low-pass filter from the IHC envelope extraction. Thus, it is more difficult to recover information from channels with high center frequency and reconstruction errors are expected to increase for higher frequency channels.

For the direct approach, errors presumably consist of estimation errors in both magnitude and phase, due to inaccuracies or round-off errors in the estimation of the gradient. For the two-step approach, however, the regularized low-pass filter inversion introduced some magnitude errors at very high frequencies but preserved the phase. As the RMS error is more sensitive to errors in phase than in magnitude, this could explain why the RMS error in the two-step approach was much lower than in the direct approach. The lower RMS_n value, along with a much reduced computational load, favors the two-step approach over the direct approach. However, the direct approach has the advantage that it does not require any tuning of parameters, such as G_{max} in (3.15). This may be advantageous when the original signal is unknown and highly non-stationary, such that the stability of the reconstruction cannot be easily assessed.

3.6.2 Loudness-based objective function for reconstruction from STFT magnitude

A reconstruction method was also developed and successfully implemented for the more common case of STFT magnitude envelope representations (i.e., square-root spectrogram). A human listener's perception of reconstruction errors is not a linear function of the absolute error magnitude. Typically, an error of a given magnitude in a "loud" segment of the signal might be inaudible (i.e., masked), whereas the same amount of error in a "quiet" segment might be perceptually significant. Thus, we proposed a compressed objective function where individual time-frequency bins of the envelope contribute with regard to their approximate loudness (i.e., raised to the power $2/3$, according to Stevens' power law for loudness, (Stevens, 1957)) instead of their energy (i.e., squared coefficients). Perceptually the benefit of using the compressed objective function is apparent when

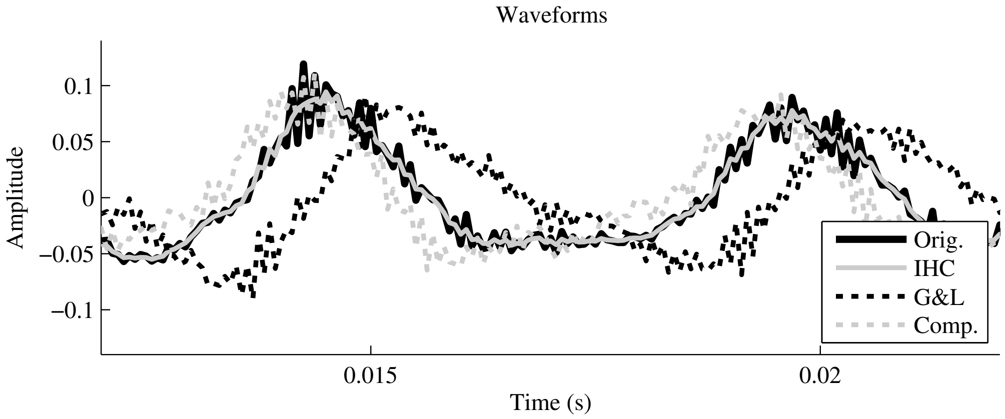


Figure 3.4: Illustration of the stagnation phenomenon: detail of waveforms of original and reconstructed signals from IHC envelope (IHC) as well as STFT magnitude for the Griffin and Lim algorithm (G&L) and the optimization algorithm with compressed objective function (Comp.).

considering the particular case when the target is obtained from a quiet yet audible signal that is embedded between two loud signals. When the compressed objective function is used, the quiet signal is reconstructed properly and audible. If the original objective function is used, this quiet signal disappears.

3.6.3 Reducing intrinsic limitations of reconstruction from STFT magnitude

As well as a perceptual benefit (reconstruction of quiet yet audible regions of the spectrogram), there appears to be a significant quality benefit when using a loudness-based objective function. The main limitation in reconstruction from STFT magnitude lies in what was originally referred to as stagnation by Fienup and Wackerman (1986): because the magnitude suppresses all information about the absolute phase, there can be a phase mismatch between local regions of the STFT of the original and reconstructed signals. Although this phenomenon may not alter perception significantly, it is visible on the waveforms of reconstructed signals: signals reconstructed from STFT magnitude can present a phase shift that does not remain constant over time.

To illustrate stagnation, Fig. 3.4 presents segments of the waveforms of original and reconstructed signals from various methods introduced in this paper. The signal reconstructed from IHC envelope (using the two-step approach in this case) is superimposed with the original signal. However, the two signals reconstructed from STFT magnitude (G&L and Comp.) show significant phase drift from the original, with the one obtained from G&L showing a larger deviation. To further illustrate the phenomenon, Fig. 3.5 presents the phase difference (modulo 2π) between the STFT of original and reconstructed signals for the two methods. In the left plot, the signal was obtained using G&L with 1000 iterations, which is sufficiently many to assume no significant further progression of the algorithm. For this specific case, a spectral convergence of -27 dB was measured. For the right plot, the proposed method, with loudness-based objective function was used. It stopped after 467 iterations, leading to a spectral convergence of -34 dB. The phase difference in the rightmost plot is smoother than in the center plot, suggesting that the use of the loudness-based function leads to

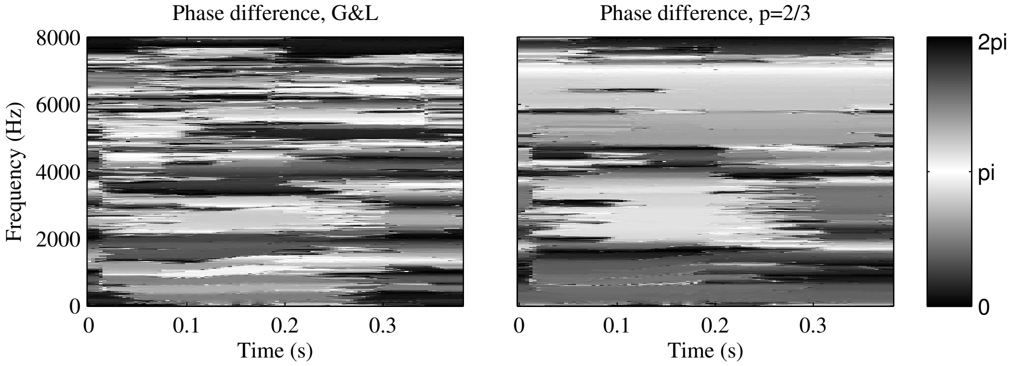


Figure 3.5: Illustration of the stagnation phenomenon: phase difference between STFTs of original and reconstructed signal for G&L and when using a compressed objective function.

signal estimates that are less prone to stagnation. This could explain the significant improvement in spectral convergence that can be observed on Fig. 3.3.

A way to account for these results is to consider the compressive behavior of the loudness-based function: the low-energy regions of the spectrogram contribute to the objective function to a larger extent than for the uncompressed function, or the G&L algorithm. It is reasonable to assume that to avoid stagnation, one needs to have a good estimate of the phase of the STFT not only in high energy regions, but over the whole time-frequency plane. By increasing the contribution of lower energy regions, the proposed method provides better reconstruction of these regions. While this is perceptually relevant as discussed earlier, it is also mathematically relevant as it provides a more consistent estimate of the phase over the whole time-frequency plane, hence limiting stagnation. The choice of $p = 2/3$ in this study was chosen based on human loudness perception, and might not be optimal. Further investigations regarding this approach might lead to better results, and determine if these results generalize to other configurations of the STFT (in terms of window duration and hop-size).

Unlike the magnitude function that removes the absolute phase of a signal, the half-wave rectification step in the IHC envelope extraction sets negative portions of channel outputs to zero, and therefore still maintains basic information regarding the absolute phase of the signal. This means there is no sign indeterminacy, no stagnation phenomenon, and a closer match between the waveforms of original and reconstructed signals can be reached (e.g., superimposed waveforms in Fig. 3.4). Although the reconstruction from IHC envelope provides signals with spectral convergence that is comparable to the one obtained when reconstructing from STFT magnitude, the absence of phase stagnation allows for very low RMS errors in the reconstruction, i.e., very similar waveforms. Hence the IHC envelope is probably better suited for reconstruction problems where accurate reconstruction of the temporal fine structure of the signal (i.e., fine details in the waveform) is critical.

3.6.4 Implications for current IHC models

The method based on IHC envelope representation was capable of reconstructing a time domain signal to a relatively high accuracy. This has implications with regard to modeling human auditory processing. While the details vary across current models of auditory processing, they all involve envelope extraction applied to the output of a filterbank. This suggests that for high-frequency channels, information regarding the temporal fine structure (i.e., the high frequency carrier fluctuations that are amplitude modulated by the envelope) is lost. However, the reconstruction method presented here suggests that this information could be recovered by processing envelopes across frequency channels. This interpretation is consistent with results from Heinz and Swaminathan (2009). In (Heinz and Swaminathan, 2009), the authors provided a theoretical framework for evaluating the neural basis for the perceptual salience of acoustic temporal fine structure and envelope cues. In their framework, temporal fine structure (carrier) information could be retrieved from (across-frequency) envelope information.

3.7 Conclusion

A general approach to reconstruct signals from an arbitrary multi-channel envelope representation was suggested. This approach was applied to both auditory and traditional spectrograms. For envelope representations computed as IHC envelope at the output of a Gammatone filterbank, signals were accurately reconstructed. This suggests that the collection of IHC envelopes provides an accurate representation of the signal, as information that is lost by the envelope extraction in individual channels can be recovered to a large extent through across-channel comparison. For STFT magnitude envelope representations, the proposed method outperformed the algorithm of Griffin and Lim (1984) for the specific STFT parameters chosen in this study (many channels and high window overlap). An analysis of the results suggested that this approach reduced the intrinsic limitations usually encountered when performing traditional spectrogram inversion.

4

Retrieval of temporal fine structure from inner hair-cell envelopes of unresolved complex tones

4.1 Introduction

In order to make use of the information available from the temporal fine structure (TFS) of a signal, a listener must be able to extract the TFS. Traditionally, it is believed that TFS information is extracted from the phase-locked response of inner hair cells (IHCs). As this phase-locking response breaks down at high frequencies, this, presumably, imposes an upper frequency limit on a listener's ability to extract TFS information. However, in the present study, we demonstrate that it is possible to reconstruct the TFS of a signal from the envelopes of a filterbank output, which we refer to as an envelope representation. This result suggests that a listener could make use of across-channel envelope processing to extract TFS information and may explain previous behavioral results. In a recent study, Santurette and Dau (2011) examined the pitch perception of a class of high-frequency complex tones. These complex tones had a periodic envelope, but the timing between the most prominent peaks in the TFS did not match the envelope period. Their study showed that some of these complex tones yielded a perceived pitch that was consistent with the timing of the peaks in the TFS rather than the envelope period. However, as the sinusoidal components in these complex tones were all sufficiently high in frequency, it is assumed that the auditory system could only extract information from the envelope of the auditory filter outputs. Based on the behavioral results, the authors suggested that TFS information might persist at higher frequencies than commonly assumed. The purpose of the current study was to investigate if, and to what extent, TFS information can be *numerically* recovered from the envelope representation of such tones. In Chapter 3, a general framework to recover a time-domain signal from an envelope representation was presented along with a specific implementation motivated by the human auditory system, where the envelope is computed from a common model of inner hair-cell (IHC) envelope, involving half-wave rectification followed by low-pass filtering. Several IHC models of this form are documented in the literature, but vary with regards to the low-pass filter parameters that are used. In this study, reconstruction from such IHC-inspired envelope representations is evaluated based on the RMS error and location of maxima in the TFS of reconstructed signals, for various IHC low-pass filter parameters.

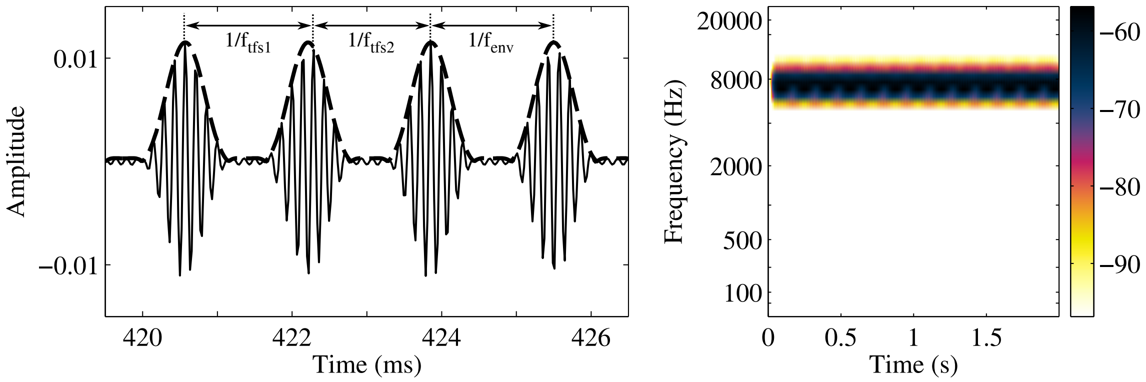


Figure 4.1: Waveform (left panel) and corresponding set of IHC envelopes (spectrogram, right panel) computed using the model of (Dau *et al.*, 1996a), for the complex tone used in this study. The alternating time intervals between the most prominent peaks of the TFS, $1/f_{tfs1}$ and $1/f_{tfs1}$, are illustrated above the waveform. Both differ from the period of the envelope, $1/f_{env}$ (dashed line in left panel plot).

4.2 Method

4.2.1 Stimulus

The complex tones from (Santurette and Dau, 2011) were generated from five sinusoidal components equally spaced in frequency, with the center component being at a frequency f_c, and the four other components at frequencies of $f_c + kf_{env}$, with $k = -2, -1, 1, 2$. The resulting signal was inharmonic, in the sense that the ratio f_c/f_{env} was not an integer. The corresponding waveform, an example of which can be observed in fig.4.1, presents a periodic envelope of period $1/f_{env}$. However, the time intervals between the most prominent peaks in the TFS differ from the envelope period, with two alternating time intervals of $1/f_{tfs1}$ and $1/f_{tfs2}$, such that $f_{tfs1} < f_{env} < f_{tfs2}$. In (Santurette and Dau, 2011), perceived pitch was measured for such signals with $f_c = 3, 4, 5, 6$ and 7 kHz and a ratio $f_c/f_{env} \approx 11.5$ and 14.5. If TFS information was not available to the listener, the perceived pitch should correspond to f_{env}. However, if the listener makes use of TFS information, then the pitch would be ambiguous and related to either f_{tfs1} or f_{tfs2}.

The complex tone with $f_c = 7$ kHz and $f_c/f_{env} = 11.5$ was chosen as the focus of the present study as it evoked a pitch that was significantly related to the TFS information rather than the period of the envelope. A version of this complex tone with duration of two seconds was generated using a sampling rate of 44.1 kHz. A smooth onset and offset was ensured by 30-ms half-raised cosine ramps. A sample of the waveform is presented in the left panel of fig.4.1, which also illustrates that the two alternating intervals between the most prominent peaks in the TFS differ from the period of the envelope.

4.2.2 Envelope extraction

The generation of the envelope representation was based on simple models of human peripheral auditory processing. The stimulus was first passed through a Gammatone filterbank composed of bandpass filters, each with a bandwidth of one equivalent rectangular bandwidth (Glasberg and Moore, 1990, ERB) and spaced one ERB apart. The envelope was extracted in each channel by half-wave rectification followed by low-pass filtering. To examine the robustness of TFS reconstruction a variety of low-pass filter parameters were tested. Thus, the envelope in each channel was computed using butterworth low-pass filters of orders 2, 4, and 6, and with cutoff frequencies decreasing from 1000 Hz to 16 Hz, in octave steps. This envelope extraction scheme corresponds to several standard IHC envelope models presented in the literature, e.g., (Lindemann, 1986; Dau *et al.*, 1996a; Breebaart *et al.*, 2001). Lindemann (1986) suggested a first order low-pass filter with a 800 Hz cutoff frequency, while Dau *et al.* (1996a) suggested a second order filter with 1 kHz cutoff frequency, and Breebaart *et al.* (2001) introduced a fifth order filter with a 770 Hz cutoff frequency. The parameters from these three models were also tested.

An example of the envelope representation (spectrogram) of the stimulus used in this study, computed using the model from (Dau *et al.*, 1996a) is illustrated in the right panel of fig.4.1. As visible in the figure, the five components of the complex tone are not individually resolved in the representation. Instead, only one component (horizontal line) spread across several channels is seen.

4.2.3 Reconstruction from the envelope

The framework introduced in Chapter 3 allows the reconstruction of time-domain signals from such auditory-inspired spectrograms following a "two-step" approach. First, a regularized inverse low-pass filter is applied in each channel. The resulting representation corresponds to a half-wave rectified output of the Gammatone filterbank and forms the target representation for the second step. There, a time-domain signal is iteratively constructed by minimizing the distance between the half-wave rectified output of the Gammatone filterbank and the target representation. This distance is based on the Frobenius norm of the difference between the two representations in their matrix forms. More details regarding this approach can be found in Chapter 3. For this study, the maximal gain allowed in the regularized inverse low-pass filter was set to 90 dB as it was found to be one of the highest possible value that does not affect the stability of the reconstruction method in the stationary sections of the stimulus. The iteration procedure was stopped after 100 iterations or when the measured distance changed by less than 10^{-9} between two iterations (which is, for the many scenarios considered in the following, between 8 and 10 orders of magnitude below the final distance value), in which case convergence was assumed.

4.2.4 Evaluating the retrieved TFS

To quantify the accuracy of the reconstructed signal, the root-mean-square (RMS) error was measured between original and reconstructed complex tones. However, an RMS metric is sensitive to errors in both timing and level. In the context of pitch perception based on TFS information, such as Santurette and Dau (2011), errors in the level (e.g., an attenuation or constant scaling of the fine

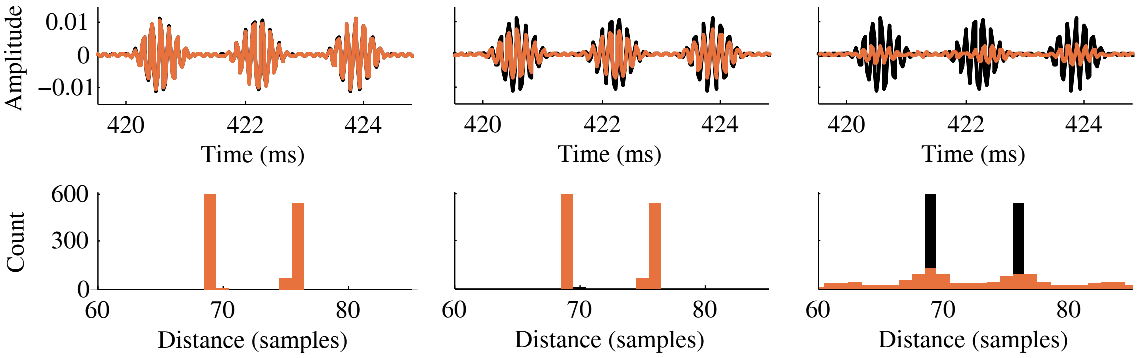


Figure 4.2: Waveforms (top three panels) of the original stimulus (black line) and signals reconstructed from the stimulus IHC envelope (orange line) for three different low-pass filtering conditions in the IHC model, fourth-order butterworth filters with cutoff frequencies of 1000 Hz (left panels), 500 Hz (central panels) and 250 Hz (right panels). The bottom three panels present the histogram of the distances between neighbor prominent peaks in the TFS for the original (black) and reconstructed (orange) signals. Note that for the bottom left and middle panels, the histograms for the original and reconstructed signals completely overlap.

structure) are less important than errors in the relative timing of peaks. In particular, the distance between the most prominent peaks in the signal for neighbor repetitions of the envelope was assumed to be the key element that needed to be accurately reconstructed in this context. Therefore a second metric was derived. In the reconstructed signal, the position of the local maximum in the TFS was noted for each period of the envelope. If reconstructed accurately, the distance between the neighbor maxima should correspond to the intervals $1/f_{tfs1}$ or $1/f_{tfs2}$ illustrated in Fig.4.1. Thus, the distribution of the distance between neighbor maxima was plotted as a histogram, and compared to the distribution of the original signal. A metric was derived by calculating the intersection of the two histograms normalized by the total number of local maxima identified. This resulted in a number between 0 and 1, with a value of 1 if both histograms were identical, i.e. fully intersecting, and 0 if they had no value in common. This metric directly relates to the proportion of correctly aligned local maxima in the TFS of the reconstructed complex tone. To avoid any bias due to instabilities of the reconstruction in the onset and offset of the signal, the first and last 30 ms of the signals were discarded in both the RMS error computation and the histogram counts.

4.3 Results

For brevity, only the results from three typical examples are illustrated in fig.4.2, here the original (black) and reconstructed (orange) waveforms (top three plots) and corresponding histograms of the distance between neighbor prominent peaks in the TFS (bottom three plots). These examples are from envelope extraction using a fourth order butterworth filter, with cutoff frequencies of 1000 Hz (left plots), 500 Hz (middle plots) and 250 Hz (right plots). For the original signal, the histogram contains two peaks distributed over four bins, with each peak having the same total count. This corresponds to the two alternating distances between neighbor prominent peaks in the fine structure of the complex tone. The leakage of each peak into two bins in the histogram is a consequence of the signal being digital, as neither of the measured distances in the corresponding continuous

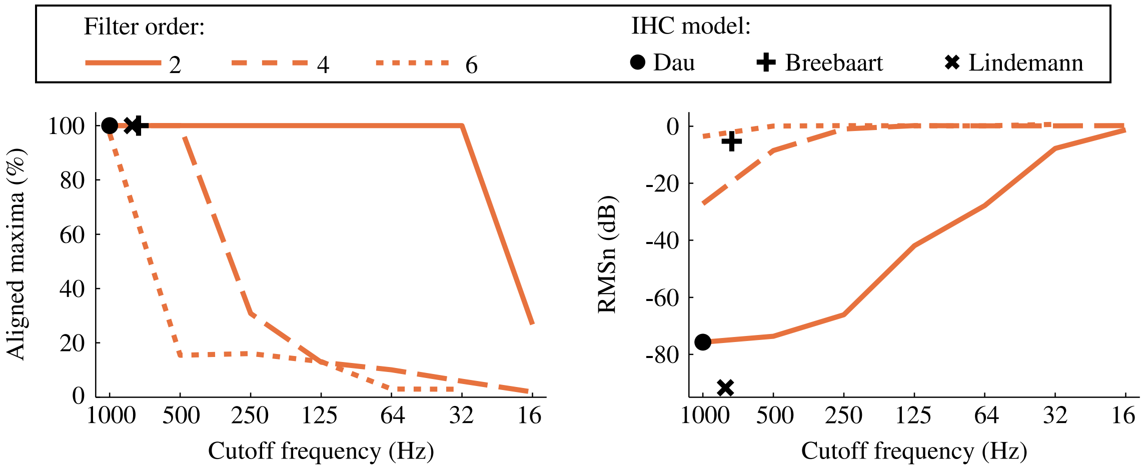


Figure 4.3: Percentage of correctly aligned prominent peaks in the TFS of reconstructed signals (left) and resulting normalized RMS error (RMS$_n$, right) for signals reconstructed from envelope representations based on different IHC low-pass filter parameters (order and cutoff frequency). Additionally, results for signals reconstructed from envelope representations based on three previously published IHC models are plotted as symbols, and labeled by the last name of the first author of the corresponding study.

analog signal is a multiple of the sampling period. The three histogram plots indicate that TFS cues could be accurately retrieved for cutoff frequencies of 1000 Hz and 500 Hz of the IHC model low-pass filter (left and middle panels where the histograms almost completely overlap) but not in the case of 250 Hz (where the histogram distributions are quite different).

The error results for reconstructions from envelope representations based on the different low-pass filter parameters tested are plotted in fig.4.3. The left panel presents the results based on the normalized histogram intersection, i.e., the proportion of local prominent peaks in the TFS of the reconstructed signal that are perfectly aligned with those of the original signal. The right panel presents the normalized RMS error (RMS$_n$) of the reconstructed signal in comparison with the original signal. Results from the (Dau *et al.*, 1996a; Lindemann, 1986; Breebaart *et al.*, 2001) filter parameters are plotted as symbols. In general, both error metrics increase as filter order is increased or filter cutoff frequency.

4.4 Discussion

The results observed in this study typically fall into one of three categories, each of which is illustrated in fig.4.2. For a given choice of low-pass filter order and cutoff frequency, the reconstructed signal could be categorized as follows: (i) A very low RMS error, thus implying perfect recovery of the TFS cues, as illustrated by the superimposed waveforms and histograms in the two left panels of fig.4.2. (ii) A significant RMS error (greater than -10 dB), yet with almost perfect recovery of the temporal position of prominent peaks (97% and above). This is illustrated by the attenuated waveform and superimposed histograms in the two center panels of fig.4.2. (iii) A large RMS error (greater than -3 dB) and a complete loss of TFS-related information, illustrated by the significant mismatch in both the waveforms and histograms distributions in the two right panels of fig.4.2.

By comparing the two plots of fig.4.3, it is possible to associate different conditions to one of these three categories. Signals are close to perfectly reconstructed from envelope representations calculated using envelope low-pass filters of lower order and/or higher cutoff frequency. On the other hand, the method fails to faithfully reconstruct signals from envelope representations computed with more restrictive low-pass filters, i.e. with lower cutoff frequency or higher order. This overall trend in the results is not surprising. However, complete retrieval of the TFS cues was achieved from envelope representations calculated using three models from the literature. Reconstruction from representations based on IHC models from Dau *et al.* (1996a) and Lindemann (1986) achieved respective normalized RMS errors of -75.7 dB and -91.7 dB. These errors are not far from the expected quantization error (-96.3 dB minimally for 16 bits), suggesting not only were some TFS cues recovered, but that near-perfect reconstruction was achieved, i.e., that no information is lost when using such models for IHC envelope extraction. These results also appear robust with regard to changes in the cutoff frequency, with the RMS error remaining below -25 dB, even when lowering this parameter by as much as four octaves. Assuming that these models are good approximations of IHC behavior, the results here implies that, based on across channel processing of the envelopes, TFS cues could be available to higher stages of auditory processing even at high audio frequencies.

The complex tone chosen reconstructed in this study exhibits several interesting properties. First and foremost, it consists of several components that appear unresolved in the envelope representation (i.e., unresolved based on place coding). Further, this signal has no energy in low frequency regions. Thus, there are no channels in the envelope representation where the TFS has not been attenuated. It could be speculated that accurate reconstruction of a signal requires at least one channel where the TFS was not heavily attenuated (i.e., easy to recover); the recovered TFS could then be used to recover the TFS in an adjacent channel and continued iteratively to higher and higher frequency channels. However, the ability to accurately reconstruct the signal used in this study refutes this claim. Thus, the accuracy of reconstructed signals in this case suggests that the missing TFS information in the spectrogram lies in the interaction between channels and in the overall consistency of the spectrogram representation.

5

Effects of manipulating the envelope signal-to-noise ratio on speech intelligibility[‡]

Jørgensen and Dau (2011) suggested a new metric for speech intelligibility prediction based on the envelope power signal-to-noise ratio ($\mathrm{SNR_{env}}$), calculated at the output of a modulation-frequency selective process. In the framework of the speech-based envelope power spectrum model (sEPSM), the $\mathrm{SNR_{env}}$ was demonstrated to account for normal-hearing intelligibility data in various conditions with stationary and fluctuating interferers as well as for conditions with linearly and nonlinearly processed noisy speech (Jørgensen *et al.*, 2013) . Here, the effect of manipulating the $\mathrm{SNR_{env}}$ on speech intelligibility was investigated by systematically varying the modulation power of either the speech or the noise before mixing the two components. A good correspondence between data and corresponding sEPSM predictions was obtained when the noise was manipulated and mixed with unprocessed speech, consistent with the hypothesis that the $\mathrm{SNR_{env}}$ is indicative of speech intelligibility. However, discrepancies between data and predictions occurred for conditions where the speech was manipulated and the noise left untouched. In these conditions distortions introduced by the applied modulation processing were detrimental for speech intelligibility but not reflected in the $\mathrm{SNR_{env}}$ metric, thus representing a limitation of the modeling framework.

5.1 Introduction

Speech intelligibility prediction has been a major research field since the first telecommunication technologies were introduced in the late nineteenth century. One of the first broadly applied methods for predicting speech intelligibility was introduced by French and Steinberg (1947) who presented the concept of the articulation index (AI). Fundamentally, the AI predicts speech intelligibility by calculating the signal-to-noise energy ratio (SNR) of the speech long-term spectrum and the background noise long-term spectrum in various frequency bands. The AI was later extended to include corrections for hearing sensitivity loss, speech level as well as upward and downward spread of masking. This has led to a revised prediction model denoted the speech intelligibility index (SII; ANSI S3.5, 1997). While this model was demonstrated to work well for predicting speech

[‡] This chapter was written in collaboration with Søren Jørgensen as a preparation to a submission to the Journal of the Acoustical Society of America.

intelligibility in conditions with stationary background noise and low- and high-pass filtering, it has limitations, for example, in reverberant conditions.

Houtgast and Steeneken (1973) demonstrated that reverberation leads to temporal smearing of the speech signal, which is not detected by the conventional SNR-metric used in the SII. Steeneken and Houtgast (1980) defined the speech transmission index (STI) as a measure of the integrity of the temporal modulations of the speech, and demonstrated that such a metric could account for the detrimental effect of stationary noise and reverberation on speech intelligibility. However, the STI-concept is also limited and fails in conditions where the noisy speech mixture has been processed by noise reduction, such as spectral subtraction (Ludvigsen *et al.*, 1993), possibly because the noise reduction affects the noise modulations as well as the speech modulations (Dubbelboer and Houtgast, 2007; Jørgensen and Dau, 2011). Several extensions to the original STI have been proposed (e.g., Payton and Braida, 2001; Goldsworthy and Greenberg, 2004) all of which, however, were based on a comparison between the clean speech and the noisy transmitted speech. Thus, none of the approaches considered the effect of the noise reduction processing on the amount of "intrinsic" modulations of the noise itself. This was done in an alternative approach by Jørgensen and Dau (2011), where it was suggested to consider the signal-to-noise ratio in the envelope domain ($\mathrm{SNR}_{\mathrm{env}}$) as a measure of the amount of useful speech modulation content available to the listener. Jørgensen and Dau (2011) quantified the modulation content of a stimulus using the power spectrum of the envelope of a stimulus relative to the DC component of the envelope's power spectrum, and demonstrated that the $\mathrm{SNR}_{\mathrm{env}}$-based metric and the STI lead to similar predictions in conditions with reverberation and stationary noise, but only the $\mathrm{SNR}_{\mathrm{env}}$ can also account for the detrimental effect of spectral subtraction on speech intelligibility. However, the relation between $\mathrm{SNR}_{\mathrm{env}}$ and speech intelligibility has not yet been evaluated in terms of explicit manipulations of the amount of $\mathrm{SNR}_{\mathrm{env}}$, while keeping the conventional energy SNR fixed. Stimuli with different $\mathrm{SNR}_{\mathrm{env}}$ but the same energy SNR can either be obtained by a modification of the modulation content of the speech signal, the noise interferer, or both. Since the $\mathrm{SNR}_{\mathrm{env}}$ is calculated from the stimuli's modulation content, this metric is sensitive to differences in the modulation content of stimuli having the same overall power, in contrast to the conventional SNR-metric used in the SII.

In the present study, the $\mathrm{SNR}_{\mathrm{env}}$ was computed using the multi-resolution version of the speech-based envelope power spectrum model (sEPSM) as presented in (Jørgensen *et al.*, 2013). The model is conceptually related to the envelope power spectrum model (EPSM Ewert and Dau, 2000) originally developed to account for psychoacoustic modulation detection and masking data. In the multi-resolution version of the sEPSM, the $\mathrm{SNR}_{\mathrm{env}}$ is estimated in short temporal segments, with segment durations inversely proportional to the center frequencies of the modulation filters considered in the processing. This model was shown to successfully predict the speech reception threshold (SRT) in a broad range of conditions with speech mixed with various stationary and fluctuating interferers as well as in conditions with noisy speech processed by spectral subtraction and reverberation (Jørgensen *et al.*, 2013). The analysis of the modulation content of a signal may be straightforward, whereas it is challenging to synthesize or process signals such that they possess prescribed modulation properties (e.g., Ghitza, 2001). Temporal modulations of a stimulus are represented by the fluctuations of its envelope (relative to the time-averaged level) after the processing through a bandpass filter. Multi-channel envelopes, obtained by passing the stimulus through a bandpass filterbank, are collectively referred to as a spectrogram. The development of spectrogram reconstruction tools (e.g. Griffin and Lim, 1984; Zhu X., 2006; Sun and Smith, 2012, or the one developed in chapter 3) makes it possible to manipulate the long-term modulation spectrum of noise or speech by reconstructing the temporal signal corresponding to a target spectrogram. Using such an approach, Elliott and Theunissen (2009) analyzed the contribution of independent temporal and spectral modulation frequency bands to the intelligibility of speech. They found that speech intelligibility remained high at about 75% words correct when restricting the temporal modulations to frequencies below 7 Hz and the spectral modulations to rates below 3.75 cycles/kHz. Restricting this "core" spectro-temporal modulation frequency range further had a large detrimental effect on intelligibility.

In the present study, the spectrogram reconstruction tool described in chapter 3 was used to generate noise backgrounds and speech stimuli with amplified or attenuated modulation content. The hypothesis was that noise with attenuated modulation content mixed with unprocessed speech as well as speech with amplified modulation content mixed with unprocessed noise should provide better intelligibility than unprocessed speech in noise. The modulation-processed stimuli were then used to evaluate the relationship between the $\mathrm{SNR}_{\mathrm{env}}$ and speech intelligibility obtained in corresponding psychoacoustic tests. Thus, the processing strategy taken here directly manipulated either the clean speech signal or the noise alone before mixing the two components, in an attempt to make the speech more intelligible in a given noisy condition. This approach differs from other modulation-domain speech enhancement strategies (Paliwal *et al.*, 2010; So and Paliwal, 2011; Wójcicki and Loizou, 2012) that focused on ways to attenuate the noise component of a noisy speech mixture. Here, the focus was to enhance the modulation content of the speech relative to the noise, before mixing the two components. Such an approach could be useful in a situation where there is access to the speech signal before it is transmitted and mixed with environmental noise, such as at a train station.

5.2 Method

5.2.1 Speech material, apparatus, and procedure

Speech reception thresholds (SRT) were measured using the material provided in the Danish Conversational Language Understanding Evaluation (CLUE; Nielsen and Dau, 2009), which is similar to the hearing in noise test (HINT; Nilsson *et al.*, 1994). The speech material in the CLUE test consists of 18 lists of ten unique sentences, recorded in anechoic conditions. Each sentence represents a meaningful everyday sentence containing five words, spoken in a natural manner, by a male speaker. The background noise in the CLUE test is a stationary speech-shaped noise (SSN) constructed by concatenating and superimposing the sentence material so as to obtain a stimulus with the same long-term spectrum as the average long-term spectrum of the sentences. Five male normal-hearing native Danish speakers, aged between 24 and 38 years, participated in the study. The subjects were sitting in a double-walled insulated booth together with the experimenter who was controlling the procedure through the dedicated MATLAB application for speech intelligibility measurement using the CLUE material. The digital signals, sampled at 44.1 kHz, were converted to analog by a high-end RME DIGI96/8 soundcard. They were presented to the subjects diotically via Sennheiser HD580 headphones. The average sound pressure level (SPL) of the stimuli in the test was 65 dB. After each presentation, the subjects were instructed to repeat the words he/she understood, with the possibility of guessing or passing on misunderstood words. The experimenter recorded the correctly understood words individually.

The SNR was controlled throughout the test by changing the level of the SSN after the listener's response, using an adaptive procedure. If all the words of a sentence were repeated correctly the SNR was lowered by 2 dB, otherwise it was increased by 2 dB. The SRT was determined as the average of the SNRs calculated after the response to the last eight sentences of a list. Further details can be found in Nielsen and Dau (2009).

5.2.2 Stimulus conditions

Two stimulus conditions were considered: (i) unprocessed speech mixed with SSN that was processed in the modulation domain as described further below (section 5.2.3) and (ii) speech that was processed in the modulation domain and mixed with unprocessed SSN. The modulation processing either attenuated or amplified the modulation power in a target modulation frequency range between 4 and 16 Hz, while providing zero gain outside the target range. Six conditions of the target modulation gain were considered when only the noise was processed: 20, 10, 0 , -5, -10 and -20 dB relative to the unprocessed noise, and seven conditions were considered when only the speech was processed: 20, 10, 5, 0, -6, -10 and -20 dB, whereby 0 dB represented the unprocessed reference condition. Smooth transitions between the amplified/attenuated modulation-frequency band and the zero-gain frequency region were obtained using raised cosine ramps in the transition bands from 1 to 4 Hz and from 16 to 22 Hz. The efficiency of the modulation processing was analyzed by computing the modulation transfer function (MTF) of the processed signals relative to the unprocessed signal as suggested by Schimmel and Atlas (2005). The MTF of a single channel, m, was defined here as

$$MTF_m = \frac{|\mathscr{F}\{|(p_m)_a|\}|}{|\mathscr{F}\{|(u_m)_a|\}|} \tag{5.1}$$

where p denotes the processed signal, u denotes the unprocessed signal, $\mathscr{F}$ denotes the Fourier transform, $(\cdot)_a$ denotes the analytical signal, and $|\cdot|$ denotes the modulus. The MTF of an entire spectrogram was taken as the average of the subchannel MTFs, where the MTF$_m$ of the individual channels were weighted according to their energy

$$MTF = \frac{1}{\sum_{m=1}^{M} \|u_m\|_2} \sum_{m=1}^{M} \|u_m\|_2 \cdot MTF_m \tag{5.2}$$

where $\|\cdot\|_2$ is the Euclidean norm. Figure 5.1 (upper left panel) shows the MTF for the processed noises for each of the five target modulation gains. The dashed curves represent the target MTF and the solid curves show the obtained "actual" MTF at the output of the modulation processing stage that is described in more detail further below. In a perfect modulation processing system, the dashed and solid lines would coincide. In the processing framework presented here, the actual gain/attenuation is smaller than the target gain/attenuation, particularly for the largest target values of +/-20 dB. The top right panel of Fig. 5.1 shows the corresponding long-term excitation patterns (Glasberg and Moore, 1990) of the processed noises, representing the total energy of the signal at the output of a bank of gammatone filters with one equivalent rectangular bandwidth spacing, plotted as a function of the center frequencies of the filters. The patterns for the three negative gains (-5, -10 and - 20 dB) coincide with the one obtained for the unprocessed signal. The pattern for 20 dB gain (light gray) lies above the unprocessed pattern (black) by more than +5dB at very low frequencies (< 50 Hz) and below the unprocessed pattern by more than -5 dB at frequencies above 1200 Hz.

The lower left panel of fig.5.1 shows the target MTFs (dashed lines) and the actual MTFs (solid lines) for the six conditions where the speech was processed. Each MTF represents the average across the individual MTFs obtained for 180 sentences without noise. The obtained actual gains/attenuations are below the respective target gains/attenuations, particularly for the target gains of +/- 20 dB. Furthermore, the effective frequency region of the MTF relative to the target range (4 to 16 Hz) is shifted towards higher modulation frequencies in the conditions with modulation enhancements, which is different from the results obtained with the processed noise. The lower right panel of fig.5.1 shows the corresponding excitation patterns for the processed speech and the unprocessed speech. The maximal deviation of the patterns for processed speech from the one for unprocessed speech amounts to 5 dB at frequencies above 5 kHz.

5.2.3 Modulation processing framework

The modulation processing consisted of two parts as indicated in fig.5.2. In part A, the original unprocessed signal was first passed through a Gammatone filterbank, and the Hilbert envelope was extracted at the output of each filter. The resulting set of envelopes constituted the unprocessed spectrogram. Each envelope of the spectrogram representation was filtered by a zero-phase bandpass filter with a given target MTF. To avoid transients in the filter magnitude response, transitions between the pass-band and the filtered band were smoothed using half raised cosine windows. The filtered envelopes were then individually normalized such that they had the same root mean square (RMS) value as their unfiltered counterpart. This ensured that the total power of the envelopes in each frequency channel was only marginally affected by the envelope filtering such that the processed signal had a similar long-term audio-domain excitation pattern as the original signal (right panels of fig.5.1).

To be consistent with the definition of a spectrogram, each processed envelope had to be non-negative. However, filtered envelopes could exhibit significant negative sections, particularly when

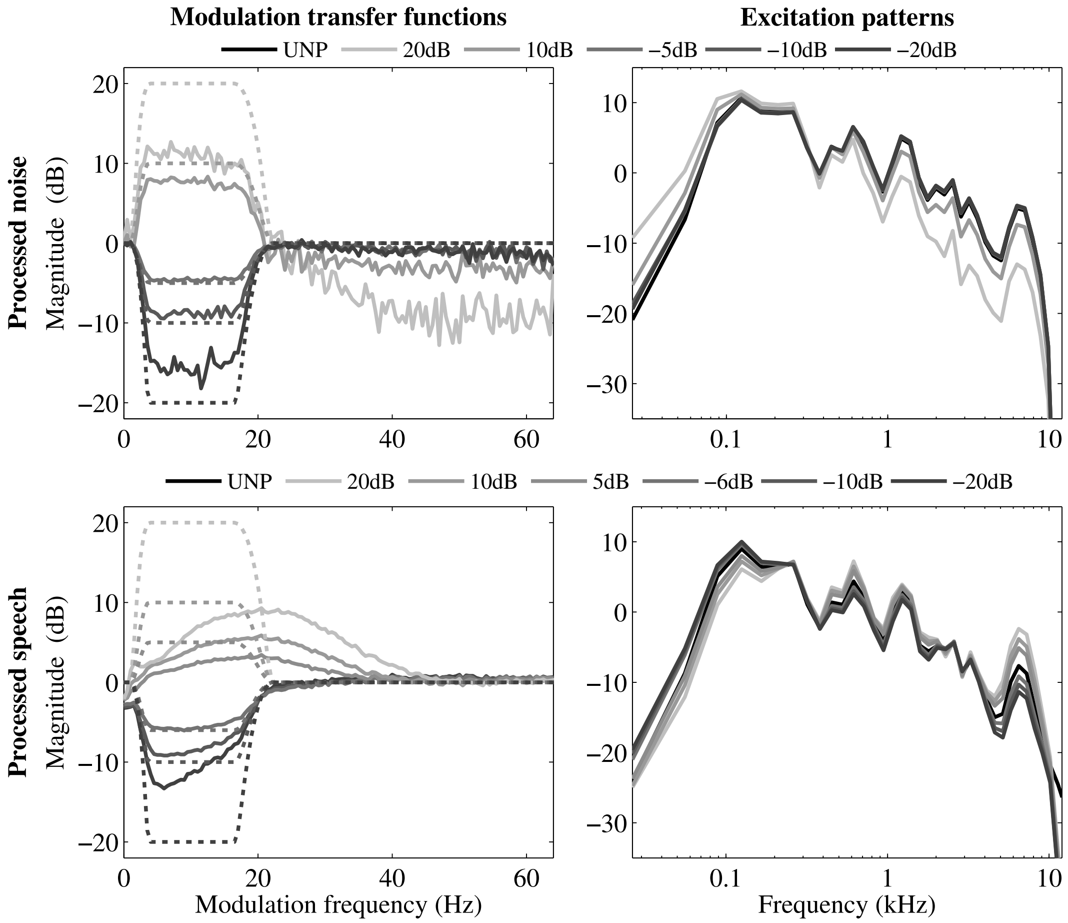


Figure 5.1: Top left: Target modulation transfer functions (MTF; dashed lines) and actual MTFs (solid lines) for the five conditions with processed noise. The grayscale corresponds to different target modulation gains (-20, -10 -5, 10, 20 dB). Each noise signal was 22 seconds long, sampled at 22.05 kHz, and the corresponding MTF was obtained as the average over 50 segments of 2 s each. Top right: Long-term excitation patterns of the five processed noises and the unmodified noise (UNP). Bottom left: Target (dashed lines) and actual (solid lines) MTFs of the processed speech for the six target modulation gains. MTFs were averaged over the 180 sentences of the CLUE material. Bottom right: Long-term excitation patterns of the speech for the six target modulation gains and the unmodified speech (UNP).

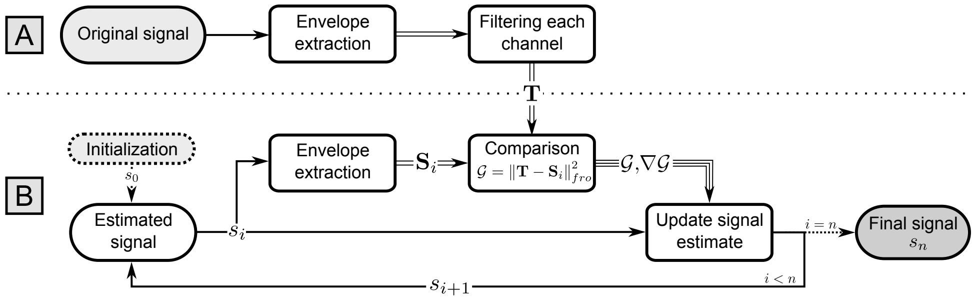


Figure 5.2: Schematic view of the two steps in the modulation processing framework. Single-lined arrows relate to time-domain signals and double-lined arrows to multi-channel envelopes (spectrograms). Step A, upper part, generates a target spectrogram **T** by separately filtering each channel of the original signalÂt's spectrogram. In step B, a signal is constructed iteratively by comparing the spectrogram $\mathbf{S}_i$ of the current signal estimate s_i to the target **T**. The distance between the two spectrograms $\mathscr{G}$ and its gradient $\nabla\mathscr{G}$ are used to update the current signal estimate until the maximum number of iterations n is reached.

large positive modulation gains were provided to signals that initially contained large envelope fluctuations, such as speech. To overcome this, the dynamic range of the envelope was limited by raising the envelope of each channel to the power of 1/3 before filtering. After filtering, the original dynamic range was restored by raising the filtered envelope to the power of 3. The resulting filtered spectrogram provided the "target" input, **T**, to the signal reconstruction stage (Part B) of the modulation processing. In the signal reconstruction, indicated as Part B in fig.5.2, a time-domain signal, s, was reconstructed iteratively, such that the difference between the spectrogram of the reconstructed signal and the target spectrogram was minimal. The procedure was initiated by a random noise signal s_0 that, for each iteration i, was updated in the direction that reduced the distance between its spectrogram $\mathbf{S_i}$ and the target spectrogram **T**. The distance, $\mathscr{G}$, between the spectrograms was given as the square of the Frobenius matrix norm of the difference between the two spectrograms:

$$\mathscr{G} = ||\mathbf{T} - \mathbf{S_i}||^2_{fro} \tag{5.3}$$

The iterative procedure was terminated after 100 iterations. Details about the signal reconstruction can be found in chapter 3.

5.2.4 Speech intelligibility prediction

The processing structure of the sEPSM is illustrated in fig.5.3. The details of the processing can be found in (Jørgensen and Dau, 2011) and (Jørgensen *et al.*, 2013). Some of the main stages are described in the following. The first stage is a bandpass filterbank consisting of 22 gammatone filters (Glasberg and Moore, 1990) with a third-octave spacing between their center frequencies, covering the range from 63 Hz to 8 kHz. The temporal envelope of each output is extracted via the Hilbert-transform and then low-pass filtered with a cut-off frequency of 150Hz using a first-order Butterworth filter. The resulting envelope is analyzed by a modulation bandpass filterbank, which consists of eight second- order bandpass filters with octave spacing, covering the range from 2 - 256 Hz, in parallel with a third-order lowpass filter with a cut-off frequency of 1 Hz.

The running temporal output of each modulation filter is divided into short segments using rectangular windows with no overlap (Jørgensen *et al.*, 2013) . The duration of the window is specific for each modulation filter, and is equal to the inverse of the center-frequency of a given modulation filter (or the cut-off frequency in the case of the 1-Hz low-pass filter). For example, the window duration in the 4-Hz modulation filter is 250 ms. For each window, the AC-coupled envelope power (variance) of the noisy speech and the noise alone are calculated separately and normalized with the corresponding long-term DC- power. The $\mathrm{SNR_{env}}$ of a window is estimated from the envelope power as:

$$\mathrm{SNR_{env}} = \frac{P_{env,S+N} - P_{env,N}}{P_{env,N}} \tag{5.4}$$

where $P_{env,S+N}$ and $P_{env,N}$ denote the envelope power of the noisy speech and the noise alone

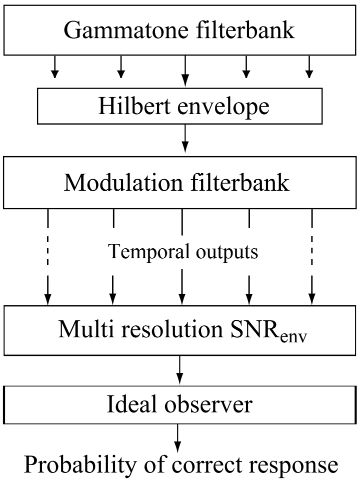


Figure 5.3: Block diagram of the processing structure of the sEPSM (Jørgensen *et al.*, 2013) . Noisy speech and noise alone are processed separately through a gammatone bandpass filterbank followed by envelope extraction via the Hilbert transform. Each sub-band envelope is further passed through a modulation bandpass filterbank. The modulation- filtered temporal outputs are segmented with a segment duration inversely related to the modulation-filter center frequency. The envelope power (variance) is computed in each segment for the noisy speech ($P_{env,S+N}$) and the noise alone ($P_{env,N}$), from which the corresponding SNR$_{\text{env}}$ is derived. The segmental SNR$_{\text{env}}$ is then averaged across segments and combined across modulation filters and audio-domain (peripheral) filters. Finally, the overall SNR$_{\text{env}}$ is converted to the probability of correct response assuming an ideal observer as in Jørgensen and Dau (2011) .

after the normalization. For each modulation filter, the running SNR$_{\text{env}}$-values are averaged across time, assuming that all parts of a sentence contribute equally to intelligibility. The time-averaged SNR$_{\text{env}}$-values from the different modulation-filters are then combined across modulation filters and across Gammatone filters, using the “integration model” from Green and Swets (1988). The combined SNR$_{\text{env}}$ is converted to the probability of correctly recognizing the speech item using the concept of a statistically “ideal observer” (Jørgensen and Dau, 2011) .

For the simulations, 150 sentences from the CLUE material were used. Each sentence was mixed with a noise token (randomly selected from the full-length noise files) over a discrete range of SNRs. For a given SNR-value, the final percent correct prediction was computed as the average predicted score across all sentences of a given speech material. The prediction at each SNR was then connected by straight lines, resulting in a continuous psychometric function, from which the SRT was estimated as the SNR corresponding to 50% correct. The values of the parameters in the model were kept fixed in all conditions and corresponded to those given in Table II in Jørgensen *et al.* (2013).

5.3 Results

Figure 5.4 shows the results for the conditions where the noise interferer was processed and the speech left untouched. The open symbols show measured speech intelligibility data, represented as the change in SRT (ΔSRT) relative to the unprocessed condition. ΔSRT is shown as a function of the target modulation gain, and a positive ΔSRT reflects worse intelligibility compared to the

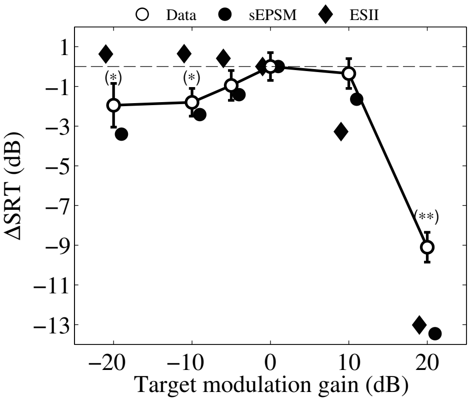


Figure 5.4: Change of the speech reception threshold, ΔSRT, relative to the unprocessed condition (0 dB target gain), as a function of the target modulation gain applied to the noise interferer (but not the speech signal). Open circles represent the measured data, with error bars indicating standard errors. Asterisks indicate a statistically significant difference to the reference condition, with $p < 0.1$ represented as $(*)$ and $p < 0.01$ represented as $(**)$. The filled circles show predictions obtained with the sEPSM, and filled diamonds show predictions using the ESII.

unprocessed condition. An analysis of variance was conducted to assess the statistical significance of the measured data. The statistical results are presented as asterisks above the data points, with $p < 0.1$ for $(*)$ and $p < 0.01$ for $(**)$ indicating significant differences from the unprocessed condition. A non-monotonic relationship between the obtained SRT and the target modulation gain was observed. In the range of conditions with negative gain, i.e. with attenuated noise modulations, the SRT decreased slightly (up to about 2 dB) with decreasing gain. In the conditions with positive gains, i.e. amplified modulations, a large decrease of the SRT of about 9dB was observed for the target gain of 20 dB. The filled circles represent the predictions obtained with the sEPSM. The predictions were in good agreement with the data, although the model slightly overestimated the effect at large positive gains, i.e. slightly overestimated the benefit of enhancing the noise modulation on intelligibility. For direct comparison, predictions obtained with the extended SII (ESII[1] Rhebergen *et al.*, 2006), using a stationary noise as the speech signal, are also shown in fig.5.4 and indicated by the filled diamonds. The ESII predicted the trend in the data for positive modulation gains, but predicted a positive ΔSRT, i.e. a slight decrease of speech intelligibility, for negative gains, in contrast to the measured data.

Figure 5.5 shows the results obtained for the conditions with processed speech (in the presence of unprocessed noise). The open symbols show the measured data. The SRT increased by 1.5 dB for a target modulation gain of 10 dB and by 5.5 dB for a gain of 20 dB, i.e. representing a decrease of intelligibility. Similarly, in the conditions with negative gains, the SRT increased by 2.7 dB for a target gain of -10 dB and by 7.3 dB for a target gain of -20 dB. Thus, the intelligibility decreased in all conditions where the speech (alone) was processed.

The predictions obtained with the sEPSM are shown by the filled circles. The results are essentially independent of the amount of negative gain, in clear contrast to the data. Moreover,

[1] The ESII used here corresponds to the method described in Rhebergen *et al.* (2006). SSN was used as probe for the speech material.

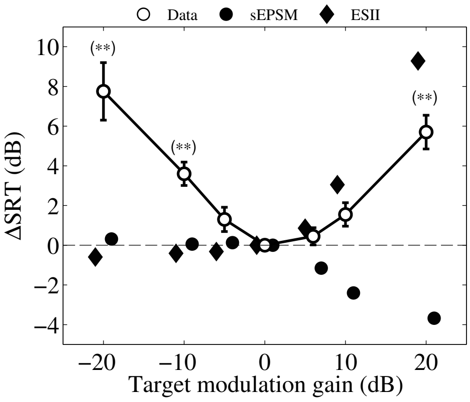


Figure 5.5: ΔSRT relative to the unprocessed condition (0 dB target gain) as a function of the target modulation gain applied to the speech (mixed with unprocessed noise). The open symbols represent the measured data, the filled circles show the sEPSM predictions, and the filled diamonds show predictions using the ESII. The error bars represent standard errors. Asterisks indicate a statistically significant difference to the reference condition, with $p < 0.1$ for (∗) and $p < 0.01$ for (∗∗).

the model predicted a decrease in SRT for the conditions with positive gains. This reflects the underlying assumption of the model linking enlarged modulation power of the speech signal (cf. bottom left panel of fig.5.1) to increased speech intelligibility. However, this relation is not supported by the data for the conditions with modulation enhanced speech. For comparison, predictions obtained with the ESII are also shown in the figure and indicated by the filled diamonds. Since the ESII procedure applies a stationary noise as a probe for the speech signal, the simulations for the processed speech conditions essentially represent a mirrored pattern of the simulations for the processed noise conditions (around the 0-dB ΔSRT axis). The ESII appeared to account at least partly for the reduced intelligibility for positive gains. However, it is unclear if the trend predicted by the ESII reflects the underlying cause for the reduced intelligibility observed in the data or if it is a consequence of the special noise-probe signal used in the ESII calculation procedure.

5.4 Discussion

5.4.1 Modulation processing of the noise interferer

The predictions obtained with the sEPSM were in good agreement with the measured data in the conditions where the noise was processed and the speech was left unchanged. This suggests that the improvement in intelligibility observed in the data can be accounted for by a greater $\mathrm{SNR}_{\mathrm{env}}$ after the modulation processing. For negative gains, the modulation power of the noise was effectively reduced in the range from 4-16 Hz, leading to a temporally more “steady” noise. Thus, in the framework of the model, the lower SRTs obtained in these conditions are caused by a release from modulation masking. In the case of positive modulation gains, the amplification of the noise modulations led to strong amplitude variations of the noise, similar to amplitude modulated noise (e.g., Festen and Plomp, 1990). The lower SRT in these conditions can also be explained in terms of a release from modulation masking, but at different modulation frequencies than in the case of the negative gains. Jørgensen *et al.* (2013) demonstrated in their model analysis that the greater intelligibility obtained in conditions with a fluctuating noise compared to a stationary noise may be based on the availability of high-frequency (>30 Hz) speech envelope fluctuations. The ESII was able to account for the results obtained with positive modulation gain, consistent with earlier studies that demonstrated its usefulness for predicting speech intelligibility in the presence of a fluctuating interferer. However, the ESII failed to account for the conditions with negative modulation gains, which could be accounted for by the sEPSM. This may be explained by the different intelligibility metrics underlying the predictions obtained with the ESII and the sEPSM. In the case of the ESII, the intelligibility metric is based on the conventional (short-term) energy of the stimuli, which is similar across the conditions with attenuated noise and, thus, does not capture the reduced modulation energy of the noise. The results obtained in the conditions with reduced noise modulations can therefore not be accounted for by the concept of "glimpsing" (e.g., Cooke, 2006).

5.4.2 Modulation processing of clean speech

Ideally, according to the concept of the $\mathrm{SNR}_{\mathrm{env}}$, amplifying the modulations of the speech signal by a given amount (before mixing it with the unprocessed noise) should lead to a similar effect on intelligibility as attenuating the modulations in the noise by the same amount (before mixing the noise with the unprocessed speech). The predictions with the sEPSM, indeed, showed the same SRT when applying the target modulation gain of 20 dB to the speech as when applying a gain of -20 dB to the noise. However, the measured data obtained in the conditions where the modulations of the clean speech were amplified differed strongly from the conditions where the modulations of the noise alone were attenuated. This difference may be explained by distortions of the speech signal resulting from the modulation processing. The degree of speech distortion after modulation processing was assessed here using the Perceptual Evaluation of Speech Quality (PESQ; ITU-T P.862, 2001) method. Figure 5.6 shows the speech signal distortion, defined here as the inverse of the PESQ-rating, scaled to a range between 0 and 1, for the different conditions of modulation-processed speech. The distortion is zero for the zero-gain reference condition and increases with increasing or decreasing target modulation gain. This trend corresponds well to the trend observed in the intelligibility data shown in fig.5.5. The presence of distortion and the attribute of unnaturalness in the conditions with modulation-processed speech were also reported qualitatively by the listeners, even at high SNRs.

5.4.3 Limitations of modulation enhancement in speech processing

There were at least two sources of distortions observed when amplifying the modulation of natural speech. First, the step in the modulation filtering process that generated the target spectrogram represented a "blind" process, i.e., the filtering process had no *a priori* information about the initial spectro-temporal structure of the speech signal.

Hence, the amplified modulation was not constrained to follow the natural temporal structure

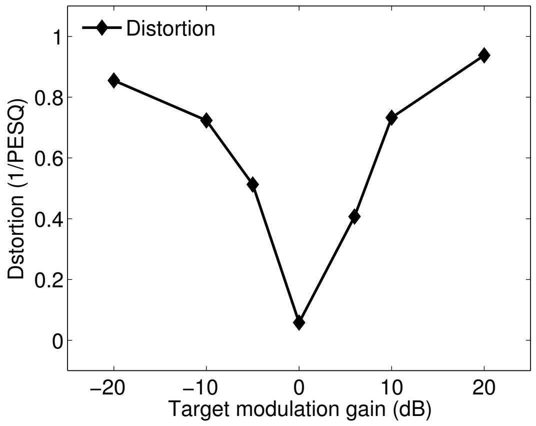


Figure 5.6: Objectively measured distortion of the speech signal for the different conditions of modulation processing. The distortion is defined here as the inverse of the PESQ measure scaled to the range between 0 and 1, and each point represent the average across 50 sentences. A metric such as the $\mathrm{SNR}_{\mathrm{env}}$ assumes that all audible speech modulations contribute to speech intelligibility. The metric cannot account for distortions that are not resulting from the processing applied to the mixture of speech and noise (such as spectral subtraction).

of the speech, but could represent a strong modulation component not related to speech at all. Moreover, the larger the modulation gain/attenuation was, the further the filtered spectrogram deviated from a faithful representation of speech, i.e., the target spectrogram could represent distorted speech rather than enhanced speech. A second source of distortion was related to the iterative reconstruction process. Since the spectrogram representation considered here was of higher dimensionality than a real time-domain signal, many theoretical spectrograms could result from the modulation-filtering process, for which no corresponding real time-domain signal would exist (Le Roux *et al.*, 2010). Thus, the target spectrogram obtained in fig.5.2 (Part A) did not necessarily correspond to an actual time-domain signal. This implied that the objective function in (5.3) might never reach zero, even if the reconstruction was successful in the sense that it reached the minimum of the objective function. The remaining distance between the target and the obtained spectrograms would translate into an error in the reconstructed signal, in the form of uncontrolled distortions.

5.4.4 Usefulness of modulation processing for speech intelligibility enhancement

While the success of the present modulation processing to enhance speech intelligibility was limited to the conditions with processed noise alone, the modeling results suggested that there could be a potential improvement of intelligibility by enhancing the modulation of the speech material used in the present study. It would be interesting to investigate if the model would account for speech intelligibility data obtained with speech containing naturally enhanced modulations, such as clearly articulated speech, which is free from distortions. Such stimuli have been considered in the study of Payton *et al.* (1994) and Payton and Braida (1999), showing improvements in the word score of 20-30 % for clearly spoken speech compared to conversational speech. This corresponds to an improvement of SRT of roughly 9 dB based on the difference between two psychometric functions (as described by Wagener *et al.*, 2003) fitted to the data from Payton *et al.* (1994) (not shown here explicitly). Such an improvement may well represent the upper limit of the potential benefit provided by an artificial speech modulation enhancement approach. The benefit of processing either the noise or the speech alone before mixing appeared to be modest compared to other approaches of speech intelligibility improvement, where the mixture of speech and noise is modified. For example, Wójcicki and Loizou (2012) demonstrated that discarding noise-dominated modulation-domain spectral components based on an SNR_{env}-like metric led to improvements of SRT up to 13 dB. Their approach was fundamentally different from the one considered in the present study, since they targeted noise-removal rather than modifying the unmixed signals to optimize intelligibility. The benefit of the present approach is that the enhancement could, in principle, be obtained by a filter-like operation, without having to estimate the noise component. In practice, the enhancement could be performed as a kind of pre-processing of the speech signal that could optimize intelligibility, before it is mixed with noise. Moreover, noise-removal could, in principle, also be achieved with the present modulation processing framework, by modifying the setup of the target spectrogram (Part A of fig.5.2) such that it produces a target that is a noise-reduced version of the noisy speech mixture.

5.5 Summary and conclusions

The effect of manipulating the SNR_{env} on speech intelligibility was investigated by systematically varying the modulation power of either the speech or the noise before mixing the two components. Improvements of the SRT for normal-hearing listeners were obtained in conditions where the modulation power of the noise was modified, leaving the speech untouched. Predictions from the sEPSM accounted well for this effect, supporting the hypothesis that the SNR_{env} is indicative of speech intelligibility. However, a large detriment of the SRT was obtained when the speech modulation power was modified and the noise left untouched, which could not be accounted for by the sEPSM. This discrepancy might be explained by distortions introduced by the modulation processing, which were not reflected in the SNR_{env} metric, thus, representing a limitation of the modeling framework.

6

Insights in spectrogram filtering: directions for improvements

The previous chapter presented an application of spectrogram reconstruction methods to perform modulation filtering. A time-domain signal reconstructed from a filtered spectrogram representation presents new modulation patterns that correspond, to some extent, to the filter applied to each channel of the original spectrogram. However, this approach has some limitations, and it was shown that the effective change in modulation is often less than what would be expected from the filter used. Such situations of less effective modulation filtering were mostly observed when processing speech, and particularly when trying to enhance the modulation content of speech. In the previous chapter's discussion, some issues that are likely responsible for these limitations were identified. It was noted that speech has a far more complicated time-frequency structure than noise, and that a processing framework that was applied "blindly" with regards to this structure could impair the processing's efficiency. Additionally, it will be shown here how the non-negativity nature of envelopes can be an issue when filtering spectrograms, impairing the whole process prior to the actual reconstruction. In this chapter, two preliminary studies are presented to assess these possible limitations. The first section presents a short-time based implementation of the spectrogram filtering framework, intending at better-accounting for the intrinsic structure of the signal being processed, limiting the negative effects of what will be introduced as non-stationarity. The second section suggests a method to filter a spectrogram according to a given filter magnitude response while retaining non-negative, i.e., envelope compliant, outputs. The last, short, section motivates a conceptual approach to the modulation filtering scheme that does not involve a direct reconstruction from a filtered spectrogram but still relies on the iterative construction of a time-domain signal. Although the concepts developed in this chapter did not lead to a concrete improvement of the modulation filtering framework, they are documented here as they offer insights in some of the issues faced with modulation manipulation, and form a good basis for potential future directions.

6.1 Short-time based spectrogram filtering

From the experimental results obtained in chapter 5, one of the most straightforward observations was that the modulation filtering framework performed far more efficiently when applied to noise than when applied to speech. Speech embeds an intricate underlying structure, mixing intonation dependent harmonics, transients, high frequency bursts as well as silence. In brief, speech signals are complex and cannot be easily described. On the other hand, noise signals involve stochastic processes for their generation. Hence, faithful descriptors of a noise signal are often found in its statistical properties (e.g., its variance, long-term spectrum, probability distribution, higher-order moments, etc...). Such statistical descriptors for the noise signal used in this document are relatively stable over time. This speech-shaped noise can be considered as a *stationary* signal. In contrast, speech is not stationary. However, speech is often considered to be a *quasi-stationary* signal: its statistical properties vary relatively slowly with time, and short individual segments of speech can be assumed to be stationary.

The modulation filtering framework presented in chapter 5 involves the reconstruction of a time-domain signal from a filtered spectrogram. The filtering was performed on a long-term basis, and the resulting modulation transfer functions (MTFs) for speech (e.g., bottom-left plot in fig.5.1) were derived for the whole signals. Because speech signals are non-stationary, the MTF of a short segment of processed speech will deviate significantly from the MTF of the whole signal, i.e., long-term based MTFs are not representative of the effect of the processing in shorter speech segments. In this section, a short-time based filtering approach is suggested to generate filtered spectrograms with changes in their modulation properties that are more consistent over shorter time segments. The reconstruction of a time-domain signal from the filtered spectrogram is unchanged and still as described in chapter 3.

6.1.1 Method

Short-time temporal modulation transfer function

The modulation transfer function was introduced in chapter 5 as an analysis tool to illustrate the effects of the modulation filtering framework in the modulation domain. As defined in (5.2), it relied on the ratios between the Fourier magnitudes of processed and unprocessed signals for each individual channel. These ratios were averaged over the channels to provide the MTF. When computing the Fourier magnitudes, the Fourier transform of the whole channel was computed. For non-stationary signals, the Fourier transform of a short segment of the signal will be significantly different from the Fourier transform of the whole signal. The MTF for this short segment will therefore deviate from the MTF of the whole signal, indicating that the long-term MTF as defined in (5.2) is not necessarily a good measure of the efficiency of the modulation filtering for local sections of the signal.

To analyze the modulation filtering efficiency for local sections of the signal, a short-time based MTF can be adapted from the MTF definition previously used. The original and processed signals are decomposed into several overlapping short segments. The MTF for each segment is computed, and averaged over the segments to provide a short-time MTF average (stMTF). In addition, the standard deviation of the short-time MTFs can also be measured for each modulation frequency

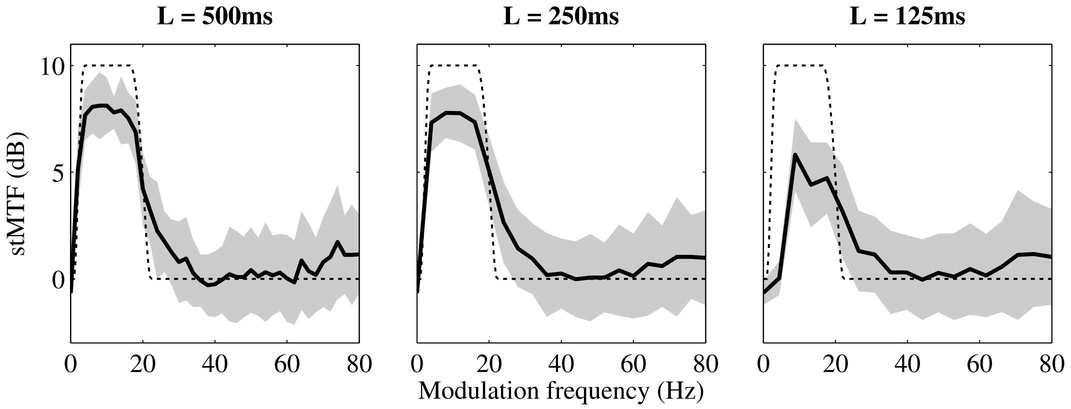


Figure 6.1: Short-time modulation transfer function average (stMTF) of modulation-filtered noise (solid line) showing plus/minus one standard deviation (gray area) and the modulation filter magnitude response (dashed line), plotted for three different analysis window duration *L*.

bin. A low standard deviation would indicate that the processing is performed consistently over short segments. Conversely, a high standard deviation would suggest that the modulation content of short segments was not processed as expected.

The segmentation of original and processed signals is performed using a sliding window. The window has a fixed duration, and is time-shifted by a fixed amount between neighbor segments, i.e., with a given window repetition rate. To limit spectral leakage from onset and offset discontinuities when measuring the MTFs of individual segments, a Hanning window is used. The window duration determines the resolution of the computed stMTFs. A window that is *L* seconds long will yield stMTFs with a resolution of $1/L$ Hz per modulation frequency bin, independent of the sampling frequency. The MTF is concerned with illustrating modulation properties which are, for speech at least, concentrated in relatively low frequencies. Therefore the modulation frequency resolution of the MTF needs to be high enough if one wants to observe the processing effect on low modulation frequencies. This implies a lower bound on the duration *L* of the analysis window.

Figure 6.1 presents the short-time MTF average (stMTF, solid line) computed with three different analysis window durations *L*, for modulation-filtered noise. The gray area indicates one standard deviation interval above and below the average, and the dashed line the modulation filter magnitude response. It illustrates the lower bound requirement on *L*. The processing in this case typically involves enhancement of the modulation content between 4 and 16 Hz, suggesting a resolution of the stMTF of at least 4 Hz, i.e. a window of at least 250 ms. The stMTF obtained with 125 ms window (right plot in fig.6.1) is clearly degraded in comparison to those obtained with 250 ms and 500 ms long windows (respectively center and left-most plot), and its resolution is not sufficient to accurately represent modulation frequencies below 8 Hz. Note that for practical reasons, the implementation of the filterbank allows only for signals longer than 5000 samples (i.e., about 227 ms at the sampling rate of 22050 Hz used here). Hence, if shorter windows are used, such as 125 ms here, zero-padding of individual segments is necessary before computing their MTF. It can be observed on fig.6.1 for the 125 ms-window where the actual frequency resolution is not the expected 8 Hz but is instead 4.41 Hz (i.e., corresponding to segments of 5000 samples at this sampling rate).

The stMTF obtained with 250 ms long window, despite its lower resolution, appears smoother than the one obtained with a longer window. This is a consequence of the windowing. The window used here, in the frequency domain, has a main lobe width which is comparable to the frequency range of interest. Typically the main lobe of a Hanning window of L seconds has a -6 dB cutoff frequency of $\frac{2}{L}$Hz (Harris, 1978), i.e., 4 Hz for $L = 500$ ms and 8 Hz for $L = 250$ ms. The windowing in the time domain translates to a convolution in the frequency (or modulation frequency) domain, and here the non-negligible width of the window's frequency response acts as a low-pass filter on the resulting MTFs, which will appear smoother for shorter windows. Eventually, too short windows, such as for $L = 125$ ms, will not only smoothe the MTF but will also present an attenuated response in the region from 4 to 16 Hz, as it acts as a low-pass filter whose cutoff frequency is below the bandwidth of this region. This can be observed in the right plot of fig.6.1, where the stMTF in the 4-16 Hz band is significantly attenuated. In the following, a Hanning window of 500 ms is chosen to compute stMTFs, as it offers a good compromise between a short window duration and detrimental effects in the representation of the MTF.

The stMTF presented in fig.6.1 exhibits a large standard deviation, indicating that in individual segments, local MTFs can vary significantly from the final averaged stMTF. The purpose of a short-term implementation of spectrogram filtering is to generate filtered spectrograms which will present a lower standard deviation of their stMTF, i.e., for which the outcome of the filtering procedure is more consistent across time. It is not expected that such an implementation will greatly reduce the stMTF's standard deviation for stationary signals such as the noise. However the effect is expected to be significant for non-stationary signals such as speech, assuming that non-stationarity is the main cause of the deviations of local short-term MTFs to the averaged stMTF.

Weighted overlap-add filtering approach

Short-term signal processing can often be conducted using a traditional overlap-add (OLA) method (Allen, 1977). The OLA approach consists in extracting consecutive overlapping segments of the signal by using a sliding window. Each segment is processed according to its extraction order, and adequately (i.e., with correct time alignment) added to the output from previously processed segments. Although the OLA approach, assuming a reasonable choice of window, offers perfect reconstruction of the original signal when no processing is involved, it might introduce additional artifacts for certain processing schemes. In the present case, when applied with a Hanning window, this approach yielded unwanted transients in the filtered signals, e.g., when one segment had a significantly higher energy content than its neighbor. Hence a weighted overlap add method (Crochiere, 1980) was applied instead. The WOLA approach is similar to the OLA, the only difference being that processed segments are windowed again by a synthesis window before being added to previous processed segments. In the present case, this removed unwanted transients in the filtered signals.

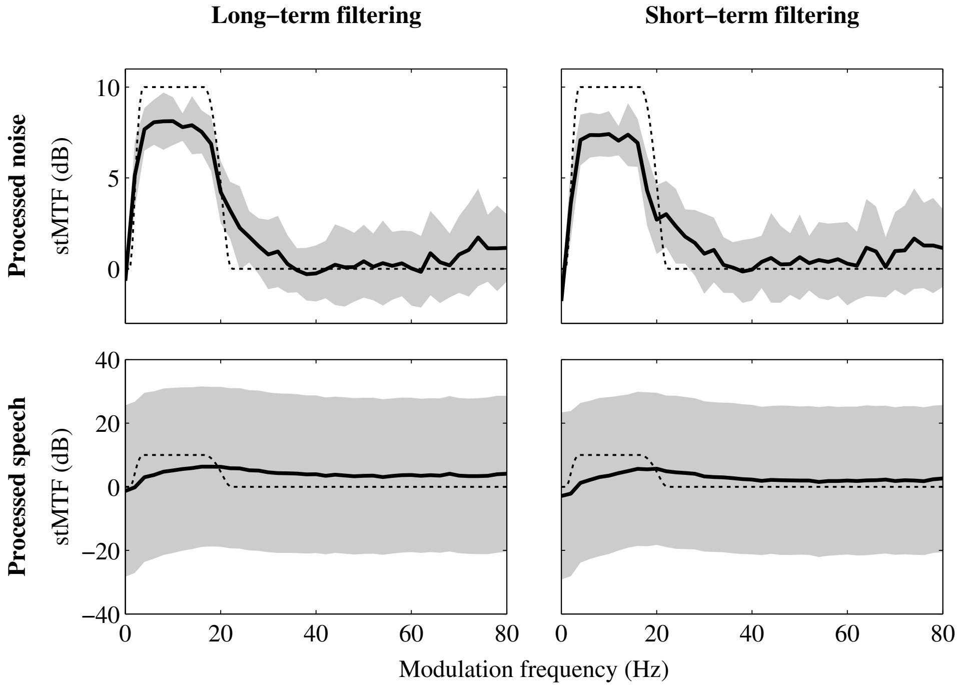


Figure 6.2: Short-time modulation transfer function average (solid lines) of modulation-filtered noise (top plots) and speech (bottom plots), when the spectrogram filtering was conducted on a long-term (left plots) or short-term basis (right plots). Additionally, one standard deviation interval (gray area) around the average is shown, as well as the modulation filter magnitude response (dashed line).

Traditionally, Hamming or Hanning windows are used in connection to WOLA. Here however, the use of such windows yielded large artifacts in the MTFs at a modulation frequency corresponding to the window repetition rate. Replacing the non-linear filtering scheme (compression followed by filtering and expansion) with standard filtering removed these artifacts. Similarly, these artifacts also disappeared when a rectangular window (i.e., no weighting) was used. Although the exact cause of the phenomenon could not be identified, these results suggest that it originates from a non-linear interaction of compression and the use of a non-rectangular window. Although the use of non-rectangular windows would have been preferable to limit spectral leakage in the modulation domain, it proved unpractical in this specific case. Hence the procedure was conducted using rectangular windows, for which WOLA performs the same operation as OLA (since no synthesis window is applied).

The window duration is subject to the same considerations as for the short-time analysis framework. Although a short window would be preferable to ensure segments of speech to be quasi-stationary, it is necessary to have a window of at least 500 ms to process modulation frequencies below 4 Hz. In the end, rectangular windows of 500 ms are used in the following.

6.1.2 Results

Figure 6.2 presents stMTFs for processed noise (top plots) and speech (bottom plots) when the filtering was performed on a long-term basis (left plots), i.e., as in chapter 5, or on a short-term

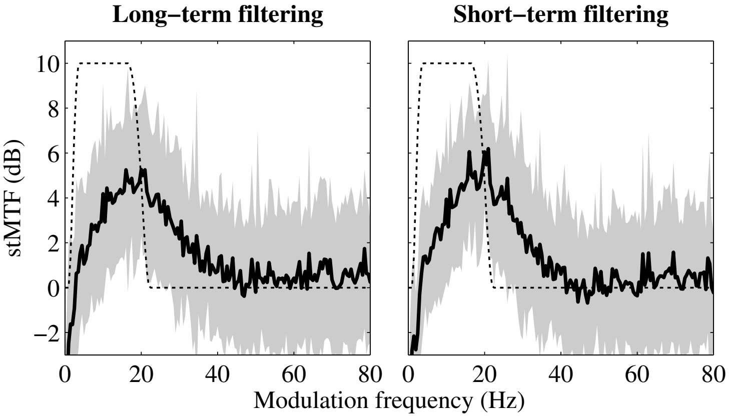


Figure 6.3: Short-time modulation transfer function average (solid lines) of modulation-filtered speech, when the spectrogram filtering was conducted on a long-term (left plot) or short-term basis (right plot), computed with a 2 s long window.

basis using an OLA approach (right plots). The processing condition is the same as one used in chapter 5, a 10 dB enhancement of the modulation content in the 4-16 Hz modulation frequency band. This condition was chosen because it yielded significant deviations of the MTF from the filter response, although not as extreme as that observed for 20 dB enhancement. To obtain sufficiently many segments to have a good estimate of the stMTF and its standard variation, the speech signal used here is a 30 s long concatenation of several sentences from the CLUE speech test (Nielsen and Dau, 2009). There is a striking similarity between the stMTFs obtained with long-term and short-term based filtering. As mentioned before, this was expected in the case of noise, due to its stationarity. However the similar observation for speech was unexpected. First, the stMTF for long-term processed speech (bottom-left plot in fig.6.2) is relatively flat and has a very large standard deviation. This is in accordance with the hypothesis which stated that for non-stationary signals such as speech, the MTF of short segments of the signal could differ significantly from the long-term MTF. However, there is no significant change in either the averages or standard deviations of the stMTF when introducing a short-time based spectrogram filtering procedure (bottom-right plot in fig.6.2), suggesting that short-term filtering of the spectrogram is not an effective solution to the variation of MTF in short local segments of speech.

Figure 6.3 presents the same results as the two bottom plots of fig.6.2, but for a longer, 2 s analysis window for the computation of the MTFs. On this time scale, the short-time spectrogram filtering approach does not have a significant impact on the standard deviation of the stMTF either.

6.1.3 Conclusion

In an attempt to reduce the variance of MTFs computed for short segments of speech, a short-term spectrogram filtering method was considered as an alternative to generate a target spectrogram. This short-term approach was shown to be inefficient at this task, yielding signals with very similar stMTFs as the previously used long-term approach, both in average and standard deviation. The fact that no significant difference was observed suggests that even though the inconsistency of the MTF of short segments of speech might be explained by the non-stationarity of said speech, the short-term approach to filtering cannot provide consistent modulation filtering across time. This could indicate that the filtering procedure on a short-time scale is subject to significant errors leading to large deviations of the local MTFs. Results from fig.6.3 show that such sources of error are likely to be localized in time, as they do not affect significantly the long-time average MTFs. One candidate for such errors is the non-negativity constraint of the filtered spectrogram, and will be addressed in the next section.

On a practical note, this study revealed that short-time analysis and processing of modulation properties poses issues that are not usually encountered in short-time analysis of sounds in general. In the modulation domain, we are concerned with much lower frequencies than we are in the traditional frequency domain. The resulting lower bound on the short segments duration is very restrictive as it does not allow the use of segments shorter than about 250 ms, which for speech would still be considered too long to be considered "quasi-stationary". This is a very restrictive constraint to a short-time approach to modulation processing, which significantly limits the scope of its application.

6.2 Non-negative realizations of filtered spectrograms

The modulation filtering framework presented in chapter 5 relies on two steps. First, each of the channels of the spectrogram of an original signal are filtered, resulting in a target spectrogram. The "target" denomination refers to the second step, where a time-domain signal is constructed such that its spectrogram is as close as possible to this target. When first considering the errors introduced in the whole modulation filtering scheme, the first part of the process, concerned with setting up a target, was ignored. It was assumed that the second step, the reconstruction from the target spectrogram, would be the limiting factor. However a significant issue regarding target spectrogram set up is that the frequency channels of the target spectrogram should correspond to individual envelopes of the outputs from a filterbank and hence, as envelopes, should be non-negative. This is a matter of concern, as the basic zero-phase modulation filters applied individually to each envelopes will not yield non-negative outputs consistent with the definition of envelope.

The reconstruction method is based on the absolute value of the resulting target spectrogram (see equations (3.20) and (3.23)), which forces non-negativity. However taking the absolute value post filtering has dramatic effect with regard to the efficiency of the filtering process. The resulting frequency response of the actual procedure (i.e., filtering followed by absolute value) differs significantly from the expected frequency response (i.e., that of the filter only). For speech, when the filter applied consists of a positive gain in a given modulation frequency range, this situation is prone to yield negative sections in the filtered envelopes.

In the study from chapter 5, this problem was circumvented by limiting the dynamic range of the envelopes through a compression scheme (raising them to the power $1/3$, filtering them, and then expanding the output back by raising it to the power 3). It was empirically found that this strategy provided more contrasted frequency responses (i.e., higher effective gain in the amplified modulation frequency band) than other means of forcing non-negativity such as absolute value or half-wave rectification. This was done at the cost of a loss of control over the modulation frequency band that was enhanced. This can be seen in fig.5.1, where the resulting modulation transfer functions (MTFs) of the processed signals significantly spread beyond the intended modulation frequency range of processing. This strategy was chosen as it was believed to be more crucial to have a significant contrast in the MTFs in order to test the influence on intelligibility. However more work was conducted towards devising a better solution to the target spectrogram setup problem. This section investigates other approaches for imposing a given frequency response, further denoted as *requisite* magnitude response, to the channels of the original signal's spectrogram that would still provide non-negative outputs.

6.2.1 Problem formulation

Given the individual channels $\{s_m\}_{1\leq m\leq M}$ of the spectrogram of a signal s and a modulation filter impulse response h, the *target* spectrogram channels $\{t_m\}_{1\leq m\leq M}$, are generated as follows:

$$t_m = s_m * h, \quad \forall m \in [1\,..M] \tag{6.1}$$

The problem investigated here can be stated as finding a suitable impulse response h such that the following two conditions are respected:

(i) The magnitude of the frequency response associated to h, $|\mathscr{F}\{h\}|$, equals the requisite magnitude response H_r

(ii) The filtered channels are non-negative signals: $t_{m,k} \geq 0, \quad \forall k \in \mathbb{Z}$ and $\forall m \in [1\,..M]$

6.2.2 Feasibility of a non-negative impulse response modulation filter

A first approach to solving the problem is to consider the general case where no assumptions are made regarding the original signal's spectrogram, i.e., where the aforementioned condition (ii) stands for any signal. This is then equivalent to dropping the channel dependency and saying the following:

> "Given a requisite magnitude response H_r, find a filter impulse response h, associated with the magnitude frequency response H_r, such that for any non-negative signal s, $s*h$ is non-negative as well."

An interesting preliminary result is that given any non-negative signal s, $s*h$ is non-negative if and only if h is non-negative. The sufficient condition is straightforward and it is clear that if h is non-negative, then $s*h$ is non-negative. The necessary condition is slightly less trivial and can be proven by contraposition. If the impulse response h is not non-negative (i.e., it contains at least one

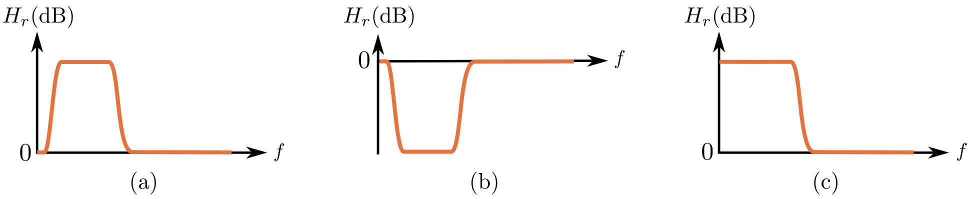


Figure 6.4: Examples of requisite filter magnitude frequency responses H_r as a function of the frequency f. Examples (a) and (b) were the designs used in the study from chapter 5. Note that H_r is given in dB, hence the zero line shows frequencies with no gain applied.

negative sample) then there exist some non-negative signals s such that $s*h$ will not be non-negative. A convincing example is to consider the signal s given by the discrete-time Dirac delta function. It is a non-negative signal, but results in $s*h = h$ which is not, by definition, non-negative. The problem therefore simplifies to the following:

> "Given a requisite magnitude response H_r, find a non-negative filter impulse response h associated with the magnitude frequency response H_r."

Non-negative impulse response filters have been a topic of study in the digital signal processing community for many decades. Such filters are generally involved in applications where non-negativity of the output of the filter is a strict requirement, e.g., "in control systems, electronic amplifiers, and many other industrial applications" (Meadows, 1972). Despite a long history of research in this field, Liu and Bauer (2010b) state that "unfortunately, the relationship between a non-negative impulse response and its corresponding frequency response is still unclear". Even though a complete understanding of such a relationship has not yet been achieved, the study in (Liu and Bauer, 2010b) provides sufficient elements for evaluating the feasibility of a non-negative impulse response modulation filter, for the particular case of our study.

Panels (a) and (b) of fig. 6.4 illustrate the class of filter magnitude frequency responses, H_r, that were used in chapter 5. We would like to realize these H_r using a non-negative impulse response filter. The very first result of (Liu and Bauer, 2010a, *Lemma 1*) provides an upper bound on H_r and states that if the impulse response of a filter is non-negative, i.e. $h[k] \geq 0$, for $k \in \mathbb{Z}$, then[1]

$$|H_r(f)|^2 \leq H_r^2(0) \tag{6.2}$$

In other words, the maximum of H_r has to be reached for $f = 0$. This result proves that designing a non-negative impulse response filter having a magnitude frequency response in the shape of fig. 6.4(a) is impossible, as the response at $f = 0$ is not a global maximum.

The design from fig. 6.4(a) is not achievable using a non-negative impulse response filter. However one could imagine a small modification, as illustrated in fig.6.4(c). Here the 0 dB gain constraint on the lowest frequency band is removed. In this case, the condition given by (6.2) would be fulfilled. However, a further result in (Liu and Bauer, 2010b) compromises the feasibility

[1] The notations have been adapted from (Liu and Bauer, 2010a) to be consistent with the rest of the description.

of the designs given by fig. 6.4(b) & (c). A consequence of their *Lemma 4* is that non-negative impulse response filters cannot have a gain of 0 dB over a given frequency range, as is the case for all the filter responses in fig.6.4. Examples of impossible designs which are given in (Liu and Bauer, 2010b, fig.4) can be related to the designs (b) and (c) presented here in fig.6.4. Although the designs illustrated in fig.6.4 could in principle be modified to comply with the conditions provided in (Liu and Bauer, 2010b), it is a requirement in this study to achieve a flat response at 0 dB in a given range of modulation frequency. Hence, trying to design a non-negative impulse response modulation filter for the purpose of our study is a dead-end.

6.2.3 Channel dependent non-negative filtering

The argumentation in the previous section is based upon having the condition (ii) defined in section 6.2.1 standing for any signal s. It was shown how this assumption led to designing non-negative impulse response filters, which was then proven impossible given the magnitude frequency responses we want to achieve. If this constraint (for *any* signal s) is released, the problem can be rephrased as follows:

> "Given a requisite magnitude response H_r, find a filter impulse response h, associated to the magnitude frequency response H_r, such that for *a given* non-negative signal s, $s*h$ is non-negative as well."

This problem is much less constrained, as it involves the design of a h that is specifically "tailored" to a given s, such that $s*h$ is non-negative and $|\mathscr{F}\{h\}| = H_r$. We suggest to reformulate the problem by circumventing the impulse response h and focusing on the resulting signal $t = s*h$:

> Given a requisite magnitude response H_r and a non-negative signal s, find a non-negative signal t such that

$$\left|\frac{\mathscr{F}\{t\}}{\mathscr{F}\{s\}}\right| = H_r \tag{6.3}$$

In this section, a method is suggested to approach this result, given that the non-negativity of t is a strict constraint and that the constraint given by (6.3) is loose, i.e., that the ratio of the Fourier magnitudes of t and s should be *as close as possible* to H_r.

Consistency and bounded multipliers

The approach proposed here is to consider the non-negative signal t as a variable in an optimization procedure which will aim to minimize an objective function based on (6.3). Before defining a suitable objective function, some considerations of the problem boundaries need to be assessed. The final goal is to process speech. As mentioned in section 6.1, speech has an intricate temporal structure. It is clear that, for the purpose of this study, this structure should be altered as minimally as possible. This consideration becomes crucial when recalling that several channels have to be filtered. Operations such as time-shift or time reversal will not change the Fourier magnitude response and hence still fulfill (6.3) in individual channels. However, mis-aligned or time-reversed channels are clearly undesirable for the reconstruction of a time signal that follows. Here, we introduce an additional constraint that aims at maintaining a certain degree of consistency between s and t. We suggest the sought signal t to take the form of a term-by-term product of the original signal s with a vector of bounded non-negative multipliers c, i.e., for a given sample k,

$$t[k] = c[k] \cdot s[k], \text{ with } 0 \leq c[k] \leq c_{max} \tag{6.4}$$

The lower bound of such multipliers, 0, ensures that t is non-negative, while the upper bound c_{max} (with $c_{max} > 1$) ensures some consistency with s is maintained. This approach, given a sufficiently low c_{max}, should maintain basic features of the original signal, such as regions of high-energy modulations or silence. The signal given by c, is then considered as a variable in an optimization algorithm.

Objective function

We define here an objective function based on (6.3):

$$\mathscr{G}(t) = \| |\mathscr{F}\{t\}| - H_r \cdot |\mathscr{F}\{s\}| \|_2^2 \tag{6.5}$$

where we introduce the bounded multipliers from (6.4):

$$\mathscr{G}(c) = \| |\mathscr{F}\{c \cdot s\}| - H_r \cdot |\mathscr{F}\{s\}| \|_2^2 \tag{6.6}$$

Because the Euclidean norm is positive definite, for any vector c, $\mathscr{G}(c) \geq 0$. $\mathscr{G}(c) = 0$, if and only if (6.3) is verified. Hence, we can find the argument that minimizes the function $\mathscr{G}$ using an optimization approach, in the same fashion as done in chapter 3. Given that the discrete signal s has a finite number of K elements, the discrete Fourier transform (DFT) applies:

$$[\mathscr{F}\{s\}]_n = \sum_{k=0}^{K-1} s_k e^{-2\pi i n \frac{k}{K}}, \forall n \in [0..K-1] \tag{6.7}$$

Deriving analytically the gradient of $\mathscr{G}$ with regard to c can be done in a way that was done for the objective functions in chapter 3, where the filterbank operator is replaced by the DFT, resulting in the following formula:

$$\nabla\mathscr{G}(c) = 2s \cdot \Re\left(\mathscr{F}\left\{(|\mathscr{F}\{c \cdot s\}| - H_r|\mathscr{F}\{s\}|) \cdot |\mathscr{F}\{c \cdot s\}|^{-1/2} \cdot \overline{\mathscr{F}\{c \cdot s\}}\right\}\right) \tag{6.8}$$

An additional constraint needs to be loosened. In situations where the requisite magnitude response H_r represents an enhancement of modulations in a given modulation frequency region, such as the design in fig.6.4(a), it is unrealistic to require a magnitude response of 0 dB at a modulation frequency of 0 Hz. An enhancement would result in a global increase of energy in the filtered signal,

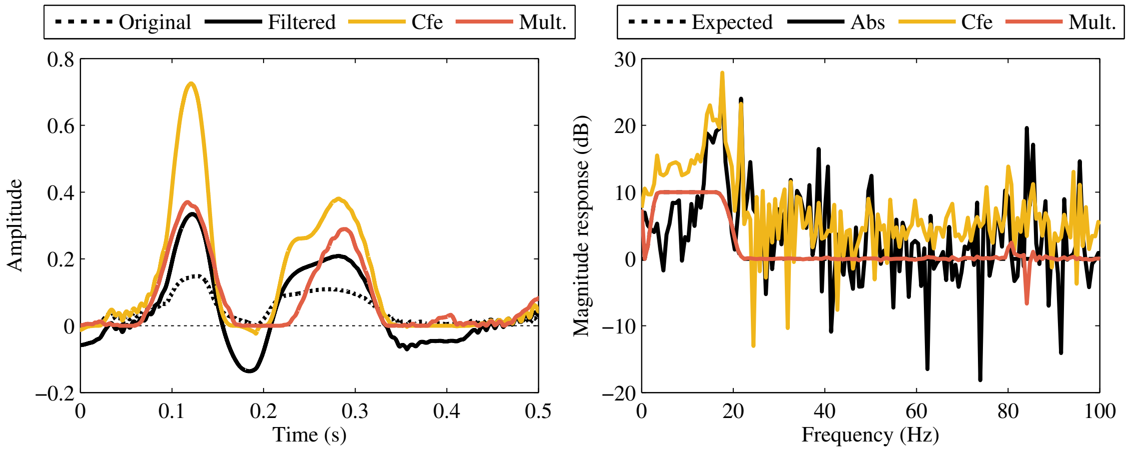


Figure 6.5: *Left panel*, waveforms of original signal (dashed line) and filtered signals using standard filtering (black line), the dynamic range compression scheme from chapter 5 (orange line) and the bounded multipliers method (red line). *Right panel*, resulting magnitude frequency responses when forcing non-negativity using the absolute value (black line), dynamic range compression (orange) and the bounded multipliers (red). The dashed line (mostly superimposed with orange line) represents the requisite filter response H_r.

and because the signals involved are non-negative, this translates to an increase of the average value (DC component). Forcing $H_r(0) = 0\text{dB}$ will result in a ill-posed overconstrained problem. It is suggested here to release the constraint on the DC component by removing its contribution in the objective function by considering the following:

$$\mathscr{G} = \|w \cdot (|\mathscr{F}\{c \cdot s\}| - H_r \cdot |\mathscr{F}\{s\}|)\|_2^2 \tag{6.9}$$

where w is a weighing vector of only ones, except for $w_0 = 0$.

A notable difference in the optimization procedure applied here, in comparison to the problems solved in chapter 3, is that the multipliers c are bounded. The optimization problem is therefore constrained, with $c \in [0, c_{max}]^K$. An implementation of the l-BFGS method for constrained optimization problems available in (Schmidt, 2008) will be applied here.

Implementation and results for one individual channel

Figure 6.5 presents results of the filtering for one channel of the spectrogram of a speech sample, for a magnitude filter response ideally presenting a 10 dB enhancement in the range of 4 to 16 Hz. The speech sample is a sentence taken from the CLUE material (Nielsen and Dau, 2009) and the channel was chosen as the one where the non-negativity requirement was the most problematic, i.e., the one presenting the largest negative sections after traditional filtering. It is centered at an audio frequency of 164 Hz and has strong modulation content in the modulation frequency range that is enhanced by the filter. The left panel of fig.6.5 presents the first 0.5 s of the original waveform of the envelope in this channel (black dashed line). The solid black line presents the output from the zero-phase filtering and exposes the issue presented here: the filtered signal has many regions where it is negative-valued. If no effort is made to force non-negativity, the reconstruction procedure that follows will consider the absolute value of the signal. The absolute value of the waveform is not plotted here as it mostly overlaps the filtered output. However, its effect on the magnitude of the frequency response is plotted in a black line on the right panel of fig.6.5, and the deviations to the expected filter response (dashed line) are significant.

The strategy used in chapter 5 was based on compressing the envelope, filtering it and then expanding it back. It was empirically found to provide an almost non-negative output while presenting a good compromise between efficiency of the filtering and control over the processed modulation frequency band. The waveform and the resulting magnitude of the frequency response for this approach, with a compression factor $p = 1/3$, are plotted in orange lines in fig.6.5. It appears from the resulting waveform that the non-negativity issue is avoided to a large extent, with only a small negative-valued segment around 0.2 s. However the effect on the frequency response is significant, as can be observed on the right panel plot. Note that fig.6.5 presents results in only one channel, for one signal. Hence, it should be kept in mind that such responses are not averaged over channels nor over many signals such as the responses presented in fig.5.1 and a direct comparison of the two should be avoided.

The results for the bounded multipliers method introduced here are presented in fig.6.5 (red lines). This result was obtained after 150 iterations of the optimization algorithm using the objective function from (6.9). The waveform is by definition non-negative, and is consistent with the original signal, in the sense that peaks and silences in both signals are occurring at approximately the same instants in time. The resulting magnitude frequency response can be observed in the right panel of fig.6.5 and mostly overlaps the requisite response H_r plotted as a dashed line. The release of the constraint on the DC offset discussed before can be observed on the outcome magnitude response since the response presents a positive gain (about 7dB) at zero frequency. These results are satisfying as it appears that the problem given by (6.3) can be solved using such an optimization approach.

Results for a target spectrogram

The suggested approach for non-negative filtering based on optimal bounded multipliers was applied individually to each channel of the spectrogram of a speech signal. The speech sample is the same sentence as used in fig.6.5. Figure 6.6 presents the resulting magnitude frequency responses on the right hand side, with individual magnitude responses in the top plot, and the magnitude frequency response averaged over all the channels in the bottom plot. The left-hand side plots correspond to the result obtained using the dynamic range compression filtering scheme from chapter 5, for the same signal. For clarity, the waterfall plots present only every second channel. However the average responses are computed based on all the channels.

It appears from the top-right plots in fig.6.6 that the suggested method based on bounded multipliers performs more consistently across channels than the dynamic range compression approach, although the constraint given by (6.3) is better satisfied in channels with lower center frequencies. Overall, the individual magnitude frequency responses plotted in 6.6 provide a solid

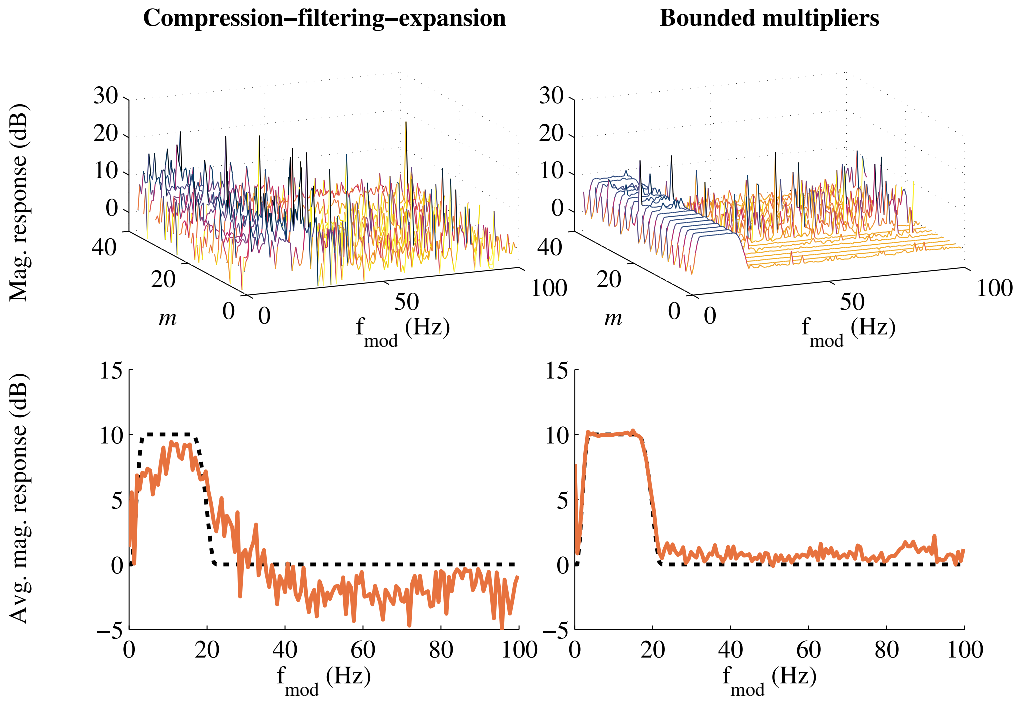


Figure 6.6: Magnitude frequency responses of filtered channels of the spectrogram of a speech sample for every channel m (top waterfall plots) and averaged over the channels (bottom plots), when the non-negative filtering strategy employs dynamic range compression (left plots) or the design of optimal bounded multipliers (right plots). For clarity, only every second channel was plotted in the waterfall plots. In the bottom plots, the requisite filter magnitude H_r is plotted as a dashed line.

argument in favor of the method proposed here: the optimal design of bounded multipliers allows for a non-negative signal to have the magnitude of its Fourier transform modified in order to achieve a given magnitude response, for signals of various bandwidths (channels with different center frequencies), to a much better extent than a traditional filtering approach combined with a method for forcing non-negativity (such as absolute value or dynamic range compression).

Reconstruction of a target spectrogram

The results presented in fig.6.6 illustrate the efficiency of the bounded multipliers method in solving the problem given by (6.3). However the final application of the non-negative realization of filtered outputs was to generate a target spectrogram consistent with the non-negativity property of a spectrogram, such that it could be inverted into a time-domain signal. Figure 6.7 presents the resulting modulation transfer functions of the time domain signals that were reconstructed from target spectrograms obtained using the dynamic range compression filtering paradigm (black line) and the multipliers approach (orange line). These results are for the same filter magnitude response (dashed line) as used in fig.6.6, and were obtained by averaging the results across 50 sentences of the CLUE material.

In chapter 5, the non-negativity of filtered spectrograms was dealt with by using a compression-filtering-expansion scheme which limited the dynamic range of the channels in the spectrogram

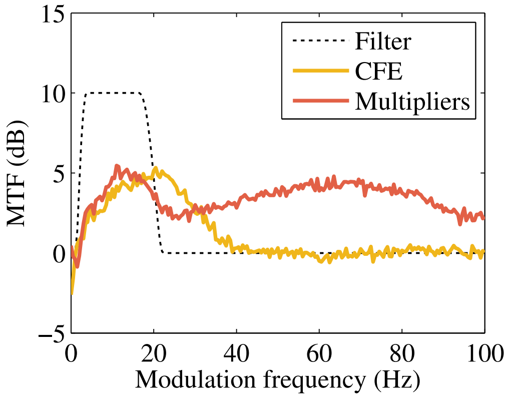


Figure 6.7: Resulting MTF for time-domain signals reconstructed from target spectrograms obtained through the compression-filtering-expansion approach (labeled CFE, yellow line) and the bounded multipliers approach (orange line). The dashed line represents the requisite filter magnitude response. These results were averaged over 50 sentences.

prior to filtering them. It was found to yield larger effective gains in the processed band than taking the absolute value post filtering, although the processed band became wider than expected. This is illustrated in fig.6.7. Signals reconstructed from target spectrograms generated using the bounded multiplier approach resulted in the orange line in fig.6.7. This method provides time-domain signals whose MTFs are significantly different from the one expected from the requisite magnitude response (dashed line in fig.6.7), with both a lower effective gain in the processed band, and a large erratic gain in the upper passband (i.e., above 20 Hz). It appears that although this approach offered a good solution to the non-negativity requirement of the channels in the target spectrogram, as illustrated in fig.6.6, it does not integrate well in the reconstruction procedure that follows. In the end, the compression-filtering-expansion scheme initially chosen actually offered a better compromise between effective processing in the processed band and a flat response at 0 dB in the upper passband.

These results illustrate that it is not sufficient to consider the problem of setting up a suitable non-negative target by itself. It has to be considered simultaneously with the reconstruction method that will follow. Both steps of target setup and reconstruction from the target introduce errors (i.e., deviations to the expected processing). The "target-related errors" are errors introduced when trying to enforce a given magnitude frequency response to a set of non-negative signals, while keeping the outputs non-negative. They are observable (e.g. in the bottom plots of fig.6.6) as the deviations of the actual response (orange line) to the filter response. Time-domain signals are then reconstructed from these target spectrograms, yielding additional errors. These "reconstruction errors" are the differences between the orange curves from fig.6.6 and the final corresponding MTF curves from fig.6.7. The results provided by these two figures form a clear proof that target-related errors and reconstruction errors are not independent. The results for the bounded multipliers, in particular, show that the method used to set up suitable target spectrograms will influence the reconstruction framework. In this case, the method seems to solve the target setup problem, but provides target spectrograms that result in poor reconstruction. Achieving a complete understanding of the relationship between the two types of errors introduced is unlikely, due to the high complexity of the problem. However, a safe assumption to explain why the bounded multipliers method does not integrate well with the following reconstruction is the fact that channels were processed independently. A filtering scheme that does not operate consistently over the channels is not likely to maintain the redundancies in the spectrogram that allow for its inversion.

6.3 Towards a target-less approach to modulation filtering

The problems presented earlier in this chapter were all stated in the context of generating a "better" target spectrogram. It appeared that setting up a suitable target is of major relevance to achieve modulation filtering based on spectrogram reconstruction. It is as crucial as the reconstruction method itself. Thus far, target spectrogram setup and reconstruction of a time-domain signal from this target were considered as two independent sub-problems. However, the choice of a given target spectrogram will condition the results obtained in the reconstruction, thus they are not independent problems. This interdependency was exposed in the previous section.

An ideal processing framework that would consider this interaction would imply a "perfect" target setup. The target spectrogram generated, independently of the processing that was performed, would be a *consistent* spectrogram, i.e., the spectrogram of an actual time-domain signal. This signal would then be recovered by the reconstruction algorithm, up to the limitations mentioned in chapter 3. How to modify an existing spectrogram to enforce given properties (e.g. a given magnitude frequency response) while still maintaining a consistent spectrogram is the problem that would have to be solved. However, current understanding of the internal redundancies in a spectrogram representation do not allow for such design.

Instead, we suggest here a conceptual approach that does not include a separate step for target setup. The modifications imposed by the target setup could be replaced by additional constraints in the reconstruction step. Instead of basing the modulation filtering method on equation (5.3) involving reconstruction of a target spectrogram $\mathbf{T}$ where the channels were generated from an original signal's spectrogram $\mathbf{E}_s$ filtered to achieve a requisite magnitude frequency response H_r, one could consider minimizing the objective function given as follows:

$$\mathscr{G}(s) = \|\mathscr{F}\{\mathbf{E}_s\} - H_r \cdot \mathscr{F}\{\mathbf{E}_{s_0}\}\|_{fro}^2 \tag{6.10}$$

where s_0 is the original signal, s the sought signal, and $\mathbf{E}_{s_0}$ and $\mathbf{E}_s$ their respective spectrograms. This approach always considers the distance (in the Fourier domain) between two consistent spectrograms, $\mathbf{E}_s$ and $\mathbf{E}_{s_0}$. Thus, it circumvents the issues implied by setting up a consistent target spectrogram. Given that the optimization problem can be solved numerically, it would provide an output signal s that is the optimal signal given a magnitude response H_r.

This approach avoids the significant constraints involved when modulation filtering using filtered spectrogram reconstruction. However the expression of the objective function in (6.10) is too intricate to derive a reasonably simple expression of the gradient, making it impossible to apply a l-BFGS optimization procedure. Further work on the analytic expression of the gradient might reveal simplifications allowing for such an implementation. It is, however, impossible to implement as such. While a complex expression of the gradient can be derived, as it does not simplify and involves nested sums over many arguments, it would result in a buildup of approximation and rounding errors that would bias the gradient value too much to obtain a functioning implementation.

7

General discussion

7.1 Results overview

Common signal analysis procedures often include the extraction of an envelope and its complementary temporal fine structure (TFS). The envelope of a signal is an intuitive concept, and various mathematical definitions of envelope extraction coexist. The envelope, as a concept, is appropriately defined only for narrow-band signals. The envelope of wide-band signals, such as speech or in general most naturally occurring sounds, is only clearly defined after decomposing such signals into narrow-band components, e.g., by means of a bank of bandpass filters. The collection of envelopes of channels at the output of a filterbank is what is then commonly referred to as the spectrogram of the signal. For narrow-band signals, and for most definitions of envelope, a certain degree of dichotomy exists in such a decomposition: envelope and TFS are both needed to faithfully represent the signal. However, spectrograms are obtained from the output of a filterbank whose filters overlap to a given extent in frequency. This overlap implies that the envelope from a given channel will convey information from neighboring channels, making the spectrogram redundant. For cases with sufficient redundancy, it was shown that the dichotomy between envelope (spectrogram) and TFS can disappear, meaning that the spectrogram, by itself, represents entirely the signal it was extracted from.

This thesis revolved around a novel approach to reconstruct time-domain signals from a multichannel envelope representation, i.e., a spectrogram. It suggested that a time-domain signal could be iteratively constructed to minimize the distance between its spectrogram and the spectrogram to reconstruct from, the *target* spectrogram. This suggestion made no specific assumptions with regard to the spectro-temporal analysis method that was adopted, i.e., which type of filterbank was used to provide sub-channels, and which envelope extraction strategy was chosen to extract the envelope in individual channels. It is in this sense a conceptual approach where applicability in practice will depend on the choice of analysis framework. It was shown how this framework could be applied to standard analysis filterbanks where outputs are given by convolution of the signal with a set of analysis windows, in the way expressed in (3.1). Such designs are common and, for this study, adequate design of these analysis windows allowed for the use of a Gammatone filterbank, whose filters' bandwidth, roll-off, and frequency spacing mimic those of human auditory filters, providing a well-accepted model of the peripheral auditory system (Glasberg and Moore, 1990).

Many alternatives exist for the choice of the envelope extraction strategy. As the envelope is mainly an intuitively defined attribute of a signal, there are several mathematical definitions (see section 2.1). The present study was focused on auditory-inspired spectrogram representations, hence the choice of a Gammatone filterbank and a class of inner hair-cell (IHC) envelope models based on channel half-wave rectification followed by low-pass filtering. Additionally, The more traditional spectrogram representation consisting in the magnitude of the short-time Fourier transform (STFT) coefficients was also investigated as it is the most common case in previous literature. Chapter 3 detailed how the proposed general framework could apply to the reconstruction of time-domain signals from such auditory spectrograms and "traditional" spectrograms. These two cases proved themselves practical, as the mathematics behind the minimization problem could be simplified, allowing for an efficient implementation. Reconstruction from STFT-based spectrograms has been extensively studied, with efficient methods specifically tailored to this problem having been developed. A simple modification of the proposed framework to account for the perceptual consequences of the reconstruction error allowed to generate time-domain signals showing better accuracy than more traditional methods. It appeared that this approach led to a reduction in the intrinsic limitations that the problem of STFT-based spectrogram inversion usually faces, although a larger-scale study would be needed to validate this result over different configurations of the STFT. Conversely, application to IHC envelope-based spectrograms has received little attention, with only few mentions in the literature. It was shown how these limitations that are usually faced in the reconstruction from STFT-based spectrograms completely disappear when an IHC-based envelope is used, allowing for far more accurate, near-perfect, reconstruction. Accurate reconstructions obtained for both definitions of the envelope provide evidence supporting the main hypothesis, that the spectrogram faithfully represents signals.

In particular, near-perfect reconstruction of speech signals from a spectrogram computed using the IHC envelope model from (Dau *et al.*, 1996a) provided numerical evidence of a strong interdependency between IHC envelope and temporal fine structure (TFS). This would suggest that such simple models of cochlear processing would not discard any TFS-related information. To further investigate this phenomenon, the study conducted in chapter 4 assessed the robustness of the reconstruction framework with regards to the parameters involved in the extraction of the IHC envelope. This study was based on a particular stimulus, a complex tone introduced in (Santurette and Dau, 2011), which presented peculiar properties. The complex tone, based on five sinusoidal components equally spaced in frequency, exhibited a periodic envelope. However, local maxima of the envelope were not aligned with local maxima in the TFS. The results from (Santurette and Dau, 2011) showed that the pitch perception of this tone was mostly related to the cues pertaining to the TFS, even though the high frequency content of the tone (the components spanning a range from around 5.8 kHz to 8.2 kHz) is usually assumed to induce an envelope-based pitch perception. Among the arguments raised in their discussion, Santurette and Dau (2011) mention TFS recovery from the output of hair-cell transduction as a possible explanation for the perceived pitch of their tones. Here, to investigate the degree to which TFS-related information was retained in a modeled internal representation, this tone was reconstructed from its IHC-based spectrogram for various parameter settings of the IHC envelope extractor, and the accuracy of the reconstruction was assessed. It was shown that perfect reconstruction of the original signal, hence of TFS-related information, was achievable for IHC models involving low order or high cutoff frequency low-pass filters, such as the ones from (Dau *et al.*, 1996a) or (Lindemann, 1986). Increasing the sharpness of this filter first yielded reconstructed signals where the TFS cues were recovered, although the waveforms were not perfectly reconstructed. This was also the case for the IHC model from (Breebaart *et al.*, 2001). Further increase of the filter's order or decrease of its cutoff frequency finally resulted in noisy reconstructed signals with no evidence of any TFS recovery. This study therefore suggested that the auditory spectrograms obtained from a Gammatone filterbank and IHC envelope models computed with physiologically plausible parameters, as the three models mentioned here, still preserved TFS-related information. While we do not argue that this information is indeed extracted by higher processing schemes in the auditory pathway (we do not provide physiological evidence for the results from (Santurette and Dau, 2011)), we provide numerical evidence that TFS-related information is present in this class of models.

The proposed framework, applied to the reconstruction of time-domain signals from original spectrograms, provided insights towards the understanding of the relationship between envelope and TFS in a multiple channel representation. Another application of the reconstruction tool lies in modulation filtering, or more generally, modulation manipulation. A time-domain signal could be reconstructed from a *modified* spectrogram, i.e. a spectrogram which has been subject to some manipulations prior to the reconstruction. In doing so, the reconstructed signal would be expected to exhibit changed properties in its modulation content concordant with the manipulations performed on the spectrogram. Many studies in the literature have considered the modulation content of speech to be one of the key elements for its understanding. For example, Jørgensen and Dau (2011) suggested a model to estimate speech intelligibility in a range of adverse conditions, which is based solely on the concept of signal-to-noise ratio in the modulation domain. Chapter 5 investigated how the spectrogram reconstruction framework could be applied to perform modulation filtering and generate, by modulation filtering noise or speech, mixtures with a controlled signal-to-noise ratio in the modulation domain. Such signals would be ideal for a systematic evaluation of the model from (Jørgensen and Dau, 2011). The intelligibility of mixtures obtained with either modulation filtered noise or speech was measured on human subjects, and compared to predictions from the model proposed in (Jørgensen and Dau, 2011). A good match between predictions and data was obtained for mixtures of modulation filtered noise and unprocessed speech. This suggested that modulation filtering of noise was a suitable approach to control the signal-to-noise ratio in the modulation domain of a mixture, and that the modulation filtering framework, when applied to noise, was relatively efficient. Conversely, significant deviations between predictions and data were observed for mixtures of modulation filtered speech and unprocessed noise. This was partly explained by the lower efficiency of the modulation filtering framework when applied to speech.

Chapter 6 addressed issues limiting the efficiency of modulation filtering using a spectrogram reconstruction method. This approach, by considering the whole spectrogram of a signal, operates on long time scales. Shorter sections of processed signals might therefore not exhibit the requested modulation properties. To test this, a short-time based implementation for spectrogram filtering was suggested. However, the modulation-frequency content of spectrogram channels is much lower than the usual audio frequency domain. Unfortunately, processing low modulation frequencies (below 4 Hz) is not compatible with the extraction and processing of very short segments (under 250 ms long) of signals, making a short-time approach to spectrogram filtering impractical. However, a significant compromise on the duration of segments allowed for an implementation. No significant difference was observed between signals reconstructed from short-term vs. long-term filtered spectrograms, suggesting that other sources of error are responsible for short sections of the processed signal that do not exhibit the expected and consistent changes in modulation. The sub-channel envelopes of speech signals undergo large variations in amplitude in relatively short time. Filtering these channels to obtain a target spectrogram can yield negative sections in the output, which are inconsistent with the definition of the envelope as a non-negative signal. An ideal solution to this problem would have been the design of a non-negative impulse response filter. However, this was shown to be impossible given the desired types of magnitude responses. A channel-dependent approach was proposed to produce non-negative outputs that present the expected magnitude frequency response. While this approach was shown to solve this specific problem, it disrupts the subsequent reconstruction step. Processing the sub-channel envelopes independently appears to compromise the original channel interaction which allows recovery of the TFS. Finally, a conceptual approach to circumvent this problem was suggested, but could not be implemented practically due to significant mathematical challenges. If these challenges could be overcome, it could form a good basis for future improvement of the method.

7.2 Additional discussions

In this final section, we provide further discussions in the form of the answers to four questions. These questions relate to the modulation filtering approach that was considered in chapters 5 and 6:

- to what cases can it be applied?
- what are its limitations and in which directions should further investigations be conducted?
- why the method was not implemented with the "well-behaved" modern definitions of envelopes presented in chapter 2?
- how does it compare to previous attempts to perform modulation filtering?

Which scenarios can benefit best from modulation filtering?

In chapter 5, we performed modulation filtering on individual elements of a mixture of speech and noise, as our objective was to control the SNR$_{env}$ of the mixtures. Having access to either the speech or the noise is uncommon in most practical applications. Being able to process speech prior to its mixture with noise might be possible in a few scenarios, for example when speech is delivered through loudspeakers in an airport or a train-station which presents an ambient noisy background. In these specific cases, one could imagine processing modulations in the recorded speech so that its intelligibility, once mixed with ambient noise, would be higher. But in such cases, the speech prior to processing is, at worse, "conversational speech". A well-designed modulation processing scheme might be able to improve the intelligibility of said speech, but at best to the level of "clear speech" or maybe "hyperarticulated speech". When processing speech only, the improvement cannot be expected to be better than what an actual speaker could perform, if he/she was confronted with this given noisy background. For research applications though, the interest in processing speech or noise prior to mixing is clear. As was carried out in chapter 5, this allows to generate stimuli with controlled modulation properties for investigating their perceptual influence.

A much more practical scenario, which was not thoroughly investigated in this thesis, is the processing of the mixture itself. Typically, this case is interesting for hearing aid development. A microphone will pick up the mixture of speech and background noise and the hearing aid will not have access to individual elements of the mixture for processing them. Noise removal in the modulation domain has been performed in the past and led to significant improvement of intelligibility (e.g., Wójcicki and Loizou, 2012). Modulation filtering approaches would have the advantage over noise removal methods of being systematic (i.e., independent of the input signal) and hence maybe simpler to implement in a hearing assistive device.

Which directions to take for improving further this modulation filtering approach?

In chapters 3 and 4, satisfying results were obtained with regard to the accuracy of reconstructed signals. This indicates that, given the actual spectrogram of a signal, i.e., a *consistent* spectrogram, the reconstruction method performs well and can recover the original time-domain signal accurately. In chapter 5 the results were more mixed. Although it appeared that the modulation content of noise could be manipulated effectively, processing speech led to artifacts which proved detrimental to intelligibility. As mentioned in the discussion of this chapter (section 5.4.3), two sources of artifacts can be identified and both regard the processing scheme prior to the spectrogram reconstruction. This target spectrogram generation does not account for the following:

- the internal characteristics of a *speech* spectrogram, i.e., the resulting target spectrogram might correspond to distorted speech instead of clear speech.
- the consistency of the resulting spectrogram, i.e., the filtered spectrogram is not associated with any time-domain signal.

Hence, the processing performed in chapter 5 could be considered naive. A more informed processing scheme with regard to the two aforementioned points is more likely to improve the results. Thus, investigating "smarter" ways to generate target spectrograms is a promising direction for improving the modulation filtering framework.

There is, however, a fundamental limit to this approach. The spectrogram extraction operator, by associating say an M-channel spectrogram to any time-domain signal of say L samples, maps the domain $\mathbb{R}^L$ to the M-times higher dimensional domain $\mathbb{R}^{M\times L}$. Assuming that the spectrogram extraction is an injective operation (the spectrogram faithfully represents the signal), it means that there is a probability of unity that any set of coefficients from $\mathbb{R}^{M\times L}$ is *not* part of the image of the spectrogram operator, i.e. that it is *not* a consistent spectrogram. This means that any arbitrarily manipulated spectrogram will not result in a consistent spectrogram, and hence cannot be exactly inverted back to a time-domain signal. In other words, arbitrary modifications of the modulation content of a signal cannot be performed with perfect accuracy; an *approximation* has to be made at some point. This compromise has to be considered when developing better methods for target spectrogram generation.

Would the use of more "modern" envelope definitions overcome the limitations faced for modulation filtering?

In section 2.1.3, we described studies which aimed at refining the definition of envelope in order to overcome the limitations faced by the Hilbert envelope. As such envelopes "behave" better, the question of using them in the context of modulation filtering comes naturally. Signed envelope (Cohen *et al.*, 1999; Li and Atlas, 2004) or complex-valued envelope (e.g., Atlas *et al.*, 2004) are based on estimation of carrier wave frequency in each channel and therefore the implementation in the framework of our method is difficult. If we assume anyway that the implementation is feasible, there is no doubt that modulation filtering carried out with these definitions of envelopes could be more accurate (see, e.g., results from (Clark and Atlas, 2009)) as it would at least circumvent the non-negativity issue discussed in section 6.2. However, there are two arguments against the use of such envelopes in the context of this thesis, both relating to a certain degree of discrepancy that these definitions present with regard to the intuitive, *perceptual* envelope.

First, discrepancies in the frequency content of the envelopes can be illustrated by the example presented in fig.1.1. When the 440 and 444 Hz tones are superimposed, we perceive a pulsating tone at a rate of 4 Hz, corresponding to the absolute value of the cosine decomposition in (1.2). However the cosine component itself, which would be the signed envelope, has a frequency of 2 Hz (the mean between the frequencies of the two components). The signed envelope, in that case, does not reflect our perception. As there is no mapping between frequencies of the "perceived" envelope and "mathematical" envelope, accurate and controllable manipulation of the *perceptual* envelope is not possible. Although "mathematical" modulation filtering would be possible, and possibly more accurate than what was achieved here, "perceptual" modulation filtering would not. And in the context of this thesis, we are interested in the latter.

A second limitation originates in the definition of such envelopes holding for very narrow-band channels of the filterbank. In an auditory-inspired filterbank (e.g., Patterson *et al.*, 1988) channels with high center frequency will have a wide bandwidth. The complex-valued envelope used in (Clark and Atlas, 2009) is defined from its associated carrier. The carrier wave itself is defined as a complex exponential at a frequency that is the "spectral center of gravity" of the sub-band signal in a close vicinity around each instant in time. In wider-band channels, the instantaneous frequency can drift significantly from this spectral center of gravity. These changes in frequency are carried in the envelope. It was observed in such cases that the real-part and imaginary-part of the complex-valued envelope could vary quickly, although its magnitude was almost remaining constant: the complex-valued envelope is rotating around the origin of the complex plane without significant changes in its magnitude. This strongly impairs its applicability to modulation filtering: in that case, applying a low-pass filter to the complex-valued envelope would essentially set the output to zero, although the perceived envelope should not have been affected by the filter. This problem is very similar to the first limitation developed above, though it is not directly caused by the definition of the envelope but rather by using it outside of its valid range (i.e., not for a narrow-band filterbank).

How does the modulation filtering framework presented here compare to "vocoding" approaches?

As mentioned in the second point of this section, a perfectly accurate implementation of any *arbitrary* manipulation of the modulation content is impossible. Some approximation has to be made. The main difference between our approach and "vocoding" approaches lies in where in the process this approximation is made. Here, by vocoder approaches, we include any of the previous methods in the literature which involved the use of extracted envelopes as modulators for a newly generated carrier, be it a narrow-band noise or a sine-wave of the instantaneous frequency or channel's center frequency. This obviously includes the literal vocoders studies from (Dudley, 1939), (Flanagan *et al.*, 1965) and (Shannon *et al.*, 1995), but also the psycho-acoustically motivated studies from (Ghitza, 2001), (Smith *et al.*, 2002), (Zeng *et al.*, 2004), (Gilbert and Lorenzi, 2006), and following. Such vocoding approaches amount to providing a model of the input signal in a domain where it can be more easily manipulated. For example, the "model" of the signal suggested by Ghitza (2001) consists in dichotically interleaving channels generated by modulating a pure cosine carrier by the speech envelope in that channel. In that domain, envelope recovery is eliminated and Ghitza (2001) shows that manipulation of the speech envelope is made possible. The main advantage of such methods is that, if the vocoding strategy (i.e., the model of the signal) is well-designed, processing will be performed flawlessly *in that domain*. The approximation, in that case, lies in the time-frequency representation of the signal (i.e., the model of the signal).

In contrast, the method we suggest is based on an actual spectrogram and not a "model" of the signal. What is approximated is not the signal, as it is for the vocoder, but rather the processing that is performed. When manipulating the actual signal, we do not perform the required manipulations exactly, but attempt to get as close as possible to it. The advantage is that the envelope recovery issue mentioned in (Ghitza, 2001) is circumvented. As the spectrogram reconstruction is based on minimizing a criterion itself defined in the spectrogram domain, the envelope recovery is indirectly taken into account: one could say that the stimuli is manipulated directly in the recovered envelope domain. Consequently, an additional advantage is that, while vocoding approaches try to reduce interaction between channels (e.g., the interleaved channels in (Ghitza, 2001)) in order to limit the band-widening effects of the Hilbert envelope, our approach fully accounts for it and hence can apply to more realistic models of auditory filterbanks that include a large amount of overlap between channels.

Overall, vocoding approaches are probably a safer choice for simple manipulations in psychoacoustical studies, as they offer better control over the manipulations performed. But modulation filtering through spectrogram reconstruction is a promising technique. With further understanding of the intricate structure of auditory spectrograms, the target spectrogram generation could be largely improved from the simple processing carried out in chapter 5 of this thesis. Eventually, it could allow for generating stimuli with requested modulation properties *at the output of the cochlea*, and without making approximations on the input signal.

Bibliography

Achan K., Roweis S.T., Frey B.J. (2004). Probabilistic inference of speech signals from phaseless spectrograms. *Advances in Neural Information Processing Systems (NIPS)*, 16

Allen J.B. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, Sig. Process.*, 25(3):235–238.

ANSI S3.5 (1997). Methods for the calculation of the speech intelligibility index. Acoustical Society of America, New York.

Atlas L., Li Q., Thompson J. (2004). Homomorphic modulation spectra. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2:761.

Baer T., Moore B.C.J. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *J. Acoust. Soc. Am.*, 94(3):1229–1241.

Balan R., Casazza P., Edidin, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. A.*, 20(3):345–356.

Balan R. (2010). On signal reconstruction from its spectrogram. *Conf. Info. Sci. Sys.*, 2010.

Bargmann V. (1961). On a Hilbert space of analytic functions and an associated integral transform. *Commun. Pure Appl. Math.*, 14:187–214.

Beauregard G.T., Zhu X., Wyse L. (2005). An efficient algorithm for real-time spectrogram inversion. *Proc. Int. Conf. Digital Audio Effects DAFx*, 116–118.

Bodmann B.G., Hammen N. (2013). Stable phase retrieval with low redundancy frames. *arXiv preprint*, arXiv:1302.5487.

Bouvrie, J. and Ezzat, T. (2006). An incremental algorithm for signal reconstruction from short-time fourier transform magnitude. *Ninth Int. Conf. Spoken Language Processing*.

Breebaart J., Van De Par S., Kohlrausch A. (2001). Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.*, 110(2):1074–1088.

Candes E., Strohmer T.,Voroninski V. (2011). PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *arXiv preprint*, arXiv:1109.4499

Chassande-Mottin E., Daubechies I., Auger F., Flandrin P. (1997). Differential reassignment. *IEEE Sig. Proc. Letters*, 4(10):293–294.

Chi T., Ru P., Shamma S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.*, 118(2):887–906.

Clark P., Atlas L. (2009). Time-frequency coherent modulation filtering of nonstationary signals. *IEEE Trans. Signal Proc.*, 57(11):4323–4332.

Cohen L, Loughlin P., Vakman D. (1999). On an ambiguity in the definition of the amplitude and phase of a signal. *Signal Processing*, 79(3):301–307.

Cooke M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119(3):1562–1573.

Crochiere R.E. (1980). A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(1):99–102.

Dau T., Püschel D., Kohlrausch A. (1996a). A quantitative model of the effective signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622.

Dau T., Püschel D., Kohlrausch A. (1996b). A quantitative model of the effective signal processing in the auditory system. II. Simulations and measurements. *J. Acoust. Soc. Am.*, 99(6):3623.

Decorsière R., Søndergaard P.L., Buchholz J., Dau T. (2011). Modulation filtering using an optimization approach to spectrogram reconstruction. *Proc. Forum Acusticum*

D. of Veterans Affairs (2006). Speech recognition and identification materials. Disc 4.0 (CD). *Department of Veterans Affairs Medical Center*, 2006.

Drullman R., Festen J., Plomp R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064.

Dubbelboer F., Houtgast T. (2007). A detailed study on the effects of noise on speech intelligibility *J. Acoust. Soc. Am.*, 122(5):2865–2871.

Dudley H. (1939). Remaking speech. *J. Acoust. Soc. Am.*, 11(2):169–177.

Dugundji J. (1958). Envelopes and pre-envelopes of real waveforms. *IRE Trans. on Information Theory*, 4(1):53–57.

Elliott T. M., Theunissen F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* 5(3):e1000302.

Ewert S., Dau T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J. Acoust. Soc. Am.* 108(3):1181–1196.

Festen J., Plomp R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.* 88(4):1725–1736.

Fienup J.R., Wackerman C.C. (1986). Phase-retrieval stagnation problems and solutions. *J. Opt. Soc. Am. A*, 3(11):1897–1907.

Flanagan J.L., Meinhart D., Golden R.M., Sondhi M. (1965). Phase vocoder. *J. Acoust. Soc. Am.*, 38(5):939–940.

Flanagan J.L., Golden R.M. (1966). Phase vocoder. *Bell System Technical Journal*, 45:1493–1509.

French N., Steinberg J. (1947). Factors governing intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119.

Gabor D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–457.

Gerchberg R.W., Saxton W.O. (1972). A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*,35(2):237–250.

Ghitza O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J. Acoust. Soc. Am.*,110(3):1628–1640.

Gilbert G., Lorenzi C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *J. Acoust. Soc. Am.*,119(4):2438–2444.

Glasberg B.R., Moore B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47(1):103–138.

Green D. M., Swets J. A. (1988). Signal detection theory and psychophysics Penisula Puplishing , Los Altos California.

Griffin D., Lim J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(2):236–243.

Harris F.J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83.

Hayes M., Lim J., Oppenheim A. (1980). Signal reconstruction from phase or magnitude. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(6):672–680.

Heinz M.G., Swaminathan J. (2009). Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *J. Assoc. Res. Otolaryngol.*, 10(3):407–423.

Hopkins K., Moore B.C.J., Stone M.A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *J. Acoust. Soc. Am.*, 123(2):1140–1153.

Houtgast T., Steeneken H. J. M. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66–73.

ITU-T P.862 (2001). Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *International Telecommunication Union, Geneva, Switzerland.*

Jørgensen S., Dau T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, 130:1475–1487.

Jørgensen S., Ewert S. D., Dau T. (2013). A multi-resolution envelope power based model for speech intelligibility. *J. Acoust. Soc. Am.*, 134:??–??.

Le Roux J., Kameoka H., Ono N., Sagayama S. (2010). Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. *Proc. Int. Conf. Digital Audio Effects DAFx*, 10:397–403.

Le Roux J., Ono N., Sagayama S. (2008). Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. *Proc. Statistical and Perceptual Audition*, 23–28.

Li Q., Atlas L. (2004). Over-modulated AM-FM decomposition. *Proc. SPIE*, 5559:172.

Li Q., Atlas L. (2005). Properties for modulation spectral filtering. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 4:521–524.

Licklider J., Pollack I. (1948). Effects of differenciation, integration, and infinite peak clipping upon the intelligibility of speech. *J. Acoust. Soc. Am.*, 20(1):42–51.

Lindemann W. (1986). Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6):1608–1622.

Liu D.C., Nocedal J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1):503–528.

Liu Y., Bauer P.H. (2010-a). Fundamental properties of non-negative impulse response filters. *IEEE Trans. Circuits Syst.*, 57(6):1338–1347.

Liu Y., Bauer P.H. (2010-b). Frequency domain limitations in the design of nonnegative impulse response filters. *IEEE Trans. Signal Process.*, 58(9):4535–4546.

Logan B.F. (1977). Information in the zero crossings of bandpass signals. *Bell System Technical Journal*, 56:487–510.

Lorenzi C., Gilbert G., Carn H., Garnier S., Moore B.C.J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc. National Academy of Sciences*, 103(49):18866.

Loughlin P., Tacer B. (1996). On the amplitude- and frequency-modulation decomposition of signals. *J. Acoust. Soc. Am.*, 100(3):1594–1601.

Ludvigsen C., Elberling C., Keidser G. (1993). Evaluation of a noise reduction method-comparison between observed scores and scores predicted from STI. *Scand. Audiol. Suppl. 38*, 22:50–55.

Meadows N.G. (1972). In-line pole-zero conditions to ensure nonnegative impulse response for a class of filter systems. *Int. J. Control*, 15(6):1033–1039.

Moore B.C.J., Glasberg B.R. (1989). Mechanisms underlying the frequency discrimination of pulsed tones and the detection of frequency modulation. *J. Acoust. Soc. Am.*, 86(5):1722–1732.

Moore B.C.J., Glasberg B.R. (1996). A revision of Zwicker's loudness model. *Acta Acust. united Ac.*, 82(2):335–345.

Moore B.C.J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J. Assoc. Res. Otolaryngol.*, 9(4):399–406.

Nielsen J. B., Dau T. (2009). Development of a danish speech intelligibility test. *Int. J. Audiol.*, 48:729–741.

Nilsson M., Soli S. D., Sullivan J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95(2):1085–1099.

Paliwal K., Wójcicki K., Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Commun.*, 52:450–475.

Patterson R.D., Nimmo-Smith I., Holdsworth J., Rice P. (1988). An efficient auditory filterbank based on the gammatone function. *APU report.*

Payton K., Braida L. (1999). A method to determine the speech transmission index from speech waveforms. *J. Acoust. Soc. Am.*, 106(6):3637–3648.

Payton K., Uchanskbi R. M., Braida L. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.*, 95(3):1581–1592.

Rhebergen K. S., Versfeld N. J., Dreschler W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J. Acoust. Soc. Am.* 120(6):3988–3997.

Santurette S., Dau T. (2007). Binaural pitch perception in normal-hearing and hearing-impaired listeners. *Hear. Res.*, 223(1–2):29–47.

Santurette S., Dau T. (2011). The role of temporal fine structure for the low pitch of high-frequency complex tones. *J. Acoust. Soc. Am.*, 129(1):282–292.

Schimmel S., Atlas L. (2005). Coherent envelope detection for modulation filtering of speech. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1:221–224.

Schmidt M. (2005). minFunc: L-BFGS algorithm implementation. Available from http://www.di.ens.fr/∼mschmidt/Software/minFunc.html

Schmidt M. (2008). minConf: L-BFGS algorithm implementation for constrained problems. Available from hhttp://www.di.ens.fr/∼mschmidt/Software/minConf.html

Sell G., Slaney M. (2010). Solving demodulation as an optimization problem. *IEEE Trans. Acoust., Speech, Signal Process.*, 18(8):2051–2066.

Shannon R.V., Zeng F.G., Kamath V., Wygonski J., Ekelid M. (1995). Speech recognition with primarily temporal cues. *Science*, 270:303–304.

Sheft S., Ardoint M., Lorenzi C. (2008). Speech identification based on temporal fine structure cues. *J. Acoust. Soc. Am.*, 124(1):562–575.

Slaney M., Naar D., Lyon R.E. (1994). Auditory model inversion for sound separation. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2:77–80.

Slaney M. (1995). Pattern playback from 1950 to 1995. *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, 4:3519–3524.

Smith Z.M., Delgutte B., Oxenham A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416:87–90.

So S., Paliwal K. (2011) Modulation-domain kalman filtering for single-channel speech enhancement. *Speech Commun.* 53:818–829.

Steeneken H.J.M., Houtgast T. (1980). A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67:318–326.

Stevens S.S. (1957). On the psychophysical law. *Psychol. rev.*, 64(3):153–181.

Sturmel N., Daudet L. (2011). Signal reconstruction from STFT magnitude: a state of the art. *Proc. Int. Conf. Digital Audio Effects DAFx*, 375–386.

Sun D.L., Smith III J.O. (2012). Estimating a signal from a magnitude spectrogram via convex optimization. *133rd Conv. Audio Eng. Soc.*, arXiv:1209.2076.

Søndergaard P.L., Majdak P. (2013). The auditory-modeling toolbox. *The technology of binaural listening*, ch.2, - *in press*.

Søndergaard P.L., Torrésani B., Balazs P. (2012). The Linear Time Frequency Analysis Toolbox. *Int. J. Wavelets Multi.*, 10(4).

Vakman D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Trans. Signal Process.*, 44(4):791–797.

Wagener K., Josvassen J.L., Ardenkjær R. (2003). Design, optimization and evaluation of a danish sentence test in noise. *Int. J. Audiol.*, 42:10–17.

Waldspurger I., d'Aspremont A., Mallat S. (2012). Phase recovery, maxcut and complex semidefinite programming. *arXiv preprint*, 1206.0102.

Wójcicki K. K., Loizou P. C. (2012). Channel selection in the modulation domain for improved speech intelligibility in noise. *J. Acoust. Soc. Am.*, 131(4):2904–2913.

Zeng F., Nie K., Liu S., Stickney G., Del Rio E., Kong Y., Chen H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues (L). *J. Acoust. Soc. Am.*, 116(3):1351–1354.

Zhu X., Beauregard G.T., W. L. (2006). Real-time iterative spectrum inversion with look-ahead. *IEEE Int. Conf. Multimedia Expo*, 229–232.

Zwicker E., Scharf B. (1965). A model of loudness summation. *Psychol. rev.*, 72(1):3–26.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modul-ated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Christiansen*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

The end.

To be continued...

A common way of analyzing signals in a joint time-frequency domain is found in the spectrogram, which can be interpreted as a multi-channel envelope representation of the signal. The envelope, as it reflects slow changes in the amplitude of a signal, cannot by itself fully represent a signal. However there is evidence that the spectrogram, because it involves multiple channels, could be a faithful representation of the signal. In this work, a method is suggested to recover audio signals from spectrograms computed for different definitions of the envelope. For auditory-motivated spectrograms, assessing the accuracy of reconstructed signals provides insights in the informational content of the cochlear representation. Additionally, more traditionally defined spectrograms can be manipulated before reconstruction, allowing for temporal modulation filtering. The influence of modulation filtering applied to either the speech or noise component of a mixture on its intelligibility is investigated, resulting in strategies for modest intelligibility enhancement and exhibiting limitations to overcome in future work.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk

ISBN 978-87-92465-xx-x