Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices

Steven Marco Schimmel

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree: Electrical Engineering
This is to certify that I have examined this copy of a doctoral dissertation by

Steven Marco Schimmel

and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

Chair of the Supervisory Committee:

________________________
Les E. Atlas

Reading Committee:

________________________
Les E. Atlas

________________________
Jeff Bilmes

________________________
Henrique S. Malvar

Date: ______________________
In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with “fair use” as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, or to the author.

Signature______________________________

Date______________________________
Abstract

Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices

Steven Marco Schimmel

Chair of the Supervisory Committee:
Professor Les E. Atlas
Electrical Engineering

Natural signals such as speech and music can be faithfully represented by low frequency modulators which modulate higher frequency carriers. In this work, a principled signal processing framework for modulation frequency analysis and synthesis is developed. Its core components are a filterbank that separates a broadband signal into narrowband subbands, and a carrier estimator or envelope detector that decomposes each subband into a carrier and a modulator. A distinction is made between “incoherent” envelope detectors, such as the magnitude or the Hilbert envelope, and “coherent” carrier estimators, such as the instantaneous frequency carrier.

Several limitations of incoherent detectors are demonstrated, and three coherent carrier estimators are proposed and evaluated. Practical aspects of modulation filtering such as modulation filtering effectiveness and signal reconstruction from a modified time-frequency representation are also addressed.

The modulation analysis framework is applied to the problem of talker separation in the context of hearing devices such as hearing aids and cochlear implants. It is demonstrated that two co-channel talkers are largely non-overlapping in the modulation frequency domain. In preliminary experiments, talker separation is achieved with a supervised incoherent modulation masking approach, and an optimal coherent modulation filtering approach.

However, both approaches require more information about the target talker than would
be available in hearing device applications. Therefore, a target talker enhancement technique is proposed which requires only knowledge of the target talker’s fundamental frequency range. It uses incoherent modulation analysis to determine coefficients for a time-varying acoustic filter that enhances the target talker and suppresses an interfering talker, whose fundamental frequency range must also be known.

Several weaknesses of this incoherent talker enhancement technique are pointed out, and a novel coherent modulation filtering approach to target talker enhancement is proposed. The coherent approach uses custom implementations of the carrier estimator and the modulation filter, in order to be more robust to noise and free of modulation filtering artifacts. The coherent approach to talker enhancement is evaluated in a subjective listening test on normal hearing subjects over a cochlear implant simulation and on hearing aid users. The test shows a modest improvement in intelligibility for coherently enhanced speech.
# TABLE OF CONTENTS

List of Figures ................................................................. iii

List of Tables ................................................................. v

Chapter 1: Introduction ..................................................... 1
  1.1 Modulation analysis and filtering ...................................... 1
  1.2 Application to hearing devices ........................................ 3
  1.3 Organization ............................................................ 5

Chapter 2: Background ....................................................... 7
  2.1 Introduction ............................................................. 7
  2.2 The modulation concept in psychoacoustics ......................... 7
  2.3 Modulation analysis and filtering in signal processing .......... 12
  2.4 Blind source separation ............................................... 20
  2.5 Summary ............................................................... 25

Chapter 3: Theory ............................................................ 26
  3.1 Introduction ............................................................. 26
  3.2 Modulation transforms ................................................. 26
  3.3 Representations ........................................................ 36
  3.4 Inverse modulation transforms ....................................... 38
  3.5 Signal modification in the modulation spectral domain .......... 41
  3.6 Envelope detectors .................................................... 49
  3.7 Sampled transforms ................................................... 50
  3.8 Filterbank interpretation ............................................. 53
  3.9 Generalized modulation filtering framework ....................... 55
  3.10 Signal model .......................................................... 59
  3.11 Summary ............................................................... 60
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>2</td>
</tr>
<tr>
<td>3.1</td>
<td>37</td>
</tr>
<tr>
<td>3.2</td>
<td>42</td>
</tr>
<tr>
<td>3.3</td>
<td>46</td>
</tr>
<tr>
<td>3.4</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>65</td>
</tr>
<tr>
<td>4.2</td>
<td>67</td>
</tr>
<tr>
<td>4.3</td>
<td>68</td>
</tr>
<tr>
<td>4.4</td>
<td>81</td>
</tr>
<tr>
<td>4.5</td>
<td>82</td>
</tr>
<tr>
<td>4.6</td>
<td>85</td>
</tr>
<tr>
<td>4.7</td>
<td>89</td>
</tr>
<tr>
<td>4.8</td>
<td>92</td>
</tr>
<tr>
<td>4.9</td>
<td>93</td>
</tr>
<tr>
<td>5.1</td>
<td>99</td>
</tr>
<tr>
<td>5.2</td>
<td>100</td>
</tr>
</tbody>
</table>
5.3 Spectrograms of two talkers before and after manual talker separation. . . . . 101
5.4 Estimated speech reception thresholds for the hearing impaired subjects of
the listening test on optimal modulation filtered speech. . . . . . . . . . . . . . 106
5.5 Estimated speech reception thresholds for the normal hearing subjects of the
listening test on optimal modulation filtered speech. . . . . . . . . . . . . . . 107
5.6 A speech signal, its spectrogram and its fundamental frequency. . . . . . . 110
5.7 Spectrograms of speech before and after processing with the incoherent fre-
quency weighting algorithm. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
5.8 Components of the target fundamental frequency detector. . . . . . . . . . . 120
5.9 Estimated speech reception thresholds for the hearing impaired subjects of
the listening test on coherent talker enhanced speech. . . . . . . . . . . . . . . 130
5.10 Estimated speech reception thresholds for the normal hearing subjects of the
listening test on coherent talker enhanced speech. . . . . . . . . . . . . . . . . 131
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Pass-band and stop-band frequencies of the modulation filters used in the</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>analysis of signal reconstruction after modulation filtering.</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Table of means for the hearing impaired subjects of the listening test on</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>optimal modulation filtered speech.</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Repeated measures ANOVA for the hearing impaired subjects of the listening</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>test on optimal modulation filtered speech.</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Table of means for the normal hearing subjects of the listening test on</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>optimal modulation filtered speech.</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Repeated measures ANOVA for the normal hearing subjects of the listening</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>test on optimal modulation filtered speech.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Parameter values of the coherent modulation filtering algorithm as used in</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>the subjective listening test.</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Table of means for the hearing impaired subjects of the listening test on</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>coherent talker enhanced speech.</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Repeated measures ANOVA for the hearing impaired subjects of the listening</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>test on coherent talker enhanced speech.</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>Table of means for the normal hearing subjects of the listening test on co-</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>herent talker enhanced speech.</td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td>Repeated measures ANOVA for the normal hearing subjects of the listening</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>test on coherent talker enhanced speech.</td>
<td></td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I sincerely thank my advisor, Les Atlas, for his generous time, his commitment and his mentorship. His persistence was an inspiration, and his knowledge and resourcefulness were invaluable. I am also very grateful for having an exceptional doctoral committee and thank Jeff Bilmes, Rico Malvar, Marina Meilà, and Jay Rubinstein for their effort, their support and their encouragement. I thank Pamela Souza and Jay Rubinstein for the inspiring collaboration and for allowing me to use their facilities for the listening tests. I thank Kelly Fitz for the fruitful collaboration on time-frequency reassignment, and Kaibao Nie for the collaboration on talker separation. I also thank the volunteers and those who helped me on the subjective listening tests, in particular Jong Ho Won and Colin Wong, for recruiting the subjects and conducting the tests.

I extend many thanks to my colleagues and friends: to everyone in and around the interactive systems design lab, Pascal Clark, Jeff Cole, Cameron Colpitts, Jon Cutter, Eric Garcia, Nate Jacobson, Brian King, Qin Li, Dan Luong, Pat McVittie, Chris Parrish, Scott Philips, Andrew Poole, Eric Rombokas, Brandon Smith, Jeff Thompson, Rahul Vanam, Jue Wang, and Travis Wilkins for their friendship and the many fun and insightful discussions; to John Keane and Somsak Sukittanan for showing me the way; to my dear friends Anton Dochtermann, Manu Forero, Megan McCloskey, and Elisabetta Valenti, for being my family away from home; and of course to everybody from the Pacifico and Seattle Revolution soccer teams, for the necessary distraction and the good times! Finally, I thank my family for their encouragement and support. This thesis would not have been completed without it. I’m especially grateful to my partner, Marloes Maathuis, for her love, patience and understanding, and for helping me keep my life in proper perspective and balance.

This research was funded by the UW Technology Gap Innovation Fund, the Washington Research Foundation, and the UW Virginia Merrill Bloedel Hearing Research Center.
Chapter 1

INTRODUCTION

1.1 Modulation analysis and filtering

Natural signals such as speech and music can be faithfully represented by low frequency modulators which modulate higher frequency carriers. This concept is illustrated by the spectrogram of a speech signal in figure 1.1. The spectrogram in figure 1.1 represents the distribution of the signal’s energy over time and frequency on a colorscale that varies from blue for low energy through green and yellow for medium energy to red for high energy. As the spectrogram shows, most of the signal’s energy is contained in nearly horizontal lines that are equidistant in frequency. Each horizontal line roughly corresponds to a harmonic of the speech signal. Individual signals for the first four harmonics for a short segment of time (highlighted by a magenta box on the spectrogram) are shown in the middle panels of figure 1.1. These panels demonstrate that each harmonic can be regarded as the multiplication of two processes: a high-frequency carrier process that captures the temporal fine-structure of the harmonic signal, and a low-frequency envelope process that follows the slowly-varying amplitude of the harmonic signal. The decomposition of each harmonic signal into its carrier signal and its envelope or modulator signal is shown in the panels on the right in figure 1.1.

Several studies have shown that the modulators of a speech signal are most important for speech reception. For example, when the modulators of a speech signal are replaced by envelopes of constant amplitude, speech becomes unintelligible. However, speech is highly intelligible when the modulators are preserved but the carriers are replaced by white noise. Because of their importance for speech reception, modulators of speech signals have been an interesting object of study for many researchers.
Figure 1.1: Spectrogram of a speech signal, with the first four harmonics separated into carriers and modulators. (left panel) Spectrogram of a male talker saying “seven”. Distribution of the signal’s energy over time and frequency is represented on a colorscale that varies from blue for low energy through green and yellow for medium energy to red for high energy. Each red, horizontal line corresponds to a modulated carrier whose frequency and amplitude are slowly varying over time. (center panels) Individual signals of the first four harmonics of the speech signal (indicated by the magenta box in the spectrogram). (right panels) Decomposition of the individual signals of the first four harmonics into modulators (blue) and carriers (red).
A commonly used method to analyze or modify the modulators of a broadband signal is to separate the signal into narrowband frequency subbands, and to decompose each subband into a carrier and a modulator, similar to the separation and decomposition in figure 1.1. Several such methods are described in the literature, but the details of how a signal is separated into subbands, and how a subband is decomposed into a carrier and modulator vary from author to author. Moreover, a formal study of these modulation analysis and filtering systems from a signal processing perspective has never been done.

In this work, we bring the existing techniques together in a unified signal processing framework to fill this gap in the theory of modulation analysis and filtering. We define the notion of a modulation transform, and discuss the conditions under which the inverse modulation transform exists, such that the modulation transform can be used for artifact-free modulation filtering. We extend the idea of a modulation transform to a generalized modulation filtering framework that is based on an arbitrary filterbank. Both the modulation transform and the modulation filtering framework transform signals to the modulation frequency domain. This signal domain gives rise to several new signal representations, which we define and present using a consistent naming scheme.

Furthermore, we analyze the existing methods to decompose a subband into a carrier and a modulator. We show the weaknesses of the commonly used magnitude envelope detector, which we refer to as the “incoherent” approach, and we argue for a “coherent” approach to the decomposition of a subband into a carrier and a modulator that is based on a carrier estimator. We define three carrier estimators that better satisfy the physical restrictions we impose on carrier signals and modulator signals. We evaluate the incoherent and coherent modulation filtering approaches on their effective modulation frequency responses, and on the signal-to-error ratio that they achieve after reconstruction with one of four candidate signal reconstruction techniques.

1.2 Application to hearing devices

To demonstrate the relevance of modulation frequency signal processing, we apply our theory of modulation analysis and filtering to the problem of talker enhancement for hearing devices. The objective is to detect and enhance the voice of a target talker in the presence
of other interfering talkers in a single channel recording using modulation analysis and filtering. The talker enhancement problem is motivated by the common complaint among users of hearing devices (such as hearing aids and cochlear implants) that they are unable to focus on a target talker in the presence of interfering talkers. This ability, colloquially referred to as the “cocktail party effect”, comes natural to normal hearing listeners, but is not restored for the hearing impaired by current technology hearing devices.

Our reasons to apply modulation analysis and filtering to the target talker enhancement problem are threefold. First, it is generally believed that the human auditory system performs modulation frequency analysis when it processes sounds, and that the neural pulses that it sends to the brain encode information about the subband modulators. Since the (healthy) auditory system is able to perform talker enhancement in the modulation domain, it seemed logical to us to apply modulation analysis and filtering to this problem. Second, in a preliminary modulation analysis and filtering experiment, we found that the energy from two co-channel talkers was largely non-overlapping in the modulation frequency domain. We demonstrate in the experiment, which is described in chapter 5, that it is possible to separate the two talkers with a supervised masking operation in the modulation frequency domain. This indicated to us that the modulation frequency domain as we define it has potential for talker enhancement and talker separation. We therefore deemed it worthwhile to pursue research into automatic target talker enhancement techniques based on modulation analysis and filtering. Third, modulation analysis and filtering is a novel approach to the target talker enhancement problem. Hence, it adds a new dimension to a problem that has been extensively studied by many researchers. Moreover, the modulation based approach is complementary to existing approaches to talker enhancement, and does not exclude the use of other techniques to achieve additional talker enhancement.

In this work, we motivate the usefulness of modulation analysis and filtering for the problems of target talker enhancement and talker separation with two preliminary experiments. We propose an incoherent modulation analysis approach to the talker enhancement problem that transfers the experience from the preliminary supervised modulation masking experiment into an automated technique. We present some talker enhancement results based on this technique, and discuss a fundamental limitation on incoherent modulation
frequency analysis of time-varying signals that prevents further significant improvement of this technique.

We therefore propose a novel coherent modulation analysis and filtering approach to target enhancement, which overcomes this limitation. The coherent approach is based on a custom implementation of a carrier estimator and a modulation filter. The carrier estimator is robust to noise, in particular to the presence of interfering talkers in the signal. The modulation filter is based on a novel multiresonator filterbank that is able to perform high-quality modulation lowpass filtering with high precision in time. We evaluate the novel coherent modulation filtering approach to target talker enhancement in a subjective listening test.

1.3 Organization

The structure of this dissertation is as follows. In chapter 2, we present the literature background relevant to this dissertation. We address topics related to modulation analysis and filtering from the psychoacoustics and signal processing research areas. We also address topics related to target talker enhancement, which we approach from the more general problem of blind source separation.

In chapter 3, we present our theory of modulation analysis and filtering systems. We introduce notation of existing frequency and time-frequency transforms, and develop the notion of a modulation transform that is based on these transforms. We discuss inverse modulation transforms, and conditions on signal modifications in the modulation frequency domain for proper invertibility of the transform. We consider sampled versions of the modulation transforms and show how they lead to a more general filterbank interpretation of modulation transforms.

In chapter 4, we deal with practical issues of modulation analysis and filtering. We describe the existing incoherent envelope detectors, and give examples of modulation frequency analysis based on these detectors. We demonstrate the limitations of the incoherent detectors, most notably that their carriers and modulators exceed the bandwidth of the subband from which they are derived. We propose three coherent carrier estimators that satisfy the bandwidth constraint of modulators and carriers, and evaluate and compare the
incoherent and coherent modulation filtering approaches on several criteria.

In chapter 5, we apply the modulation filtering framework to the problem of co-channel talker separation and target talker enhancement. We describe the preliminary experiments that motivated us to apply the modulation analysis and filtering technique to the problem of talker enhancement. We then propose an incoherent talker enhancement technique, which is limited in its performance by a fundamental limitation of incoherent modulation analysis of time-varying signals, and a novel coherent talker enhancement approach which overcomes this limitation. We describe the novel components of the coherent approach in detail, and present the results of a listening test with normal hearing and hearing impaired subjects that we conducted to evaluate the performance of the coherent technique.

We conclude the dissertation with an overview of our contributions and with suggestions for future research in chapter 6.
Chapter 2

BACKGROUND

2.1 Introduction

In this chapter we provide an overview of previous research in the areas relevant to this dissertation. In section 2.2 we review the history of the concept of modulation as it emerged in psychoacoustic research. In section 2.3 we discuss the discovery of modulation analysis and filtering in speech and signal processing, and their uses in a variety of applications. In section 2.4 we focus on target talker enhancement. We review a number of significant contributions to the more general research area of blind source separation, and discuss the applicability of this research to our work.

2.2 The modulation concept in psychoacoustics

In the early days of psychoacoustics, researchers studied the human auditory system using basic stimuli such as clicks and sinusoids. As understanding of the auditory system increased, the stimuli that were used to study its behavior became more complex. Zwicker [163] is one of the first to systematically study the auditory system using modulated stimuli. He measures the sensitivity of the auditory system to the relative phase of neighboring tones using amplitude-modulated and frequency-modulated stimuli with identical amplitude spectra. He determines that, below a certain modulation frequency, the ear is able to distinguish between the two kinds of modulation, and perceives amplitude modulation much better. He refers to this modulation frequency as the phase-limit frequency. For low frequency tones, the phase-limit frequency is around 30 Hz, and increases to 1 kHz for tones around 10 kHz.

In a set of publications, Viemeister [150–152] uses a similar systems analysis approach to measure the temporal modulation transfer function (TMTF) of the auditory system. The TMTF is an empirical function that quantifies the modulation threshold at which
an observer is able to detect sinusoidal amplitude modulation as a function of the mod-
ulation frequency. In one of his experiments, Viemeister [152] measures the TMTF using
sinusoidally amplitude-modulated wideband noise. The use of wideband noise eliminates
all spectral cues for the detection of modulation. The TMTFs he obtained have a lowpass
characteristic; they have a modulation threshold that is constant for modulation frequencies
up to approximately 10 Hz, and then falls off at a rate of 3–4 dB/octave up to a modulation
frequency of 800 Hz. The TMTFs are largely independent of carrier level. Viemeister also
measures the TMTF for different carrier frequency regions using bandlimited sinusoidally
amplitude-modulation wideband noise. These TMTFs have the same lowpass characteristic
as the wideband TMTFs, but the cutoff modulation frequency increases with increasing
carrier frequency. Viemeister concludes that this reflects an improvement in temporal res-
olution of the auditory system with increasing spectral frequency.

Increasing sophistication of the modulated stimuli that were used in psychoacoustic
studies led to the discovery of two important phenomena of modulation detection processes
in the auditory system: comodulation masking release (CMR) and modulation detection
interference (MDI, referred to by some authors as modulation discrimination interference).

The principle of comodulation masking release, first described by Hall, Haggard and
Fernandes [43], is the effect in the auditory system that a tone, masked by bandlimited
noise, is easier to detect when the noise is amplitude-modulated than when the noise is not
amplitude-modulated. Hall et al. found that this masking release of amplitude-modulated
noise only occurs when the noise extends over more than one critical band in the auditory
system. They conclude that temporal envelope coherence between critical bands helps to
differentiate a signal from noise. In a supplemental paper, Hall et al. [42] report that
they also find a (smaller) release from masking when envelope coherent noise is presented
contralateral (i.e., to the other ear) from the tone plus noise masker stimulus. They conclude
that CMR seems to occur at more than one stage in the auditory system.

Other effects of tone and noise parameters on CMR are extensively studied by several
authors. For example, McFadden et al. [89] investigate the influence of the level of the
envelope coherent noise relative to the masker noise, the time delay between the noises, and
the duration of the tone to be detected. The effect of other CMR parameters is analyzed by
Schooneveldt and Moore [125, 126], Moore et al. [94] and Verhey et al. [149], among others.

The second across-channel modulation detection process of the auditory system is modulation detection interference. Experiments by Yost and Sheft [160, 161] and Yost et al. [162] demonstrated that it is harder to detect amplitude-modulation on a probe tone, in the presence of a masker tone in a different spectral region, when the masker tone has amplitude modulation similar to the probe tone, than when the masker tone has no amplitude modulation. They attribute this phenomenon to \textit{perceptual grouping} of sounds with common amplitude modulation in the auditory system. Perceptual grouping of sounds would make it more difficult to observe the modulation of the individual sounds.

Moore and Shailer [93] further examined the role of perceptual grouping on MDI. Some aspects of their results were consistent with the idea that perceptual grouping causes MDI. However, their results also showed that MDI takes place even when the modulation frequencies of probe and masker are very different. Since perceptual grouping of components with very different modulation is not known to occur, they suggest that perceptual grouping is not the only factor in MDI.

Given the measurements of modulation masking and modulation detection in the literature, Dau, Kollmeier and Kohlraush [20, 21] proposed a quantitative model of auditory processing of amplitude modulation that describes the experimental data. Their model is essentially the earlier model of the “effective” signal processing in the auditory system by Dau, Püschel and Kohlrausch [18, 19], with the addition of a modulation filterbank.

The basic model of Dau et al. [18, 19] consists of a preprocessing stage, an adaptation stage, internal noise and an optimal detector. In the preprocessing stage, input stimuli undergo basilar-membrane filtering, half-wave rectification and low-pass filtering. The adaptation stage compresses the preprocessed signals. Compression is linear for rapidly varying inputs, and almost logarithmical for stationary signals. This stage models adaptive properties of the auditory periphery such as temporal masking. Internal noise is added to the output of the adaptation stage to account for the fact that the auditory system can not detect every change in the input signal. The resulting signal is referred to as the internal representation of the input stimulus. Finally, the optimal detector simulates the ability of a human observer to discriminate between two auditory stimuli by analyzing the internal
representations of the stimuli.

In the extended model of Dau et al. [20, 21], a modulation filterbank is applied to the signal after the adaption stage and before internal noise is added to the signal. Their rationale for the modulation filterbank is that experimental data on modulation frequency specificity suggests that the auditory system has not a single modulation frequency band but multiple bands that are broadly tuned for modulation frequency. Dau et al. postulate that there are two regimes in the modulation filterbank. They assume that modulation filters between 0–10 Hz are uniformly spaced and have a constant bandwidth of 5 Hz, whereas filters between 10–1000 Hz are constant-Q with a Q value of 2. They find that the extended model, by using a modulation filterbank, better predicts experimental data on modulation masking than the basic model.

The concept of a modulation filterbank in the auditory system has received considerable attention in recent years, see for example [9, 22, 24, 32, 51, 92, 95, 137]. However, it remains a controversial topic that is the subject of ongoing research (e.g., [130, 131]).

Perhaps the single most important contribution to the field of psychoacoustics for our research on modulation filtering is a paper by Drullman, Festen, and Plomp [27]. In their paper, Drullman et al. measure the effect of temporal envelope smearing on the intelligibility of speech. To achieve envelope smearing, they pass speech signals through a filterbank, compute the envelope of each subband, and apply a lowpass filter to the envelopes. They recombine the filtered envelopes with the original fine structure of the subbands, and reconstruct a broadband signal by summing the modulation filtered subbands. In their experiments, they vary the bandwidth of the filterbank’s frequency bands (\( \frac{1}{4}, \frac{1}{2} \) or 1 octave), and the cutoff frequency of the lowpass filter (0, \( \frac{1}{2}, 1, 2, 4, 8, 16, 32, \) or 64 Hz). They find that modulation frequencies between 4 and 16 Hz are most important for speech intelligibility. Speech reception is severely reduced for lowpass cutoff frequencies of 0–2 Hz, and only marginally affected by cutoff frequencies above 16 Hz. Intelligibility seems to be independent of the bandwidth of the filterbank frequency bands for cutoff frequencies above 4 Hz. Below 4 Hz, a larger bandwidth improves intelligibility. In a follow-up paper, Drullman et al. [26] repeated their experiments for highpass modulation filtered speech. These experiments are consistent with their original conclusions: speech intelligibility is not affected by
a 4 Hz highpass modulation filter, but clearly reduced with a narrowband filterbank and highpass cutoff frequencies above 64 Hz.

Drullman et al. were particularly influenced by the work of Houtgast and Steeneken on the concept of the modulation transfer function (MTF) in room acoustics. Houtgast and Steeneken [52–54], Steeneken and Houtgast [135, 136] and Houtgast, Steeneken and Plomp [55] were the first to measure the intelligibility of speech in rooms using an amplitude-modulated test signal. They recognized the importance of modulation frequencies to the intelligibility of speech, and designed a method that could objectively evaluate the intelligibility of speech in rooms and auditoria. In their method, they excite a room with the amplitude-modulated test signal and measure the envelope spectrum of several frequency subbands at a target location in the room. The ratio of the test signal’s envelope spectrum and the measured envelope spectrum yields the modulation transfer function of the room. They derive a single speech transmission index (STI) number for the room by perceptually weighting the modulation transfer functions of the frequency subbands. The idea of a modulation transfer function was put in a formal system analysis framework by Schroeder [129].

The papers by Drullman et al. [26, 27] are significant for modulation filtering research because they were among the first to apply modulation filtering to speech. They clearly documented their signal processing methods, establishing a modulation filtering procedure that has subsequently been copied by many researchers. Many authors also refer to the conclusions of Drullman et al. on the relevance of specific modulation frequencies for the intelligibility of speech. Either they base their own experiments on these conclusions, or their experiments confirm Drullman’s findings.

Drullman’s modulation filtering method, however, has been questioned by Ghitza [35]. Ghitza noticed that a modulation filtered subband, when passed through the filterbank for a second time, has essentially the same envelope as the original, unmodified subband. This suggests that Drullman’s processing technique creates modulation filtered signals that still contain rich envelope information much beyond the cutoff frequency of the lowpass modulation filter. Ghitza attributes this to the dependence between the temporal envelope and the subband fine structure as they are computed by Drullman et al. Both the envelope and the fine structure of a narrowband subband are wideband signals. Reconstructing the
narrowband subband from the wideband envelope and carrier depends heavily on the can-
celation of high-frequency terms in the envelope and fine structure. When the temporal
envelope is filtered, the high-frequency terms of the fine structure are no longer canceled
by the temporal. The presence of those high-frequency terms in the reconstructed subband
usually restores the original subband envelope to a great extent, and greatly reduces the
effectiveness of the filter that was applied to the envelope. Ghitza’s observation does not
necessarily invalidate Drullman’s work, but it places doubt on their method and conclu-
sions. We share Ghitza’s view that Drullman’s modulation filtering scheme produces rich
envelopes. In our work and in this dissertation, we argue for a different kind of modulation
filtering that gives better control over the envelope bandwidth in the output signal.

Psychoacoustics is a rich field with contributions from many researchers. We have been
necessarily incomplete in our brief overview of the modulation related research here, as we
have only mentioned the contributions that have influenced our research and thinking most
directly. The interested reader can find a complete chronological background of modulation
research in psychoacoustics in the extensive review by Kay [63], and a summary of the role
of modulation in hearing in the paper by Plomp [104].

2.3 Modulation analysis and filtering in signal processing

The modulation concept in speech engineering can be traced back to Dudley’s work on
speech coding in the 1930’s and 1940’s [28, 29]. In [28], Dudley describes the vocoder,
an electrical speech synthesizer. The vocoder analyzes a speech signal and reproduces
it by selectively modifying a buzz-tone and a hiss-tone. The operation of the vocoder
is based on the fundamental assumption that speech can be decomposed into “carriers”
and “signals”. According to Dudley, the “carriers” are complex multi-frequency carriers
that are caused by the “(vocal) cord tone” and the “breath tone”. The “signals” are
the slowly-varying articulatory motions that define speech. These speech-defining signals
contain all the speech information although they themselves vary at slow, inaudible rates.
They are impressed on the voiced carrier by both frequency modulation (changes in pitch)
and amplitude modulation, and on the unvoiced carrier by amplitude modulation, and thus
become audible. In [29], Dudley elaborates on the idea of carriers and signals, specifying
more clearly what constitutes the carrier elements of speech, the voice carrier, the speech message and the voice modulators. He demonstrates the validity of his theory with the voder, a device that allows a human operator to manually modulate a buzz-tone and a hiss-tone to generate intelligible speech.

The ideas of Dudley’s voder and vocoder are advanced in the phase vocoder by Flanagan and Golden [33]. They capitalize on a decomposition of the short-time Fourier spectrum into its magnitude and phase components to achieve a reduction in transmission bandwidth and a means of time compression and expansion of speech signals. Their experience with channel vocoders shows that the short-time Fourier magnitude of speech signals may be bandlimited to 20 to 30 Hz without substantial loss of perceptual quality. The bandwidth of the subband phase signal, however, is generally not bounded, and a similar bandlimit can not be imposed on it. Flanagan and Golden speculate that the first-order time-derivative of subband phase signals is better behaved and may be bandlimited for transmission. They propose a speech transmission system where the receiver reconstructs a speech signal by summing the output of a number of oscillators modulated by the slowly-varying magnitude and phase signals that are transmitted by the sender.

Besides speech coding, modulation analysis has been successfully used in several other applications. Examples of these applications are audio coding [142, 154], audio fingerprinting [141], signal identification and classification [139, 140], speaker recognition [65], and speech detection [90]. Of particular note is the relative spectral transform - perceptual linear prediction (RASTA-PLP) speech analysis technique by Hermansky, Morgan, Bayya and Kohn [48, 49]. RASTA-PLP was one of the first signal analysis techniques to use subband spectral analysis similar to modulation analysis. It was designed as a front-end to speech recognizers to make them more robust to changes in the acoustic environment. RASTA-PLP computes the critical band power spectrum (as in the PLP technique [46]), and takes the logarithm of the spectrum. It applies a band-pass filter to the log-spectral subbands, removing slowly-varying components due to the environment and rapidly-varying components due to noise. It subsequently applies perceptual corrections to the filtered log-spectrum, such as adding the equal loudness curve and scaling the spectrum to simulate the power law of hearing. Finally, it models the resulting spectrum by an all-pole model. Hermansky et
[48, 49] report significantly reduced speech recognition error rates on small vocabulary isolated telephone quality speech and large vocabulary continuous high quality speech for RASTA-PLP in comparison with PLP, when the speech recognizer is tested in a different acoustic environments than it was trained in.

In the direct footsteps of RASTA-PLP, several researchers have also used similar modulation spectral features for automatic speech recognition, for example [61, 64]. More recently, Tyagi [145] computed RASTA-PLP speech recognition features in the cepstral domain, which resulted in lower word recognition error rates on noisy signals.

In [47], Hermansky and Morgan present a variation to the original RASTA-PLP technique, which replaces the logarithm applied to the spectral magnitude by a different compressive nonlinearity. They also extend RASTA-PLP to the *relative spectral transform* (RASTA) signal processing technique, which allows reconstruction of power spectral filtered subbands to time-domain waveforms. RASTA computes the spectrum and takes the cube-root of the magnitude of the spectrum. It band-pass filters the cubic root of power spectral magnitude, which is subsequently half-wave rectified, exponentiated to the third power and recombined with the original spectral phase. RASTA then reconstructs a time-domain signal using the standard overlap-add technique [3]. Hermansky and Morgan [47] used RASTA processing for enhancement of speech corrupted by car noise and artificially introduced impulses. In informal listening, they found that RASTA processing significantly reduced the steady background noise as well as the impulse noise, at the cost of some colored musical noise and phase artifacts. Speech intelligibility did not seem to improve, but the processed speech appeared to stand out more above the background.

Hermansky, Wan and Avendaño [50] proposed an interesting application of RASTA. They replaced RASTA’s fixed band-pass filter with a bank of Wiener-like filters. Each filter was designed to spectrally map a subband of noisy speech to the corresponding subband of clean speech in a least squared-error sense. Initial filters for the clean speech subband spectra were estimated on approximately 2 minutes of speech recorded in quiet. Hermansky *et al.* [50] tested the RASTA processing using the Wiener-like filters on noisy cellular communication. Informally they observed a noticeable improvement in subjective quality of the recordings. The processing also did not seem to impair the quality of clean speech.
Two other noteworthy modulation analysis studies are Greenberg and Kingsbury [38] and Greenberg and Arai [37]. Greenberg and Kingsbury [38] define the modulation spectrogram as an invariant representation of speech. Their definition of the modulation spectrogram involves a critical-band filterbank, followed by a subband envelope detector defined by half-wave rectification and low-pass filtering of the subband signals. Their modulation spectral representation is the energy of the subband envelope over the 0-8 Hz modulation frequency range, plotted as a two-dimensional function of time and subband center frequency on a 30 dB range.

Greenberg and Arai [37] establish a relationship between the complex modulation spectrum and speech intelligibility. They define a complex modulation spectral representation of subband envelope spectra that contains both an amplitude and a phase component. Given a speech signal, they time-reverse segments of a specific duration (in the range of 20 ms to 180 ms), and measure the intelligibility of the time-reversed speech in a subjective listening experiment. In addition, they compute the complex modulation spectrum for the time-reversed speech signals. They find that speech intelligibility has no strong correlation with the amplitude of the modulation spectrum, but a high degree of correspondence with the complex modulation spectrum (i.e., both amplitude and phase).

The use of modulation filtering techniques has been studied in the context of signal dereverberation, audio watermarking, and source separation. Langhans and Strube [75] were possibly the first to develop a modulation filtering approach to dereverberation. Based on the relation between the modulation transfer function of a room and speech intelligibility ([52–54]), and on a suggestion by Schroeder [128], they attempted to enhance speech corrupted by reverberation noise by pre- or postprocessing the signal with a suitably chosen modulation filter. They report no intelligibility improvement when filtering the linear subband envelope, but a raise in intelligibility when filtering the logarithmic subband envelope. Similarly, Kitamura et al. [66] and Kusumoto et al. [74] pre-process speech signals with data-derived or empirically-derived modulation filters to emphasize specific modulation frequencies prior to transmission in a reverberant environment. The data-derived modulation filters are defined as the ratio of the modulation frequency response of clean and reverberant speech, averaged over a large number of speech signals. The empirically-derived filters are
designed to emphasize modulation frequencies between 4 and 16 Hz, which are known to be
important for speech intelligibility. Kitamura et al. [66] report an improvement in speech
intelligibility after pre-processing, both for normal hearing and hearing impaired subjects.
Kusumoto et al. [74] tested only normal hearing subjects, and find an improvement in speech
intelligibility for consonant-vowel recognition tasks but not for word recognition tasks.

Modulation filtering has been successfully applied to audio watermarking by Dish, Herre
and Kammerl [25]. They embed a perceptually transparent watermark in an audio signal
by selectively reducing the energy in specific regions of the modulation spectrogram. By
cleverly organizing those regions in a checkerboard pattern, their method is resilient to a
wide range of signal manipulations, such as time shifting, MPEG compression and lowpass
filtering. The watermark is detected in the receiver by comparing the energy in the notched
areas to the energy in the surrounding areas.

Source separation and enhancement using modulation filtering is addressed in [4, 122,
musical instruments with differing modulation frequency characteristics through modula-
tion filtering. Schimmel, Atlas and Nie [123] demonstrate the feasibility of target talker
enhancement using modulation filtering. A novelty of these techniques is that they decom-
pose subbands into carriers and modulators using coherent carrier estimators, instead of
the magnitude/phase decomposition of subbands. Details of coherent carrier estimation are
given in sections 4.3.2 and 4.4, and details of the separation and enhancement techniques
are given in sections 5.3 and 5.4.

2.3.1 AM-FM decomposition

An important subproblem of modulation analysis and filtering is the decomposition of a
signal into its carriers and modulators. With the exception of [4, 122, 123], all of the above
modulation analysis and filtering approaches use a magnitude and phase decomposition to
separate frequency subbands into a carrier and a modulator. However, the magnitude/phase
decomposition is just one of many ways to decompose a narrowband signal into a carrier and
modulator [15]. We show in section 4.3 that the use of the magnitude/phase decomposition
is limited in practice, mostly because the magnitude envelope and phase carrier of a narrow-
band subband are generally not bandlimited. One of the contributions of our work is the
motivation and development of alternative modulator/carrier decompositions to the magni-
tude/phase decomposition that better satisfy the condition of constrained bandwidth that
we impose on modulators and carriers. Since existing research on AM-FM decompositions
of monocomponent signals (see [14] for a definition of monocomponent and multicomponent
signals) is relevant to our carrier/modulator decompositions, we review several contributions
here.

Kaiser [60] derives a simple algorithm to determine the instantaneous energy required
to generate a signal. He uses Newton’s law of motion to the motion of a mass suspended by
a spring to show that the energy associated with the oscillation of the mass is proportional
to the square of the amplitude and the square of the oscillation frequency. Based on this
observation, he defines the energy operator Ψ on a signal $s(t)$ by

$$\Psi(s) = (\dot{s})^2 - s\ddot{s},$$  \hspace{2cm} (2.1)

where $\dot{s}$ and $\ddot{s}$ denote the first and second derivative of the signal $s(t)$ with respect to time.
The energy operator $\Psi$ satisfies Newton’s law in that it is proportional to the square of the
amplitude and the square of the frequency of a sinusoid oscillating at constant amplitude
and frequency. Kaiser studies the properties of the energy operator for a few common signals
(a chirp, sums of sinusoids, etc.), and its robustness to noise. He notes that the operator is
sensitive to noise, but remarks that the bandwidth of the energy signal $E(n)$ is much lower
than the bandwidth of the signal, so that noise effects may be reduced by lowpass filtering
the energy signal.

Kaiser’s energy operator, which has become known as the Teager-Kaiser energy op-
erator, is not yet an AM-FM decomposition. Before long, however, Maragos, Kaiser and
Quatieri [85–87] introduced the energy separation algorithm (ESA), which was subsequently
refined by Bovik, Maragos and Quatieri [13]. The ESA is defined by

$$\hat{a}^2(t) = \Psi^2(s)/\Psi(\dot{s}),$$  \hspace{2cm} (2.2)
\[ \hat{\omega}^2(t) = \Psi^2(\dot{s})/\Psi(s). \]  

Bovik et al. consider the effects of noise on the ESA and demonstrate that significant noise renders the energy operator unpredictable and the ESA unreliable. However, they show that the ESA can be made highly resistant to noise by first filtering a signal through a constant-Q bank of bandpass filters. For optimal performance, the filters need to be sufficiently narrowband (increasing the signal to noise ratio in the analyzing band) and must densely sample the signal (ensuring a high signal response).

Vakman [147] approaches AM-FM decomposition from the perspective that AM and FM components of a signal can be derived from the sum of the real signal and its “conjugation”. A conjugation operator transforms the real signal into its conjugated signal. Vakman lists three physical conditions for the conjugation operator: (1) it must be continuous; (2) it must be homogeneous, i.e., the FM must be invariant to scaling; (3) it must have harmonic correspondence, i.e., a simple sinusoid must retain its constant amplitude and frequency, and only shift 90 degrees in phase. Vakman shows that only the Hilbert transform satisfies all three conditions. He compares AM-FM decomposition using the Hilbert transform to two other AM-FM decompositions, and to the Teager-Kaiser algorithm. He shows how the Teager-Kaiser algorithm violates the continuity condition. He concludes that the Teager-Kaiser algorithm gives reasonable AM-FM decomposition of most signals, but that the Hilbert method is more accurate for noisy signals and eliminates distortion in special cases.

Loughlin and Tacer [80] build on the work by Vakman. However, they propose four slightly different physical constraints that any AM-FM decomposition must satisfy: (1) if the signal is bounded in magnitude, then the magnitude of the AM must be bounded too; (2) if the signal is limited to a frequency region, then the FM must be limited to the same frequency region; (3) a simple sinusoid with constant amplitude and frequency must yield constant AM and FM; (4) if the input is scaled, the AM must scale by the same amount, and the FM must remain unchanged. They show that the analytic signal (i.e., the Hilbert method) does not satisfy conditions 1 and 2. Therefore, they propose a different AM-FM decomposition. They define the FM of a signal \( x(t) \) as the spectral center of gravity of any
positive time-frequency distribution $P(\omega|t)$ of the signal,

$$\omega(t) = \langle \omega \rangle_t = 2 \int_0^\infty \xi P(\xi|t) d\xi.$$  \hfill (2.4)

They obtain the AM of the signal from the FM through time-varying coherent demodulation of the signal $x(t)$ with the phase signal

$$\phi(t) = \int_{-\infty}^t \omega(\tau) d\tau,$$  \hfill (2.5)

as follows:

$$A(t) = \int x(\tau) e^{j\phi(\tau)} h_{lp}(t, \tau) d\tau,$$  \hfill (2.6)

where $h_{lp}(t, \tau)$ is a time-varying low-pass filter with cut-off frequency $\langle \omega \rangle_t$. They illustrate the AM-FM detector on a two-tone signal and a speech signal, showing that the proposed AM is a good match to the signal’s envelope.

Quatieri, Hanna and O’Leary [106, 107] propose an AM-FM estimation method based on the principle of FM to AM transduction by linear filters. They find motivation for their method in the hypothesis that the auditory system possibly uses a similar transduction mechanism to detect frequency modulation [118]. They derive a closed-form solution to transduction-based AM-FM separation for filters with piecewise linear frequency response, and generalize this result to an expression for filters that are not piecewise-linear. They show that a closed-form solution to this expression exists for Gaussian filters, and provide an iterative procedure to find a solution to this expression for auditory-motivated gammatone filters.

AM-FM decompositions are implicitly studied in research on instantaneous frequency (IF) estimation. The concept of the instantaneous frequency of a monocomponent signal is usually approached from a more mathematical standpoint than the physical perspective of signals that AM-FM decompositions commonly take. In both subject areas, however, authors use similar time-varying amplitude and frequency models of signals, and define methods to estimate their time-varying frequency component. Without going into their detail, we found several publications on IF estimation worth exploring in the context of
modulation analysis and filtering [11, 12, 73, 84, 103, 108].

Implicit AM-FM decompositions can also be found in publications on sinusoidal parameter estimation. Articles on this topic explore the limits of estimating the amplitude, frequency and phase of a single tone or a set of tones in the presence of noise [41, 81, 109, 114]. They invariably derive the maximum-likelihood estimator of those parameters under the conditions of their problem statement, and compare its performance to the Cramér-Rao bound. The paper by Hainsworth and Macleod [41] includes the method of frequency reassignment in its comparison of estimator performance, which makes it especially relevant since we have adopted the frequency reassignment technique as a carrier estimator in our work (see section 4.4.4).

2.4 Blind source separation

In chapter 5 of this dissertation, we apply our research on modulation filtering to the problem of enhancing a target talker in the presence of interfering talkers. This problem is usually not discussed in isolation in the literature, but considered as part of the more general and more difficult problem of blind source separation (BSS). Blind source separation techniques attempt to achieve full separation of arbitrary sound sources in one or more observed mixtures without prior knowledge of the signals. BSS techniques can be divided in three broad categories: statistical learning approaches, spatial filtering approaches, and source modeling approaches. We discuss each category in the following three sections, with the emphasis on source modeling approaches because modulation filtering techniques belong to this category.

2.4.1 Statistical learning

Statistical learning approaches make no assumption about the sources other than statistical independence. They attempt to learn source characteristics either from the observed mixture, or from a priori data from each source. For example, Roweis [117] trains a hidden Markov model (HMM) to consecutive columns of narrowband spectrograms of clean speech data from a speaker. He combines two of those speaker dependent HMMs into a factorial HMM, and uses this to estimate the probability that a time-frequency point of an observed
mixture signal belongs to either speaker. He achieves separation of the sources by reconstructing a signal from the time-frequency points that most likely belong to the desired talker, setting all other time-frequency points to zero.

Bach and Jordan [8] also treat the speaker separation problem as a time-frequency segmentation problem. Instead of learning the characteristics of individual speakers, they build feature sets based on typical time-frequency speech cues such as common onset/offset, frequency comodulation, pitch and timbre. To segment the time-frequency domain of an observed mixture signal, they use a spectral segmenter defined by a parameterized similarity matrices, one for each feature in the feature set. Values for the similarity matrix parameters are learned from training data using a spectral learning algorithm.

Another approach to speech separation in the time-frequency domain is to represent the time-frequency domain of an observed mixture by a linear combination of basis vectors that represent each source. Two popular ways to estimate basis vectors for a source are vector quantization, e.g. [31], and non-negative matrix factorization, e.g. [124, 133]. In [31], Ellis and Weiss build a vector quantization codebook for short-time spectral frames of training data from a single source. They project an observed mixture signal onto the source’s codebook vectors, and reconstruct a separated signal from the best matching vectors. They augment their technique with an HMM to implement sequential constraints between codebook vectors. They test their method on speech corrupted by speech-shaped noise. They find the results somewhat disappointing but the method an interesting approach with greater potential.

In the non-negative matrix factorization technique (e.g. [124, 133]), the short-time Fourier transform magnitude of training data from a single source is approximated by the product of a non-negative matrix of spectral patterns and a non-negative matrix of corresponding temporal profiles. The same factorization technique is then applied to the short-time Fourier transform magnitude of an observed mixture, but it is fixed to use the spectral patterns from the desired source, plus sufficient “free” spectral patterns to represent the interference. Separation is achieved by synthesizing a short-time Fourier transform magnitude from spectral patterns and temporal profiles from the desired source only, and inverting the short-time Fourier transform to a time-domain signal (using original short-time
Fourier phase and overlap-add reconstruction).

Jang and Lee [57] and Jang, Lee, and Oh [58] propose to represent a source by basis vectors in the time-domain. They learn time-domain basis functions for each source using a method from independent component analysis (ICA). Given an observed mixture signal, they compute the most likely combination of basis functions and coefficients that explains the mixture signal. They achieve separation by reconstructing from the basis functions and coefficients from a single source.

2.4.2 Spatial filtering

Spatial filtering approaches to source separation assume that different sources must occupy different locations in space. They analyze an observed mixture of signals for spatial cues, and separate signals that appear to be emitted from different locations. It usually takes two or more simultaneously observed mixtures of the source signal to estimate spatial locations, which makes this type of technique less relevant for our purposes. We mention some of these methods here for their historical value for modulation research or the insight they offer about speech signals.

The spatial filtering system proposed by Kollmeier and Koch [70] uses intensity and phase differences in the complex modulation spectrograms of stereo recordings to separate sources. They suggest that a clustering algorithm could be used to cluster the complex modulation spectrogram into regions that likely originated from the same source. In lieu of such a complicated algorithm, however, they implement a simpler technique: regions of the two modulation spectrograms for which the intensity and phase differences are consistent with a desired source location are passed, and other regions are rejected. They test their processing algorithm in a subjective listening test, and find it improves speech intelligibility and reduces listening effort under most acoustical conditions. The improvement, however, is small when compared to other binaural techniques, or when compared to the binaural performance of the healthy auditory system.

A time-frequency version of spatial filtering is known under the acronym DUET. This technique was introduced by Jourjine, Rickard and Yilmaz [59], restructured for real-time
operation by Rickard, Balan and Rosca [113], and put into a rigorous mathematical framework by Yılmaz and Rickard [159]. The DUET technique works as follows: it estimates amplitude and delay differences between spectrograms of two observed mixtures. Given the amplitude-delay parameters, it maps each point in time-frequency space to a point in amplitude-delay space. It clusters the points in amplitude-delay space into sources, and segregates points in the spectrograms based on their cluster assignment in amplitude-delay space.

Rickard and associates motivate their spectrogram segregation approach to talker separation with the concept of *W-disjoint orthogonality*. Two signals are called disjoint orthogonal if their contributions to the spectrogram of their mix are orthogonal, i.e., their time-frequency energy distributions have no overlap. They show that mixtures of up to 4 speech sources are approximately W-disjoint orthogonal, and that individual sources can be recovered with perfect intelligibility from such mixtures using ideal binary masks [159].

Essentially the same time-frequency spatial filtering approach is taken by Roman, Wang and Brown [116], but they treat clustering of the amplitude-delay space as a classification problem.

### 2.4.3 Source modeling

Source modeling approaches to source separation use a parametric model for each source. They estimate model parameters for each source from the observed mixture. They create separated signals either by generating a signal from the model of each source, or by filtering the mixture signal with a filter that is controlled by the values of a source’s model parameters.

An early example of a source modeling approach is the harmonic selection technique by Parsons [102]. He assumes an harmonic model of voiced speech, and implements a pitch estimation and tracking algorithm based on spectral peak picking that can handle two concurrent talkers. Parsons separates the two talkers by suppressing the harmonics from the interfering talker in the spectral domain and reconstructing a time-domain signal from the desired talker’s harmonics only. He claims the result sounds good, although no
formal listening test is performed. Parsons concludes with several suggestions to improve his algorithm.

The harmonic magnitude suppression (HMS) technique by Hanson and Wong [44] employs a similar strategy. It estimates the parameters of the harmonics of the interfering talker, which Hanson and Wong assume is the stronger signal in practical situations. It then subtracts the interfering harmonics from the signal to recover the target talker. Morgan et al. [97] improve on this technique by using a maximum-likelihood pitch estimator [156] and adding a spectral enhancement stage that further enhances the harmonics and formants of the target talker. Additionally, they add voiced/unvoiced detection and speaker assignment logic to fully automate the talker separation process.

Stubbs and Summerfield [138] evaluate two separation techniques that are also based on a harmonic model of speech: cepstral filtering and a hybrid technique that combines cepstral filtering with harmonic selection [102]. The rationale behind cepstral filtering is that the cepstrum of a mixture of two speech signals has a peak at the pitch of either talker. Removing one of those peaks separates one talker from the other in the mixture. In the hybrid technique, harmonic selection is augmented with cepstral filtering to better handle cases when the target talker’s speech is unvoiced but the interfering speech is voiced. Stubbs and Summerfield evaluate their separation algorithms in perceptual experiments on normal hearing and hearing impaired subjects. When tested with speech signals with monotone pitch, both methods worked well for both groups of subjects. With normally intoned speech signals, only the cepstral filtering approach gave improvement, but only for the normal hearing subjects.

A major problem for all of these methods, according to Hu and Wang [56], is their inability to adequately deal with the high-frequency part of speech. Hu and Wang distinguish between resolved harmonics and unresolved harmonics. A harmonic is called resolved if it is the dominant component in a filterbank channel; otherwise, it is called unresolved. For the auditory filterbank they use, low-frequency harmonics are resolved, whereas high-frequency harmonics are unresolved. They estimate parameters for resolved harmonics by AM-FM decomposition of the corresponding filterbank channel, and evaluate the amplitude modulation of unresolved harmonics to estimate their parameters. They assign the estimated
harmonics to different speakers based on temporal continuity, cross-channel correlation, and common amplitude modulation.

Some talker separation methods claim to model the source with a more general sinusoidal model, see for example [105] and [155]. However, these methods often add a harmonic constraint to the sinusoidal model to boost parameter estimation performance, which reduces the model to a harmonic model. The method of reconstruction of separated sources, on the other hand, is generally different for sinusoidal modeling techniques; they usually drive a bank of time-varying oscillators with the amplitude, frequency and phase parameters of harmonics from a single source to generate a separated signal.

Unfortunately, most of the source modeling techniques have limited practical value for our application, because they are often designed to handle a maximum of two simultaneous talkers. Nonetheless, the ideas that they are based on have provided useful insights about pitch estimation and talker enhancement for our research.

2.5 Summary

In this chapter we have given a survey of the literate in psychoacoustics and signal processing that is relevant to our research into modulation transforms and target talker enhancement. We discussed the modulation detection phenomena of comodulation masking release and modulation detection interference, and the possible existence of a modulation filterbank in the auditory system. Two key results in psychoacoustics related to modulation processing are the importance of modulation frequencies between 4 and 16 Hz for speech intelligibility, and the introduction of a modulation filtering technique by Drullman et al. [26, 27]. We have also provided a short historical context to modulation analysis and filtering in the field of signal processing, and discussed the most relevant theories (the (phase) vocoder, RASTA processing, the modulation spectrogram and the coherent modulation spectrogram) as well as applications of modulation analysis and filtering in audio coding, audio fingerprinting, audio watermarking, signal classification, and more. In the final part of this chapter, we reviewed some of the approaches to blind source separation, a research area that encompasses target talker enhancement. We discussed three major categories of blind source separation techniques, namely statistical learning, spatial filtering, and source modeling.
Chapter 3

THEORY

3.1 Introduction

In this chapter we identify modulation frequency analysis and filtering systems, or modulation systems for short, in an abstract way. We organize the definitions and implementations of the modulation systems that were discussed in the previous chapter, together with our own research into these systems, into a cohesive modulation analysis and filtering framework. This framework gives us a compact description of modulation systems and allows us to establish the notation and terminology necessary to discuss these systems effectively.

Furthermore, we show that modulation systems are a family of algorithms that is parameterized by a filterbank and an envelope detector. We review some important aspects of the filterbank and the envelope detector, and define desirable characteristics of modulation systems and discuss their implications. Most of the practical considerations and implementation details of modulation systems are left for the next chapter.

3.2 Modulation transforms

Our theory of modulation frequency analysis and filtering is best explained through the definition of modulation transforms. Modulation transforms are signal transformations that are based on the Fourier transform and the short-time Fourier transform. Therefore, we first establish the definition and notation of these fundamental transforms before turning to the definition of modulation transforms.

3.2.1 Existing frequency and time-frequency transforms

The Fourier transform and the short-time Fourier transform are both well understood and extensively described in the literature. However, notation may vary from author to author. We therefore give definitions of the continuous-time and discrete-time versions of these
transforms in the notation that will be used throughout this dissertation.

**Definition 3.2.1.** The *continuous-time Fourier transform* of a signal \( x(t) \) is defined by

\[
X(\Omega) = \mathcal{F}\{x(t)\} = \mathcal{F}\{x(t)e^{-j\Omega t}\} = \int x(t) e^{-j\Omega t} dt. \tag{3.1a}
\]

The unit of the time variable \( t \) is seconds, and the unit of the frequency variable \( \Omega \) is radians/second.

**Definition 3.2.2.** The *discrete-time Fourier transform* of a signal \( x(n) \) is defined by

\[
X(\omega) = \text{DTFT}\{x(n)\} = \sum_{n} x(n) e^{-j\omega n}. \tag{3.2a}
\]

The time variable \( n \) indicates sample index, and is dimensionless. The unit of the frequency variable \( \omega \) is radians/sample.

\( X(\Omega) \) and \( X(\omega) \) are said to represent the signals \( x(t) \) and \( x(n) \) in the frequency domain.

It is sometimes convenient to express \( X(\Omega) \) and \( X(\omega) \) in polar form,

\[
X(\Omega) = A(\Omega)e^{j\phi(\Omega)} \quad \text{and} \quad X(\omega) = A(\omega)e^{j\phi(\omega)}, \tag{3.3}
\]

where \( A(\Omega) = |X(\Omega)| \) and \( A(\omega) = |X(\omega)| \) are the amplitude, and \( \phi(\Omega) = |\phi(\Omega)| \) and \( \phi(\omega) = |\phi(\omega)| \) are the phase of \( X(\Omega) \) and \( X(\omega) \), respectively.

The short-time versions of the continuous-time and discrete-time Fourier transform, which we define next, use a short-time *analysis window*. For our purposes, we define an analysis window as follows.

**Definition 3.2.3.** An *analysis window* is a continuous-time signal \( w(t) \) or discrete-time signal \( w(n) \) whose energy is concentrated around the origin in time and in frequency. Specifically, \( w(t) = 0 \) for \( |t| > t_0 \), \( W(\Omega) = 0 \) for \( |\Omega| > \Omega_0 \), or in discrete-time \( w(n) = 0 \) for \( |n| > n_0 \).
and $W(\omega) = 0$ for $|\omega| > \omega_0$. Without loss of generality, we assume that analysis windows are normalized, i.e.,

$$\int w(t)dt = 1,$$

or equivalently

$$\sum_n w(n) = 1.$$

Given these definitions of the Fourier transform and of analysis windows, the short-time Fourier transforms are defined as follows.

**Definition 3.2.4.** The *continuous-time short-time Fourier transform* of a signal $x(t)$ is defined by

$$X(\tau, \Omega) = \text{STFT}\{x(t)\}$$

$$\triangleq \mathcal{F}\{x(t)w(\tau - t)\}$$

$$= \int x(t)w(\tau - t)e^{-j\Omega t}dt,$$

where $w(t)$ is an analysis window. We use the notation

$$X_\Omega(\tau) \equiv X(\tau, \Omega)$$

to denote the continuous-time short-time Fourier transform evaluated at frequency $\Omega$ as a function of time.

**Definition 3.2.5.** The *discrete-time short-time Fourier transform* of a signal $x(n)$ is defined by

$$X(m, \omega) = \text{DTSTFT}\{x(n)\}$$

$$\triangleq \text{DTFT}\{x(n)w(m - n)\}$$

$$= \sum_n x(n)w(m - n)e^{-j\omega n},$$
where \( w(n) \) is an analysis window. We use the notation

\[
X_\omega(m) \equiv X(m, \omega)
\]

(3.10)

to denote the discrete-time short-time Fourier transform evaluated at frequency \( \omega \) as a function of time.

\( X(\tau, \Omega) \) and \( X(m, \omega) \) are said to represent the signals \( x(t) \) and \( x(n) \) in the time-frequency domain.

Modulation transforms use the short-time Fourier transforms of definitions 3.2.4 and 3.2.5 to transform a signal to the time-frequency domain. They subsequently extend these linear time-frequency transforms with a non-linear envelope detector, which transforms a signal from the time-frequency domain into the modulator domain. However, before we define these concepts, we must first address the question: what is an envelope?

3.2.2 Envelopes and carriers

The term envelope commonly refers to a slowly-varying, non-negative real signal \( m(t) \) that is related to a signal \( x(t) \) by \( m(t) \geq |x(t)| \) for all \( t \) (for example in [27, 30, 71]). For real signals, an example of an envelope is the Hilbert envelope,

\[
m_H(t) = |x(t) + j\mathcal{H}\{x(t)\}|
\]

(3.11)

where \( \mathcal{H} \) indicates the Hilbert transform (see section 4.2.1 for details). For complex signals, an example of an envelope is the magnitude envelope,

\[
m_{||}(t) = |x(t)|
\]

(3.12)

In our work, we relax the non-negative real restriction on signal envelopes and allow them to be complex-valued. Since many people associate the term envelope strictly with a non-negative real quantity, we use the more general term modulator instead of envelope to avoid confusion, and to stress the broader definition of our envelopes.
Complementary to a signal’s modulator is a signal’s carrier, which is defined by

\[ c(t) = \frac{x(t)}{m(t)}. \]  

(3.13)

In theory, the carrier is not well defined when \( m(t) \) approaches 0. Therefore, we will define \( c(t) = 0 \) when \( m(t) = 0 \). In practice, we typically let \( c(t) = 0 \) for \( |m(t)| < \epsilon \), for arbitrarily small \( \epsilon \), which results in only negligible errors in \( c(t) \). The definition of (3.13) is a direct consequence of the assumed product model for a signal’s modulator and carrier, i.e.,

\[ x(t) = m(t)c(t). \]  

(3.14)

The decomposition of a signal into a modulator and a carrier is most meaningful for narrowband signals. In that case, the modulator is a low-pass signal that describes the amplitude modulation (AM) of the signal, and the carrier is a narrowband signal that describes the frequency modulation (FM) of the signal. Broadband signals typically contain multiple AM and FM components, which are inadequately described by a single modulator and carrier. Hence, decomposing a broadband signal into a modulator and carrier would result in a broadband modulator and carrier, rather than low-pass and narrowband signals, which would violate our constraint that the modulator should be slowly-varying.

3.2.3 Envelope detectors and carrier estimators

A signal’s modulator is found by applying an envelope detector to the signal. Examples of envelope detectors are the Hilbert envelope and magnitude envelope mentioned earlier. Many other decompositions of a signal into a modulator and carrier are possible [15, 80]. Much of this dissertation is concerned with defining the modulator and carrier more precisely for practical applications. Here, we present the theory of modulation transforms without specifying any envelope detector in particular. We leave the discussion of the various envelope detectors and carrier estimators that we have developed to chapter 4.

We use the following notation for the continuous-time and discrete-time envelope detector throughout this dissertation.
**Definition 3.2.6.** We denote the *continuous-time envelope detector* \( \mathcal{D} \) by

\[
m(t) \triangleq \mathcal{D}\{x(t)\},
\]

or in the time-frequency domain by

\[
M_\Omega(\tau) \triangleq \mathcal{D}\{X_\Omega(\tau)\}.
\]

**Definition 3.2.7.** We denote the *discrete-time envelope detector* \( \mathcal{D} \) by

\[
m(n) \triangleq \mathcal{D}\{x(n)\},
\]

or in the time-frequency domain by

\[
M_\omega(m) \triangleq \mathcal{D}\{X_\omega(m)\}.
\]

A signal’s carrier can be defined as the complement to the signal’s modulator, according to the product model in equation (3.14). Therefore, we introduce the following notation for a *complementary detector* in continuous-time and discrete-time.

**Definition 3.2.8.** We denote the *continuous-time complementary detector* \( \mathcal{D}^c \) by

\[
c(t) = \mathcal{D}^c\{x(t)\}
\]

\[
\triangleq \frac{x(t)}{\mathcal{D}\{x(t)\}} = \frac{x(t)}{m(t)},
\]

or in the time-frequency domain by

\[
C_\Omega(\tau) = \mathcal{D}^c\{X_\Omega(\tau)\}
\]

\[
\triangleq \frac{X_\Omega(\tau)}{\mathcal{D}\{X_\Omega(\tau)\}} = \frac{X_\Omega(\tau)}{M_\Omega(\tau)},
\]
Definition 3.2.9. We denote the discrete-time complementary detector $D^c$ by

$$c(n) = D^c\{x(n)\} \quad (3.21a)$$

or in the time-frequency domain by

$$C_\omega(m) = D^c\{X_\omega(m)\} \quad (3.22a)$$

The output of the complementary detector is the signal’s carrier, which is also referred to by some authors as the signal’s (temporal) fine structure.

It is sometimes more convenient to perform envelope detection by first estimating the carrier, and subsequently finding the modulator by using the relationship in (3.20) and (3.22). In that case, the complementary detector is no longer “complementary”, and is therefore usually referred to as the carrier estimator.

An envelope detector or carrier estimator must satisfy several properties. First, they must be non-linear operators. Would they be linear, then the entire modulation transform could be expressed in linear operations, as is for example shown by Avendaño [7, chapter 3]. In that case modulation transforms, and consequently modulation analysis and filtering, would be part of linear system theory, and would not be an extension to the time-frequency domain.

An envelope detector or carrier estimator should also be a projection, in order to be a well-behaved operator. That is,

$$D\{D\{x(t)\}\} = D\{x(t)\} \quad (3.23)$$

and

$$D^c\{D^c\{x(t)\}\} = D^c\{x(t)\} \quad (3.24)$$

should hold. We will see in chapter 4 that this condition is not easily satisfied in practice, in
the same way that practical LTI filters are not true projections. We therefore allow envelope
detectors and carrier estimators to satisfy projection in a similar approximate sense.

Furthermore, an envelope detector must be frequency shift invariant. That is, for an
analytic signal \( x^+(t) \) and radian frequency \( \omega_0 > 0 \),

\[
\mathcal{D}\{e^{\text{j}\omega_0 t}x^+(t)\} = \mathcal{D}\{x^+(t)\}
\]

(3.25)

and consequently

\[
\mathcal{D}^c\{e^{\text{j}\omega_0 t}x^+(t)\} = e^{\text{j}\omega_0 t}\mathcal{D}^c\{x^+(t)\}.
\]

(3.26)

(the same holds for \( \omega_0 < 0 \) if \( e^{\text{j}\omega_0 t}x^+(t) \) is also analytic). This property must hold for an
envelope detector because the concept of a signal’s modulator is independent of a signal’s
absolute location in frequency, and depends only on the frequencies of a signal in relation
to each other.

Finally, an envelope detector or carrier estimator should preserve bandwidth. That is,
the bandwidth of the modulator \( m(t) \) and the carrier \( c(t) \) should be of the same order as the
bandwidth of the signal \( x(t) \). This restriction is important for distortion free modulation
filtering, since we want to be able to recombine the carrier with an arbitrarily modified
modulator without introducing significant artifacts outside the signal’s original frequency
range.

3.2.4 The modulator domain

An envelope detector transforms a signal from the time-frequency domain to the modulator
domain. In section 3.2.3, however, envelope detectors are only defined to operate on functions of time. We can extend the definition of envelope detectors and carrier estimators to functions of time and frequency by agreeing to the convention that

\[
M(\tau, \Omega) = \mathcal{D}\{X(\tau, \Omega)\} = \mathcal{D}\{X_\Omega(\tau)\}
\]

(3.27)

and

\[
C(\tau, \Omega) = \mathcal{D}^c\{X(\tau, \Omega)\} = \mathcal{D}^c\{X_\Omega(\tau)\}
\]

(3.28)
for continuous-time signals, and similarly

\[ M(m, \omega) = \mathcal{D}\{X(m, \omega)\} = \mathcal{D}\{X_\omega(m)\} \quad (3.29) \]

and

\[ C(m, \omega) = \mathcal{D}^c\{X(m, \omega)\} = \mathcal{D}^c\{X_\omega(m)\} \quad (3.30) \]

for discrete-time signals. We refer to \( M(\tau, \Omega) \) and \( M(m, \omega) \) as the modulators (plural) and to \( C(\tau, \Omega) \) and \( C(m, \omega) \) as the carriers (plural) of the signal \( x(t) \) and \( x(n) \).

Before we can formally define the modulator domain, we need to introduce notation for a general signal domain.

**Definition 3.2.10.** We denote a signal domain \( S \) as the set of complex functions of variable(s) \( \lambda \) defined on \( I \), i.e.,

\[ S(I; \lambda) = \{x(\lambda) : I \rightarrow \mathbb{C}\}. \quad (3.31) \]

So far, we have encountered from the following signal domains: the time domains \( x(t) \in S(\mathbb{R}; t) \) and \( x(n) \in S(\mathbb{N}; n) \), the frequency domains \( X(\Omega) \in S(\mathbb{R}; \Omega) \) and \( X(\omega) \in S(\mathbb{R}; \omega) \), and the time-frequency domains \( X(\tau, \Omega) \in S(\mathbb{R}^2; \tau, \Omega) \) and \( X(m, \omega) \in S(\mathbb{N} \times \mathbb{R}; m, \omega) \).

Given this notation, and given the time-frequency extension of envelope detectors and carrier estimators, we define the modulator domain in continuous-time and discrete-time as follows.

**Definition 3.2.11.** The continuous-time modulator domain is defined by

\[ S_D(\mathbb{R}^2; \tau, \Omega) = \{x \in S(\mathbb{R}^2; \tau, \Omega) \mid \text{there exists an } y \in S(\mathbb{R}^2; \tau, \Omega) \text{ such that } \mathcal{D}\{y\} = x\}. \quad (3.32) \]

**Definition 3.2.12.** The discrete-time modulator domain is defined by

\[ S_D(\mathbb{N} \times \mathbb{R}; m, \omega) = \{x \in S(\mathbb{N} \times \mathbb{R}; m, \omega) \mid \text{there exists an } y \in S(\mathbb{N} \times \mathbb{R}; m, \omega) \text{ such that } \mathcal{D}\{y\} = x\}. \quad (3.33) \]
In other words, the modulator domain is the subset of functions in \( \mathcal{S}(\mathbb{R}^2; \tau, \Omega) \) that are valid modulators with respect to the envelope detector \( \mathcal{D} \).

### 3.2.5 Modulation transforms

The groundwork in sections 3.2.1–3.2.4 leads us to the following definitions of modulation transforms.

**Definition 3.2.13.** We define the *continuous-time modulation transform* of a signal \( x(t) \) by

\[
X(H, \Omega) = \mathfrak{M}\{x(t)\}
\]

\[
\triangleq \mathfrak{F}\{\mathcal{D}\{\text{STFT}\{x(t)\}\}\}\]

\[
= \int \mathcal{D}\left\{\int x(t)w(\tau - t)e^{-j\Omega \tau}dt\right\} e^{-jH\tau}d\tau,
\]

where \( w(t) \) is an analysis window, and \( H \) (capital eta) is a frequency variable representing modulation frequency in radians per second.

**Definition 3.2.14.** We define the *discrete-time modulation transform* of a signal \( x(n) \) by

\[
X(\eta, \omega) = \text{DTMT}\{x(n)\}
\]

\[
\triangleq \text{DTFT}\{\mathcal{D}\{\text{DTSTFT}\{x(n)\}\}\}\]

\[
= \sum_m \mathcal{D}\left\{\sum_n x(n)w(m - n)e^{-j\omega n}\right\} e^{-j\eta m},
\]

where \( w(n) \) is an analysis window, and \( \eta \) is a frequency variable representing modulation frequency in radians per sample.

**Definition 3.2.15.** We define the *continuous-time short-time modulation transform* of a signal \( x(t) \) by

\[
X(T, H, \Omega) = \text{STMT}\{x(t)\}
\]

\[
\triangleq \text{STFT}\{\mathcal{D}\{\text{STFT}\{x(t)\}\}\}\]

\[
= \int \mathcal{D}\left\{\int x(t)w(\tau - t)e^{-j\Omega \tau}dt\right\} e^{-jH\tau}d\tau,
\]
\[
= \int \mathcal{D} \left\{ \int x(t)w(\tau - t)e^{-j\Omega t} dt \right\} v'(T - \tau)e^{-j\Omega \tau} d\tau. \quad (3.36c)
\]

Here, \( w(t) \) and \( v(\tau) \) are an analysis window and a modulation analysis window that are concentrated around the origin in \((\tau, \Omega)\) and \((T, H)\), respectively. In addition, \( H \) (capital eta) is a frequency variable that represents modulation frequency in radians per second, and \( T \) (capital tau) represents time in seconds.

**Definition 3.2.16.** We define the discrete-time short-time modulation transform of a signal \( x(n) \) by

\[
X(l, \eta, \omega) = \text{DTSTMT} \{x(n)\} \quad (3.37a)
\]

\[
\triangleq \text{DTSTFT} \{\mathcal{D} \{\text{DTSTFT} \{x(n)\}\}\} \quad (3.37b)
\]

\[
= \sum_m \mathcal{D} \left\{ \sum_n x(n)w(m - n)e^{-j\omega n} \right\} v(l - m)e^{-j\eta m}. \quad (3.37c)
\]

Again, \( w(n) \) and \( v(m) \) are an analysis window and a modulation analysis window that are concentrated around the origin in \((m, \omega)\) and \((l, \eta)\), respectively. Additionally, \( l \) represents time in samples, and \( \eta \) is a frequency variable representing modulation frequency in radians per sample.

These definitions show that a modulation transform is actually three operations applied in sequence: a short-time Fourier transform, an envelope detector, and another Fourier transform or short-time Fourier transform. Figure 3.1 illustrates all the signal domains that a signal goes through when a modulation transform is applied.

### 3.3 Representations

Definitions 3.2.13 through 3.2.16 of the modulation transforms give rise to two modulation spectral representations of signals: the *modulation spectrogram* and the *modulation spectrum*.

The term modulation spectrogram is commonly used to refer to the output of the short-time modulation transforms \( X(T, H, \Omega) = \text{STMT} \{x(t)\} \) and \( X(l, \eta, \omega) = \text{DTSTMT} \{x(n)\} \). It can either refer to the output of the transform as a whole, to the output at a single time
instance T or l, i.e., $X_T(H, \Omega) \equiv X(T, H, \Omega)$ and $X_l(\eta, \omega) \equiv X(l, \eta, \omega)$, or to the magnitude of these quantities. The origin of the term is not entirely clear, but it appears to have been introduced by Kollmeier and Koch [70], and subsequently adopted by others (e.g., [38, 120, 145]). Occasionally, the term modulation spectrogram is also used to refer to other modulation frequency related quantities, such as the output of the modulation transforms $X(H, \Omega) = \mathfrak{M}\{x(t)\}$ and $X(\eta, \omega) = \text{DMT}\{x(n)\}$. Moreover, some authors use the term modulation spectrum instead of modulation spectrogram (e.g., [61]).

To disambiguate between these different uses of the terms modulation spectrogram and modulation spectrum, we introduce a more consistent naming scheme for the various modulation frequency representations. In our naming scheme, we use the term short-time modulation spectrum for the output of STMT\{x(t)\} and DTSTMT\{x(n)\}, both for the representations as a whole and at a single time instance. The correct term in our naming scheme for the magnitude component of $X_T(H, \Omega)$ and $X_l(\eta, \omega)$ is the short-time modulation magnitude spectrum. We also refer to this as the modulation spectrogram, since that is most consistent with prior use of this term, and corresponds with the use of the term spectrogram in the time-frequency domain. We use the term short-time modulation phase spectrum for the phase component of the short-time modulation spectrum.
In our naming scheme, the term *modulation spectrum* is used for the output of the continuous-time modulation transform \( \mathcal{M}\{x(t)\} \) and the discrete-time modulation transform \( \text{DTMT}\{x(n)\} \). It can either refer to the spectrum of a single modulator at a particular acoustic frequency \( \Omega_0 \) (or \( \omega_0 \)), i.e., \( \mathfrak{F}\{X_{\Omega_0}(\tau)\} \) (or \( \text{DTFT}\{X_{\omega_0}(m)\} \)). It can also refer to the entire joint-frequency domain \( X(H,\Omega) \) (or \( X(\eta,\omega) \)). The magnitude component of the modulation spectrum is called the *modulation magnitude spectrum*, and its phase or complex argument component is called the *modulation phase spectrum*.

### 3.4 Inverse modulation transforms

Modulation transforms are invertible when the output of the complementary detector, i.e. the carriers, of the original input to the forward modulation transform is available. The inverse modulation transform is then a simple reversal of the forward (short-time) Fourier transforms. For completeness, we first give the existing inverse Fourier transforms before defining the inverse modulation transforms.

**Definition 3.4.1.** From definition 3.2.1 of the continuous-time Fourier transform it follows that the *inverse continuous-time Fourier transform* of a spectrum \( X(\Omega) \) is given by

\[
x(t) = \mathfrak{F}^{-1}\{X(\Omega)\} = \frac{1}{2\pi} \int X(\Omega)e^{j\Omega t}d\Omega. \tag{3.38a}
\]

\[
x(t) = \frac{1}{2\pi} \int X(\Omega)e^{j\Omega t}d\Omega. \tag{3.38b}
\]

**Definition 3.4.2.** From definition 3.2.2 of the discrete-time Fourier transform it follows that the *inverse discrete-time Fourier transform* of a spectrum \( x(\omega) \) is given by

\[
x(n) = \text{IDTFT}\{X(\omega)\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega)e^{j\omega n}d\omega. \tag{3.39a}
\]

\[
x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega)e^{j\omega n}d\omega. \tag{3.39b}
\]

**Definition 3.4.3.** From definition 3.2.4 of the continuous-time short-time Fourier transform it follows that the *inverse continuous-time short-time Fourier transform* is given by

\[
x(t) = \text{ISTFT}\{X(\tau,\Omega)\} \tag{3.40a}
\]
This definition is a direct consequence of the normalization of the analysis window of equation (3.5), from which it easily follows that

\[ \int w(t - \tau)d\tau = 1 \text{ for all } t, \quad (3.41) \]

and

\[ x(t) = x(t) \int w(t - \tau)d\tau = \int x(t)w(t - \tau)d\tau. \quad (3.42) \]

Substituting \( x(t) \) in the continuous-time Fourier transform of equation (3.1)

\[
X(\Omega) = \int \left[ \int x(t)w(t - \tau)d\tau \right] e^{-j\Omega t} dt \\
= \int \int x(t)w(t - \tau)e^{-j\Omega t} d\tau dt \\
= \int \int [x(t)w(t - \tau)e^{-j\Omega t} dt] d\tau \\
= \int X(\tau, \Omega)d\tau, \quad (3.43) \]

which, when substituted in equation (3.38), yields the definition of equation (3.40).

**Definition 3.4.4.** From definition 3.2.5 of the discrete-time short-time Fourier transform it follows that the inverse discrete-time short-time Fourier transform is given by

\[
x(n) = \text{IDTSTFT}\{X(m, \omega)\} \\
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_m X(m, \omega)e^{j\omega n} \right] d\omega, \quad (3.44) \]

which follows from the normalization of the analysis window, analogous to definition 3.4.3.

The inverse modulation transforms that follow simply apply the previous definitions, combined with an inversion of the envelope detector.

**Definition 3.4.5.** Let

\[ C(\tau, \Omega) = \mathcal{D}^c \{\text{STFT}\{x(t)\}\} \]
be the carriers of the signal $x(t)$. Furthermore, let

$$X(H, \Omega) = \mathfrak{M}\{x(t)\}$$

be the modulation spectrum of the signal $x(t)$. The inverse continuous-time modulation transform is given by

$$x(t) = \mathfrak{M}^{-1}\{X(H, \Omega)\} \quad (3.45a)$$

$$= \text{ISTFT}\{\delta^{-1}\{X(H, \Omega)\}C(\tau, \Omega)\} \quad (3.45b)$$

$$= \frac{1}{2\pi} \iint \left[ \frac{1}{2\pi} \int X(H, \Omega)e^{jH\tau}dH \right] C(\tau, \Omega)e^{j\Omega t}d\tau d\Omega. \quad (3.45c)$$

Definition 3.4.6. Let

$$C(m, \omega) = \mathfrak{D}^c \{\text{DTSTFT}\{x(n)\}\}$$

be the carriers of the signal $x(n)$. Furthermore, let

$$X(\eta, \omega) = \mathfrak{M}\{x(n)\}$$

be the modulation spectrum of the signal $x(n)$. The inverse discrete-time modulation transform is given by

$$x(n) = \text{IDTMT}\{X(\eta, \omega)\} \quad (3.46a)$$

$$= \text{IDTSTFT}\{\text{IDTFT}\{X(\eta, \omega)\}C(m, \omega)\} \quad (3.46b)$$

$$= \frac{1}{2\pi} \sum_m \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\eta, \omega)e^{j\eta m}d\eta \right] C(m, \omega)e^{j\omega n}d\omega. \quad (3.46c)$$

Definition 3.4.7. Define the carriers of the signal $x(t)$ as in definition 3.4.5, and let

$$X(T, H, \Omega) = \text{STMT}\{x(t)\}$$

be the short-time modulation spectrum of the signal $x(t)$. The inverse continuous-time
short-time modulation transform is given by

\[ x(t) = \text{ISTMT}\{X(T, H, \Omega)\} \]
\[ = \text{ISTFT}\{\text{ISTFT}\{X(T, H, \Omega)\}C(\tau, \Omega)\} \]
\[ = \frac{1}{2\pi} \iint \left[ \frac{1}{2\pi} \iint X(T, H, \Omega)e^{j\tau r}dTdH \right] C(\tau, \Omega)e^{j\Omega t}d\tau d\Omega. \] (3.47c)

**Definition 3.4.8.** Define the carriers of the signal \( x(n) \) as in definition 3.4.6, and let

\[ X(l, \eta, \omega) = \text{DTSTMT}\{x(n)\} \]

be the discrete-time short-time modulation spectrum of the signal \( x(n) \). The inverse discrete-time short-time modulation transform is given by

\[ x(n) = \text{IDTSTMT}\{X(l, \eta, \omega)\} \]
\[ = \text{IDTSTFT}\{\text{IDTSTFT}\{X(l, \eta, \omega)\}C(m, \omega)\} \]
\[ = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_m \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} X(l, \eta, \omega)e^{j\eta m}d\eta \right] C(m, \omega)e^{j\omega n}d\omega. \] (3.48c)

Note that the inverse modulation transforms given in this section are the left inverses of the modulation transforms given in section 3.2.5. That is, they are only the true mathematical inverses for modulation spectra and modulation spectrograms that were obtained by applying a modulation transform to a time-domain signal. For modified or synthesized modulation spectra and modulation spectrograms, they act as pseudo-inverses, which is the topic of section 3.5.

### 3.5 Signal modification in the modulation spectral domain

The definitions of the inverse modulation transforms tell us how to invert the corresponding forward transform. What they do not tell us, however, is whether it is possible to reconstruct a time-domain signal from a modified modulation spectrum or a modified modulation spectrogram, or whether it is possible to synthesize a time-domain signal from an arbitrary modulation spectrum or spectrogram. To address these questions, we must take a closer
look at the conditions under which the inverse transforms exist.

### 3.5.1 Valid subspaces

Figure 3.2 illustrates the conditions that must be satisfied for the inverse transform to exist. It introduces a subspace for each signal domain that was shown in figure 3.1, except for the time domain. We define these subspaces here for the continuous-time case; the discrete-time subspaces are defined similarly, and are omitted for brevity.

Recall definition 3.2.10 of a general signal domain, \( \mathcal{S}(I; \lambda) = \{x(\lambda) : I \to \mathbb{C}\} \), such that the time-frequency domain is defined as \( \mathcal{S}([\tau]; \Omega) \). Then the subspaces shown in figure 3.2 are defined as follows.

**Definition 3.5.1.** We define the valid time-frequency domain \( \mathcal{R}_{\text{STFT},\mathcal{S}} \) as the set of time-frequency functions that are valid STFTs. A time-frequency function is a valid STFT if it is the STFT of a time domain signal \( x(t) \). Formally,

\[
\mathcal{R}_{\text{STFT},\mathcal{S}} = \{x \in \mathcal{S}([\tau], \Omega) \mid \text{there exists an } y \in \mathcal{S}(\mathbb{R}; t) \text{ such that } \text{STFT}\{y\} = x\}.
\]  

(3.49)
Definition 3.5.2. We define the valid modulator domain \( R_D \) as the set of time-frequency functions that are valid modulators. A time-frequency function is a valid modulator if it is the modulator of a valid STFT. Formally,

\[
R_D = \{ x \in S(\mathbb{R}^2; \tau, \Omega) | \text{there exists an } y \in R_{STFT,S} \text{ such that } \mathcal{D}\{y\} = x \}. \tag{3.50}
\]

Definition 3.5.3. We define the valid carrier domain \( R_{Dc} \) as the set of time-frequency functions that are valid carriers. A time-frequency function is a valid carrier if it is the carrier of a valid STFT. Formally,

\[
R_{Dc} = \{ x \in S(\mathbb{R}^2; \tau, \Omega) | \text{there exists an } y \in R_{STFT,S} \text{ such that } \mathcal{D}^c\{y\} = x \}. \tag{3.51}
\]

Definition 3.5.4. We define the valid joint-frequency domain \( R_{\mathcal{F},R_D} \) as the set of joint-frequency functions that are valid modulation spectra. A joint-frequency function is a valid modulation spectrum if it is the Fourier transform of a valid modulator. Formally,

\[
R_{\mathcal{F},R_D} = \{ x \in S(\mathbb{R}^2; H, \Omega) | \text{there exists an } y \in R_D \text{ such that } \mathcal{F}\{y\} = x \}. \tag{3.52}
\]

Definition 3.5.5. We define the valid short-time joint-frequency domain \( R_{STFT,R_D} \) as the set of time-joint-frequency functions that are valid modulation spectrograms. A time-joint-frequency function is a valid modulation spectrogram if it is the short-time Fourier transform of a valid modulator. Formally,

\[
R_{STFT,R_D} = \{ x \in S(\mathbb{R}^3; T, H, \Omega) | \text{there exists an } y \in R_D \text{ such that } \text{STFT}\{y\} = x \}. \tag{3.53}
\]
3.5.2 Valid modifications

In this dissertation we are interested in modifications of a signal’s modulators. Given the subspaces defined above, we therefore define valid modulator modifications as follows.

**Definition 3.5.6.** We define a *valid modulator modification* as a modification of the modulators in the valid modulator domain $\mathcal{R}_D$, the valid joint-frequency-domain $\mathcal{R}_{\bar{D},\mathcal{R}_D}$, or the valid short-time joint-frequency domain $\mathcal{R}_{\text{STFT},\mathcal{R}_D}$ such that the modified modulators remain inside the corresponding valid subspace.

For example, given the modulation spectrum $\mathfrak{M}\{x(t)\} = X(H, \Omega) \in \mathcal{R}_{\bar{D},\mathcal{R}_D}$, the modification

$$\tilde{X}(H, \Omega) = X(H, \Omega)G(H, \Omega)$$

is considered valid if $\tilde{X}(H, \Omega) \in \mathcal{R}_{\bar{D},\mathcal{R}_D}$, since that guarantees, by definition, that

$$\mathfrak{F}^{-1}\{\tilde{X}(H, \Omega)\} \in \mathcal{R}_D.$$ 

3.5.3 Inverting the envelope detector

It is clear that the definition of a valid modulator modification ensures the invertibility of the (short-time) Fourier transform between the modulator domain and (short-time) joint-frequency domain after modification of the modulators. The definition also establishes the invertibility of the short-time Fourier transform between the time-domain and the time-frequency domain after modification. However, it is not yet clear if the “invertibility” of the envelope detector after modification of the modulators is also guaranteed.

Recall from equations (3.45)–(3.48) that the envelope detector is “inverted” by recombining the modulator with the carrier. Formally, let $X(\tau, \Omega) = \text{STFT}\{x(t)\}$, and let $M(\tau, \Omega) = \mathcal{D}\{X(\tau, \Omega)\}$ and $C(\tau, \Omega) = \mathcal{D}^c\{X(\tau, \Omega)\}$. Then $M(\tau, \Omega) \in \mathcal{R}_D$ and $C(\tau, \Omega) \in \mathcal{R}_{\mathcal{D}^c}$ by definition, and

$$X(\tau, \Omega) = M(\tau, \Omega)C(\tau, \Omega)$$

by equation (3.20). Recombining the unmodified modulator $\tilde{M}(\tau, \Omega) = M(\tau, \Omega)$ with the
original carrier $C(\tau, \Omega)$ restores the original STFT,

$$\tilde{X}(\tau, \Omega) = \tilde{M}(\tau, \Omega)C(\tau, \Omega) = M(\tau, \Omega)C(\tau, \Omega) = X(\tau, \Omega),$$

and hence $\tilde{X}(\tau, \Omega) \in \mathcal{R}_{\text{STFT},S}$.

We do not necessarily have the guarantee that $\tilde{X}(\tau, \Omega) \in \mathcal{R}_{\text{STFT},S}$ when the carriers are recombined with an arbitrary modulator $\tilde{M}(\tau, \Omega) \in \mathcal{R}_D$. Since the modulators are modified independently from the carriers, it could very well be that in general

$$\tilde{X}(\tau, \Omega) = \tilde{M}(\tau, \Omega)C(\tau, \Omega) \notin \mathcal{R}_{\text{STFT},S}$$

This leads to the concept of a subspace of the time-frequency domain of signals that, when recombined with a signal’s original carriers, produce valid STFTs. We can formally define this subspace as follows.

**Definition 3.5.7.** Given the carriers $C(\tau, \Omega) = \mathcal{D}^c\{X(\tau, \Omega)\}$ of a signal $X(\tau, \Omega)$, we define the **valid modified modulators subspace** $\mathcal{R}_{\tilde{M}(\tau, \Omega)} \subset \mathcal{S}(\mathbb{R}^2; \tau, \Omega)$ such that

$$\tilde{M}(\tau, \Omega) \in \mathcal{R}_{\tilde{M}(\tau, \Omega)} \text{ if and only if } \tilde{M}(\tau, \Omega)C(\tau, \Omega) \in \mathcal{R}_{\text{STFT},S}. \quad (3.54)$$

The relationship of the valid modified modulators subspace with the other time-frequency subspaces is illustrated in figure 3.3. Ideally, modulator modifications should be restricted to produce only modulators for which $\tilde{M}(\tau, \Omega) \in \mathcal{R}_{\tilde{M}(\tau, \Omega)} \cap \mathcal{R}_D$. Such modified modulators are valid modulators with respect to the envelope detector, because they are in $\mathcal{R}_D$, and they recombine with the signal’s carriers into a valid STFT, because they are in $\mathcal{R}_{\tilde{M}(\tau, \Omega)}$. It is clear that this intersection of the valid modified modulators subspace $\mathcal{R}_{\tilde{M}(\tau, \Omega)}$ and the valid modulators $\mathcal{R}_D$ is not empty, since it contains at least the original modulators, i.e., $M(\tau, \Omega) \in \mathcal{R}_{\tilde{M}(\tau, \Omega)} \cap \mathcal{R}_D$. At this time, however, a complete characterization of this subspace for any envelope detector remains an open problem.

Figure 3.3 suggests that there exist time-frequency functions $X(\tau, \Omega)$ outside the modulator domain, but inside the valid modified modulator subspace $\mathcal{R}_{\tilde{M}(\tau, \Omega)}$. This means that
in theory we could allow modulator modifications to produce invalid modulators, i.e., time-frequency functions that are not in the modulator domain, but that still map to valid STFTs when recombined with the signal’s original carriers $C(\tau, \Omega)$. This would be a considerable relaxation of the restriction on modulator modifications that $\tilde{M}(\tau, \Omega) \in \mathcal{R}_{\tilde{M}(\tau, \Omega)} \cap \mathcal{R}_D$. However, when the envelope detector represents a physical envelope property of a signal, such as its energy or its density, there seems little reason to accept such non-envelope like modulators.

If a modulator modification results in modulators $\tilde{M}(\tau, \Omega) \in \mathcal{R}_{\tilde{M}(\tau, \Omega)}$, we are guaranteed by definition that the modified modulators $\tilde{M}(\tau, \Omega)$, when recombined with the signal’s original carriers $C(\tau, \Omega)$, result in a valid STFT $\tilde{X}(\tau, \Omega) \in \mathcal{R}_{\text{STFT},S}$. This in turn guarantees that we can invert the STFT to obtain a time-domain signal $\tilde{x}(t)$. From definitions 3.2.4 and 3.4.3 we have that $\tilde{x}(t)$ and $\tilde{X}(\tau, \Omega)$ are short-time Fourier transform pairs, i.e.,

$$\tilde{x}(t) \xrightarrow{\text{STFT}} \tilde{X}(\tau, \Omega). \quad (3.55)$$
By definition, we also have that $X(\tau, \Omega)$ and $(M(\tau, \Omega), C(\tau, \Omega))$ are envelope detector pairs,

$$X(\tau, \Omega) \xrightarrow{\mathcal{D}} (M(\tau, \Omega), C(\tau, \Omega)).$$

(3.56)

However, we are not given that $\tilde{X}(\tau, \Omega)$ and $(\tilde{M}(\tau, \Omega), C(\tau, \Omega))$ are envelope detector pairs,

$$\tilde{X}(\tau, \Omega) \xrightarrow{\mathcal{D}} (\tilde{M}(\tau, \Omega), C(\tau, \Omega)),$$

(3.57)

unless the envelope detector is modulator modification invariant, which we define as follows.

**Definition 3.5.8.** Let $\mathcal{D}\{X(\tau, \Omega)\} = M_1(\tau, \Omega)$ be the modulators and $\mathcal{D}^c\{X(\tau, \Omega)\} = C_1(\tau, \Omega)$ be the carriers of a signal $X(\tau, \Omega)$, such that

$$X(\tau, \Omega) = M_1(\tau, \Omega)C_1(\tau, \Omega).$$

(3.58)

Consider the modified modulators $\tilde{M}_1(\tau, \Omega)$, such that

$$\tilde{X}(\tau, \Omega) = \tilde{M}_1(\tau, \Omega)C_1(\tau, \Omega).$$

(3.59)

Let $\mathcal{D}\{\tilde{X}(\tau, \Omega)\} = M_2(\tau, \Omega)$ be the modulators and $\mathcal{D}^c\{\tilde{X}(\tau, \Omega)\} = C_2(\tau, \Omega)$ be the carriers of the modified signal $\tilde{X}(\tau, \Omega)$. The envelope detector $\mathcal{D}$ or carrier estimator $\mathcal{D}^c$ is modulator modification invariant if and only if

$$M_2(\tau, \Omega) = \tilde{M}_1(\tau, \Omega) \text{ and } C_2(\tau, \Omega) = C_1(\tau, \Omega).$$

(3.60)

### 3.5.4 LTI modulation filtering

The intention of this dissertation is to extend the concepts of linear system theory to the modulator domain and the modulation frequency domain. We are therefore particularly interested in one kind of modulator modification: filtering a signal’s modulators with an LTI filter. We distinguish two types of LTI modulation filters: frequency independent filters,

$$\tilde{M}_\Omega(\tau) = M_\Omega(\tau) * g(\tau),$$

(3.61)
and frequency dependent filters,

\[ \tilde{M}_\Omega(\tau) = M_\Omega(\tau) \ast g_\Omega(\tau). \] (3.62)

According to our definition of valid modulator modifications (definition 3.5.6), both type of filters are valid if and only if the modulator domain \( \mathcal{R}_\mathcal{D} \) is closed under convolution. If we also take the invertibility of the envelope detector in consideration, then LTI filters are valid modifications if and only if \( \mathcal{R}_\mathcal{D} \cap \mathcal{M}(\tau, \Omega) \) is closed under convolution.

3.5.5 Pseudo-inverses

The conditions imposed on valid modifications in sections 3.5.2–3.5.4 are difficult to meet in practice. To our knowledge, no envelope detector has been found for which \( \mathcal{R}_\mathcal{D} \cap \mathcal{M}(\tau, \Omega) \) is closed under convolution. As a result, practical modulation filtering systems produce modified modulators that are invalid in the sense of our definition.

Even though the modified modulators may be invalid, these systems can still be used in a meaningful way because the inverse transforms defined in equations (3.45)–(3.48) act as pseudo-inverses on invalid data. For example, for a signal \( X(\tau, \Omega) \in \mathcal{R}_{\text{STFT,ST}} \) we have

\[ \text{ISTFT}\{X(\tau, \Omega)\} = x(t) \quad \text{and} \quad \text{STFT}\{x(t)\} = X(\tau, \Omega) \] (3.63)

but for a signal \( X(\tau, \Omega) \in \mathcal{S}(\mathbb{R}^2; \tau, \Omega) \setminus \mathcal{R}_{\text{STFT,ST}} \) we have

\[ \text{ISTFT}\{X(\tau, \Omega)\} = x'(t) \quad \text{and} \quad \text{STFT}\{x'(t)\} = X'(\tau, \Omega) \] (3.64)

where \( X'(\tau, \Omega) \neq X(\tau, \Omega) \). The inverse short-time Fourier transform still produces a time-domain signal \( x'(t) \) for \( X(\tau, \Omega) \in \mathcal{S}(\mathbb{R}^2; \tau, \Omega) \setminus \mathcal{R}_{\text{STFT,ST}} \), but \( x'(t) \) and \( X(\tau, \Omega) \) no longer form a short-time Fourier transform pair. Pseudo-invertibility also applies to the second short-time frequency transform (from the time-frequency domain to the short-time joint-frequency domain).

The pseudo-invertibility of the short-time Fourier transform has been studied carefully by Griffin and Lim [39] for continuous-time, and by Dembo and Malah [23] for discrete-time.
Griffin and Lim [39] defined an iterative procedure to approximate the time-domain signal $x'(t)$ with an STFT that is closest to the desired $X(\tau, \Omega)$ in the mean-squared error sense. In section 4.6 we discuss our study of their technique when applied to modulation filtering systems. Our study shows that the pseudo-inverse of the short-time Fourier transform produces relatively minimal approximation errors in practice.

### 3.6 Envelope detectors

So far we have discussed modulation transforms in an abstract way, without specifying the envelope detector. We generally distinguish two types of envelope detectors, which we call *incoherent* and *coherent* detectors. The difference between the two types is that incoherent detectors estimate the modulators using a magnitude or magnitude-like operation, whereas the coherent detectors use a carrier estimator based on an instantaneous frequency property of the signal.

The most prominent example of an incoherent envelope detector is the magnitude detector,

$$D_{||}\{X_\Omega(\tau)\} = |X_\Omega(\tau)|.$$  \hspace{1cm} (3.65)

The magnitude detector’s modulator domain is not closed under convolution, because $\tilde{M}_\Omega(\tau) = |X_\Omega(\tau)| \ast g(\tau)$ is not guaranteed to be non-negative for all $g(t)$. For example, it is very likely that $\tilde{M}_\Omega(\tau) < 0$ for some $\tau$ when $g(t)$ is a high-pass modulation filter that attenuates or removes the DC component of $|X_\Omega(\tau)|$. Furthermore, modulators of the magnitude detector are not necessarily bandlimited to the bandwidth of the signals it operates on [30], which violates our interpretation of modulators (see section 3.2.2). Despite their obvious drawbacks, magnitude detectors and magnitude-like detectors are often used in practice (e.g., [27, 110, 134]) because they are easy to compute, and because they represent the non-negative instantaneous energy of the underlying signal.

Much of our research efforts have been focused on defining coherent envelope detectors that are free of the problems associated with the incoherent detectors. All coherent estimators estimate the instantaneous frequency of the signal $X_\Omega(\tau)$ by using their own kind of
The coherent envelope is then found by equation (3.20) or (3.22). By estimating the instantaneous frequency of the signal, coherent detectors are able to control the bandwidth of the carrier, and thereby control the bandwidth of the modulator. In addition, their modulator domain is closed under convolution because coherent envelopes are complex functions of time. The various coherent envelope detectors that we have defined in our work are discussed in section 4.4.

### 3.7 Sampled transforms

The discrete-time modulation transforms and inverses presented so far in this chapter are defined in terms of the continuous frequency variables $\omega$ and $\eta$. In digital implementations of the transforms, these variables must be sampled to be representable on a computer. In this section we introduce discrete versions of the modulation transforms. First, however, we establish the notation of the discrete Fourier transforms that they are based on.

#### 3.7.1 Discrete Fourier transforms

**Definition 3.7.1.** The *discrete Fourier transform* is defined by

\[
X(k) = \text{DFT}\{x(n)\} \triangleq \sum_n x(n) e^{-j2\pi nk/K}, \text{ for } k = 0, \ldots, K - 1. \tag{3.67b}
\]

**Definition 3.7.2.** The *discrete short-time Fourier transform* is defined by

\[
X(m, k) = \text{DSTFT}\{x(n)\} \triangleq \sum_n x(n) w(m - n) e^{-j2\pi nk/K}, \text{ for } k = 0, \ldots, K - 1, \tag{3.68b}
\]
where $w(n)$ is an analysis window. We use the notation

$$X_k(m) \equiv X(k,m)$$

(3.69)

to denote the discrete short-time Fourier transform evaluated at frequency index $k$ as a function of time.

### 3.7.2 Discrete modulation transforms

Given the definitions of the discrete Fourier transforms, we define the discrete modulation transforms as follows.

**Definition 3.7.3.** We define the **discrete modulation transform** by

$$X(i,k) = \text{DMT}\{x(n)\}$$

$$\triangleq \text{DFT}\{\mathcal{D}\{\text{DSTFT}\{x(n)\}\}\}\right\} \right\}$$

$$= \sum_m \mathcal{D} \left\{ \sum_n x(n)w(m-n)e^{-j2\pi nk/K} \right\} e^{-j2\pi mi/I},$$

(3.70a)

(3.70b)

(3.70c)

where $w(n)$ is an analysis window.

Both $X(i,k)$ and $X_k(i) \equiv X(i,k)$, the spectrum of the modulator at the $k$-th acoustic frequency, are referred to as the **(discrete) modulation spectrum**.

**Definition 3.7.4.** We define the **discrete short-time modulation transform** by

$$X(l,i,k) = \text{DSTMT}\{x(n)\}$$

$$\triangleq \text{DSTFT}\{\mathcal{D}\{\text{DSTFT}\{x(n)\}\}\}\right\} \right\}$$

$$= \sum_m \mathcal{D} \left\{ \sum_n x(n)w(m-n)e^{-j2\pi nk/K} \right\} v(l-m)e^{-j2\pi mi/I},$$

(3.71a)

(3.71b)

(3.71c)

for $i = 0, \ldots, I - 1$ and $k = 0, \ldots, K - 1$, where $w(n)$ and $v(m)$ are analysis windows. We use the notation

$$X_l(i,k) \equiv X(l,i,k)$$

(3.72)
to denote the discrete short-time modulation transform at a particular time instance \( l \).

### 3.7.3 Multi-rate implementation

The discrete short-time modulation transform defined above is highly redundant, since for every sample in the input \( x(n) \) there is a \( I \times K \)-point modulation spectrogram \( X_l(i, k) \). For implementation efficiency, the DMT and DSTMT can be based on a multi-rate version of the discrete short-time Fourier transform that is decimated in \( m \) by a factor \( R < K \), defined as follows.

**Definition 3.7.5.** The *decimated discrete short-time Fourier transform* is defined by

\[
X(mR, k) = \text{DSTFT}_R\{x(n)\} \triangleq \sum_n x(n)w(mR - n)e^{-j2\pi nk/K}, \text{ for } k = 0, \ldots, K - 1.
\]  

**Definition 3.7.6.** We define the *decimated discrete modulation transforms* by

\[
X(i, k) = \text{DMT}_R\{x(n)\} \triangleq \text{DFT}\{\mathcal{D}\{\text{DSTFT}_R\{x(n)\}\}\}
\]

\[
= \sum_m \mathcal{D}\left\{ \sum_n x(n)w(mR - n)e^{-j2\pi nk/K} \right\}e^{-j2\pi mi/I},
\]

Similarly, the short-time modulation spectrogram time variable \( l \) is often decimated by a factor \( S < I \) in practice, resulting in the decimated discrete modulation transform.

**Definition 3.7.7.** The *decimated discrete short-time modulation transforms* is defined by

\[
X(lRS, i, k) = \text{DSTMT}_{R,S}\{x(n)\} \triangleq \text{DSTFT}_S\{\mathcal{D}\{\text{DSTFT}_R\{x(n)\}\}\}
\]

\[
= \sum_m \mathcal{D}\left\{ \sum_n x(n)w(mR - n)e^{-j2\pi nk/K} \right\}v(lS - m)e^{-j2\pi mi/I}.
\]

If the input signal \( x(n) \) was sampled at \( F_s \) Hz, the sampling frequency of the DSTFT
time variable \( m \) after decimation becomes \( F_s/R \), which corresponds to the point \( \eta = 2\pi \) in the DMT and DSTMT.

### 3.7.4 Reconstruction from a sampled transform

The discrete short-time modulation transform is invertible when \( K \) and \( I \) are chosen suitably large for the analysis windows \( w(n) \) and \( v(m) \). Reconstructing a time-domain signal from a modified DMT or DSTMT follows the procedures given in section 3.4. The validity constraints on reconstruction from the continuous-frequency modulation transforms, as discussed in section 3.5, also apply to reconstruction from a sampled modulation transform. Reconstructing the multi-rate versions of these transforms, however, is different in one aspect. The multi-rate modulation transforms require the inversion of one or two decimated DSTFTs, as equations (3.74b) and (3.75b) show. Inverting a decimated DSTFT requires interpolation and anti-imaging by a synthesis window that is appropriate for the analysis window and downsampling factor that was used in the forward transform. The conditions on the analysis window, synthesis window and the downsampling factor to properly invert the decimated DSTFT are well documented by for example Crochiere and Rabiner [17, pp. 326–346]. Although these conditions are an important practical aspect of modulation transforms, we will not discuss them here, but rather recommend the interested reader to consult this excellent book.

### 3.8 Filterbank interpretation

In sections 3.2–3.7, we have presented modulation transforms as a cascade of a short-time Fourier transform, an envelope detector, and a second transform such as the (short-time) Fourier transform. We have used the short-time Fourier transform as the initial transform for clarity of presentation. It allowed us to develop our modulation analysis and filtering theory, and discuss the dependencies between the signal domains encompassed by modulation transforms, without being distracted by details of the initial transform. In this section we develop an alternative and more general view on modulation transforms by the well-known interpretation of the DSTFT as a filterbank.
It is well-known that the DSTFT can be interpreted as a filterbank (see for example [17, pp. 296–302] or [146, pp. 116–117]). To interpret the DSTFT as a filterbank in our notation, we need to slightly manipulate equation (3.68) into the right form. Recall that the DSTFT as a function of time is defined by

$$X_k(m) = \sum_n x(n) w(m-n) e^{-j2\pi nk/K}$$  \hspace{1cm} (3.76a)

$$= x_k(n) * w(n),$$  \hspace{1cm} (3.76b)

for \( k = 0, \ldots, K - 1 \), where

$$x_k(n) = x(n) e^{-j2\pi nk/K}.$$  \hspace{1cm} (3.77)

The \( k \)-th frequency channel of the DSTFT can therefore be regarded as the original signal, shifted downward in frequency by \( \omega_k = 2\pi k/K \) and lowpass filtered by \( w(n) \). If we multiply the right hand side of equation (3.76) by \( e^{-j2\pi mk/K} e^{j2\pi mk/K} = 1 \), we obtain

$$X_k(m) = e^{-j2\pi mk/K} \sum_n x(n) w(m-n) e^{j2\pi (m-n)k/K}$$  \hspace{1cm} (3.78a)

$$= e^{-j2\pi mk/K} \hat{X}_k(m), \ k = 0, \ldots, K - 1,$$  \hspace{1cm} (3.78b)

where

$$\hat{X}_k(m) = \sum_n x(n) w(m-n) e^{j2\pi (m-n)k/K}$$  \hspace{1cm} (3.79a)

$$= x(n) * w_k(n),$$  \hspace{1cm} (3.79b)

with

$$w_k(n) = w(n) e^{j2\pi nk/K}.$$  \hspace{1cm} (3.80)

Through a change of variable, equation (3.79a) can be written as

$$\hat{X}_k(m) = \sum_p x(p+m) w(-p) e^{-j2\pi pk/K}$$  \hspace{1cm} (3.81a)

$$= \mathcal{F}\{x(n+m) w(-n)\}.$$  \hspace{1cm} (3.81b)
Equation (3.81) is sometimes referred to in the literature as the moving window transform (MWT) of $x(n)$ [6]. As equations (3.79b) and (3.80) show, the moving window transform of $x(n)$ can be regarded as the original signal filtered by a set of bandpass filters, which establishes the interpretation of the DSTFT as a filterbank. Each bandpass filter is a modulated version of the analysis window $w(n)$, centered at $\omega_k = 2\pi k/K$.

A more general view of modulation transforms can be obtained from the filterbank interpretation by relaxing the strict relationship of equation (3.80) between the subband filters. In this relaxed view of modulation transforms, our definition of modulation systems includes all of the modulation analysis and filtering approaches we reviewed in chapter 2. For example, by replacing the DFT filterbank with a critical band filterbank, our definition of modulation transforms includes the RASTA modulation filtering technique [47]; replacing it with a filterbank with $\frac{1}{4}$, $\frac{1}{2}$ or 1-octave filter spacing yields the technique by Drullman et al. [27].

We will refer to a modulation transform that is based on a filterbank other than the DFT filterbank as a generalized modulation transform. In generalized modulation transforms, the envelope detector and carrier estimator operate on the filterbank subbands rather than on the time-trajectories of the short-time Fourier transform. To stress the subband structure of generalized modulation transforms, it is convenient to restate their definition in continuous-time notation.

### 3.9 Generalized modulation filtering framework

The generalized modulation filtering framework that follows from the filterbank interpretation of modulation transforms is depicted in the signal flow diagram in figure 3.4. In the framework, the broadband signal $x(t)$ is passed through a filterbank that is characterized by a set of LTI bandpass filters $h_k(t)$,

$$x_k(t) = x(t) \ast h_k(t), \text{ for } k = 1, \ldots, K.$$  \hspace{1cm} (3.82)

The resulting subband signals $x_k(t)$ are passed to the envelope detector $\mathcal{D}$ and carrier estimator $\mathcal{D}^c$, which estimate a modulator signal $m_k(t)$ and a carrier signal $c_k(t)$, independently.
for each subband signal,

\[ m_k(t) = \mathcal{D}\{x_k(t)\}, \quad (3.83) \]
\[ c_k(t) = \mathcal{D}^c\{x_k(t)\}, \quad (3.84) \]

satisfying the product model

\[ x_k(t) = m_k(t)c_k(t). \quad (3.85) \]

Next, the modulator signals are filtered by LTI filters \( g_k(t) \),

\[ \tilde{m}_k(t) = m_k(t) * g_k(t). \quad (3.86) \]

In practical modulation systems care should be taken to realign the modulators after filtering, since each filter \( g_k(t) \) may have a different group delay. Furthermore, the carriers should also be delayed appropriately before recombining them with the filtered modulators. After filtering, the modulators are recombined with the original subband carriers into modified subband signals,

\[ \tilde{x}_k(t) = \tilde{m}_k(t)c_k(t). \quad (3.87) \]

From the modified subband signals, a modulation filtered broadband signal is obtained. This process, commonly referred to as filterbank reconstruction, is done by the filterbank
summation method, which is given by

$$\tilde{x}(t) = \sum_{k=1}^{K} \tilde{x}_k(t). \quad (3.88)$$

The filterbank summation method is only a proper inverse of the filterbank if the filterbank satisfies \textit{perfect reconstruction}. Perfect reconstruction is satisfied when

$$x(t) - \sum_{k=1}^{K} x_k(t) = x(t) - \sum_{k=1}^{K} x(t) * h_k(t) = 0. \quad (3.89)$$

This expression can be rewritten as

$$x(t) - \sum_{k=1}^{K} x(t) * h_k(t) = x(t) - x(t) * \left[ \sum_{k=1}^{K} h_k(t) \right] \quad (3.90a)$$

$$= x(t) * \left[ \delta(t) - \sum_{k=1}^{K} h_k(t) \right] \quad (3.90b)$$

$$= 0, \quad (3.90c)$$

which implies that for perfect reconstruction the filters in the filterbank should sum to an impulse,

$$\sum_{k=1}^{K} h_k(t) = \delta(t). \quad (3.91)$$

In our theory on modulation analysis and filtering we assume that the filterbank satisfies perfect reconstruction. For some applications, however, a near-perfect reconstruction filterbank, satisfying

$$\left| x(t) - \sum_{k=1}^{K} x(t) * h_k(t) \right| < \epsilon, \quad (3.92)$$

for \( \epsilon \) small may be acceptable.

In the generalized modulation filtering framework, signal modifications are subject to the constraints defined in section 3.5, with the exception that they are not based on the
valid STFT domain but on the valid subbands domain, which we define next. Let
\[ \{s_k(t)\}_{k=1}^{K} = \{s_1(t), \ldots, s_K(t)\} \quad (3.93) \]
denote a set of \( K \) arbitrary time-domain signals. Let equation (3.82), the operation of applying a filterbank to the signal \( x(t) \), be denoted compactly by
\[ \text{FB}\{x(t)\} = \{x_k(t)\}_{k=1}^{K}. \quad (3.94) \]
The subband signals \( \{x_k(t)\}_{k=1}^{K} \) are said to represent the signal \( x(t) \) in the subband domain, which we denote by \( \{S_k(\mathbb{R}; t)\}_{k=1}^{K} \). The valid subbands domain is defined as follows.

**Definition 3.9.1.** We define the **valid subbands** domain \( \mathcal{R}_{\text{FB,}S} \) as the set of subband signals that are subband signals of a time domain signal \( x(t) \). Formally,
\[ \mathcal{R}_{\text{FB,}S} = \{ x \in \{S_k(\mathbb{R}; t)\}_{k=1}^{K} \mid \text{there exist an } y \in S(\mathbb{R}; t) \text{ such that } \text{FB}\{y\} = x \}. \quad (3.95) \]

Given this definition, the definitions of the “valid domains” (definitions 3.5.2 through 3.5.5) and the definitions of the valid modulator modification and the valid modified modulators subspace (definition 3.5.6 and 3.5.7) can be rephrased in terms of the subband domain \( \{S_k(\mathbb{R}; t)\} \) and the valid subbands domain \( \mathcal{R}_{\text{FB,}S} \). However, we omit these definitions here for brevity.

As discussed in section 3.5.5, practical STFT-based modulation filtering systems generally produce modified modulators that violate our definition of valid modulators. That is, the modified modulators no longer constitute a valid STFT when they are recombined with the signal’s original carriers. Recall from section 3.5.5, however, that we can still use modulation filtering systems in a meaningful way in practice, because the inverse STFT acts as a pseudo-inverse for invalid STFTs.

The same reasoning holds for generalized modulation filtering systems. For arbitrary modulation filters \( \{g_k(t)\} \), the modified modulators \( \{\tilde{m}_k(t)\} \) in equation (3.86) will not be
in the valid modified modulators subspace. Hence, the modified subbands $\tilde{x}_k(t)$ are not valid subbands with respect to the modulation system’s filterbank. Nevertheless, it is easy to see from equation (3.88) that the filterbank summation reconstruction technique acts as a pseudo-inverse as well. Therefore, generalized modulation filtering systems are equally valuable in practice.

### 3.10 Signal model

The signal model that underlies both the discrete modulation transforms and their filterbank interpretation is given in its simplest form by the following expression:

$$x(t) = \sum_{k=1}^{N} m_k(t)c_k(t). \tag{3.96}$$

The usual interpretation of this model is that the carriers $c_k(t)$ are high-frequency signals that are modulated by low-frequency modulators $m_k(t)$. Furthermore, it is assumed that each carrier captures all the frequency modulation (FM) and that each modulator captures all the amplitude modulation (AM) in a narrowband region of $x(t)$’s spectrum.

This signal model resembles the framework of sinusoidal modeling [88, Eq. (2)], but there are a few key differences. First, each modulator and carrier describes the AM and FM of a part of the signal’s spectrum for the duration of the signal, unlike the explicit ‘birth’ and ‘death’ of partials in sinusoidal modeling [88, Sec. IV]. Second, all carrier/modulator pairs describe disjoint regions of the signal’s spectrum. This is in contrast with sinusoidal modeling, where partials may overlap in frequency and are not explicitly tied to a fixed part of the spectrum.

The use of a linear, time-invariant filterbank in modulation analysis and filtering implies two additional assumptions on the signal model of equation (3.96): (i) each of the modulated carriers $m_k(t)c_k(t)$ is bandlimited to the bandwidth of one of the filterbank filters, and (ii) at most one of the modulated carriers falls within each of the filterbank filters.

It is generally undesirable to have a modulation system whose parameters depend on the input signal. In our experience it seems unavoidable, however, that the number $K$ and the shape $h_k(t)$ of the filters in the filterbank depend on the characteristics of the type of input
The filterbank must reflect to some extent the expected number, frequency location and bandwidth of the carriers $c_k(t)$ in the input signal. As a general rule, the number of filters $K$ must be in the order of the expected number of carriers $N$ in (3.96). Too few (broad) filters can cause neighboring carriers to be resolved as modulators rather than as individual carriers. Too many (narrow) filters results in over-representation of the input signal in acoustic frequency, causing the envelope detector and carrier estimator to fit noise in low-energy subbands. Narrow filters also introduce a significant amount of undesirable FM to AM transduction for carriers with time-varying frequency.

### 3.11 Summary

In this chapter we presented a theory of modulation frequency analysis and filtering. The theory is based on the concept of a modulation transform, which we define in terms of existing time-frequency transforms such as the short-time Fourier transform. We defined the modulation frequency representations of a signal that can be obtained with the modulation transforms. We then specified the inverse modulation transforms, and discussed the conditions that must be satisfied for their existence. We demonstrated that, in theory, the inverse modulation transform only exist for a very limited type of envelope modifications, but that in practice more modifications are acceptable because useful signals can be reconstructed with the pseudo-inverses of the inverse modulation transform.

Next, we defined sampled versions of the modulation transforms, and demonstrated how the use of a sampled short-time Fourier transform as the time-frequency transform leads to a more general filterbank interpretation of modulation transforms. We redefined the modulation transform in a generalized modulation filtering framework, and discussed the signal model that is assumed when a signal is analyzed or filtered in the context of the generalized modulation framework.
Chapter 4

PRACTICE

4.1 Introduction

In chapter 3 we have presented our theory of modulation analysis and filtering systems. We have kept our discussion of the theory as abstract and general as possible, without an explicit choice for the envelope detector or carrier estimator. In this chapter, we present the envelope detectors and carrier estimators that we have developed.

In section 4.2, we first describe envelope detectors that are based either on the magnitude operator or the Hilbert envelope detector. The characteristic property of these detectors is that their modulators are always non-negative and real. We refer to this type of envelope detectors as “incoherent”. They are the predominant choice in the literature to date (for example [25, 33, 47, 65, 142, 154]), partly because of two key studies on the importance of specific modulation frequencies for the intelligibility of speech that were based on incoherent detectors [26, 27], and partly because they have an energetic interpretation that is easy to understand.

From a signal processing standpoint, however, the incoherent envelope detectors are only meaningful for a small class of signals. Details of the limitations of the incoherent detectors are given in section 4.3.1. To overcome the limitations of the incoherent detectors, we have introduced what we call the class of “coherent” envelope detectors. Characteristically, their modulators are complex valued. Why complex valued modulators are essential for proper modulation filtering of a larger class of signals is motivated in section 4.3.2. We present the coherent envelope detectors that we have developed over the course of our research in section 4.4. Although coherent envelope detectors are just finding their way into the literature (for example [5, 78, 98, 120, 122, 123], interest in them is increasing as more people become aware of the limitations of the incoherent detectors.

At the end of section 4.2 and 4.4, we illustrate modulation frequency analysis of a few
broad signal classes using each of the detectors and estimators discussed in those sections. Section 4.5 and 4.6 conclude this chapter with an analysis of two issues that are common to all modulation analysis and filtering systems: modulation filter effectiveness, and signal reconstruction from a modified time-frequency representation.

4.2 Incoherent envelope detection

Incoherent envelope detectors are based on the Hilbert envelope (for real-valued subbands), or the magnitude operator (for complex-valued subbands). Each type is discussed in detail in section 4.2.1 and 4.2.2.

4.2.1 Hilbert envelope

Hilbert envelope modulation filtering techniques, such as the one proposed by Drullman et al. [27], can be described as follows. Consider a generalized modulation transform (as defined in section 3.9) that is based on a regular filterbank with real-valued subband filters $h_k(t)$. The Hilbert envelope detector $\mathcal{D}_H$ is based on (and named after) the Hilbert transform,

$$\hat{x}(t) = \mathcal{H}\{x(t)\}$$

$$\triangleq \frac{1}{\pi} \int \frac{x(\tau)}{t - \tau} d\tau. \quad (4.1b)$$

The Hilbert transform shifts the frequency content of the signal 90 degrees in phase. The main use of the Hilbert transform in signal processing is to determine the pre-envelope or analytic signal,

$$x^+(t) = x(t) + j\hat{x}(t)$$

$$= x(t) + j\mathcal{H}\{x(t)\}, \quad (4.2b)$$

of the real signal $x(t)$. By definition, the analytic signal $x^+(t)$ is complex valued, and its Fourier transform, $X^+(\Omega)$, is related to the Fourier transform $X(\Omega)$ of the signal $x(t)$.
according to
\[ X^+(\Omega) = \begin{cases} 
2X(\Omega) & \Omega > 0 \\
X(\Omega) & \Omega = 0 \\
0 & \Omega < 0 
\end{cases} \]  
\quad (4.3)

The spectrum of the analytic signal is identical to the spectrum of \( x(t) \) (up to scaling), and vanishes for negative frequencies.

For a real-valued subband \( x_k(t) \), the Hilbert envelope detector is defined as the magnitude of the analytic subband \( x_k^+(t) \), i.e.,
\[ m_k(t) = \mathcal{D}_H\{x_k(t)\} \quad (4.4a) \]
\[ \triangleq |x_k^+(t)| \quad (4.4b) \]
\[ = |x_k(t) + j\mathcal{H}\{x_k(t)\}|. \quad (4.4c) \]

The corresponding Hilbert carrier is defined by
\[ c_k(t) = \mathcal{D}_H^c\{x_k(t)\} \quad (4.5a) \]
\[ \triangleq \cos\{\arg[x_k^+(t)]\} \quad (4.5b) \]
\[ = \cos\{\arg[x_k(t) + j\mathcal{H}\{x_k(t)\}]\}. \quad (4.5c) \]

Instead of applying the Hilbert transform to the input signal \( x(t) \), the Hilbert transform can also be applied to the subband filters \( h_k(t) \) prior to processing. Doing so converts the modulation transform filterbank into an analytic filterbank with analytic subband filters \( h_k^+(t) \). Moreover, it turns the Hilbert envelope detector into the magnitude envelope detector, which we discuss next.

4.2.2 Magnitude detection

A modulation analysis and filtering system based on the magnitude detector can be defined as follows. Consider a generalized modulation transform (as defined in section 3.9) that is based on an analytic filterbank. That is, the frequency response of the subband filters is zero for negative frequencies, such that the subband signals at the output of the filterbank are
analytic (and hence complex valued). An example of an analytic filterbank is the discrete short-time Fourier transform. The magnitude detector $D\parallel$ for an analytic subband signal $x^+_k(t)$ is defined by

$$m_k(t) = D\parallel\{x^+_k(t)\} \triangleq |x^+_k(t)|,$$

(4.6a)
(4.6b)

and the carrier estimator is defined as the complex argument or phase of the subband signal $x^+_k(t)$ as a continuous function of time,

$$c_k(t) = D^c\parallel\{x^+_k(t)\} \triangleq \exp\{j \arg[x^+_k(t)]\}.$$

(4.7a)
(4.7b)

It is easy to verify that the magnitude envelope detector (and therefore the Hilbert envelope detector) satisfies the nonlinearity and projection properties of envelope detectors (see section 3.2.3). It also satisfies frequency shift invariance, since

$$D\parallel\{e^{j\omega_0 t}x(t)\} = |e^{j\omega_0 t}x(t)|$$

(4.8a)
$$= |x(t)|$$

(4.8b)
$$= D\parallel\{x(t)\}.$$ 

(4.8c)

However, it does not satisfy the bandwidth preservation property for arbitrary signals, as we illustrate in section 4.3.

4.2.3 Examples

We illustrate the use of the incoherent modulation transform on three signal classes: amplitude modulated monochromatic carriers, modulated white noise, and speech. Figure 4.1 shows the time-domain signal, the spectrogram and the incoherent modulation spectrum for a signal from each of these classes. The amplitude modulated monochromatic carrier
Figure 4.1: Examples of incoherent modulation frequency analysis for three signal classes. (left) Amplitude modulated monochromatic carriers; (center) Modulated white noise; (right) Speech. Displayed from top to bottom are the signal’s waveform, spectrogram, and incoherent modulation spectrum.
signal has the form

\[ x(t) = \sum_{i=1}^{4} \cos(2\pi \Omega_{c,i} t) \left[ 0.5 + 0.4 \cos(2\pi \Omega_{m,i} t) \right], \quad (4.9) \]

where the carrier and modulator frequencies were arbitrarily selected to be

\{\Omega_{c,i}\} = \{1031, 1344, 2250, 2844\} and \{\Omega_{m,i}\} = \{16.6, 9.8, 19.5, 12.7\}.

In figure 4.1, the presence of amplitude modulation can be detected, but it is difficult to determine the exact carrier and modulator frequencies. In the spectrogram representation, the carrier frequencies can be determined, and the presence of the individual modulators is evident. However, the modulator frequencies remain elusive. In the incoherent modulation spectrum display, both the carrier and modulator frequencies can be easily determined.

The modulated white noise signal has the form

\[ x(t) = w(t) \left[ 0.5 + 0.2 \cos(2\pi 9.8 t) + 0.2 \cos(2\pi 19.5 t) \right], \quad (4.10) \]

where \(w(t)\) is a white noise signal. The incoherent modulation transform is capable of detecting the modulation frequencies on the broadband white noise, as the modulation spectrum representation of the modulated white noise signal in figure 4.1 shows.

The speech signal is a recording of the word “zero”, spoken by a male talker. The waveform and the spectrogram in figure 4.1 show the periodic structure of the voiced parts of the speech signal. Moreover, the spectrogram clearly shows how the formants move in frequency over time. The modulation spectrum of the speech signal shows the amplitude variation of the speech formants as low modulation frequency energy at the formant frequencies.

The incoherent modulation transform can be used in two different ways to analyze signals: narrowband modulation frequency analysis and wideband modulation frequency analysis. Examples of both types of analysis are given in figure 4.2. Narrowband modulation frequency analysis mode uses an analysis window that is long in time, and hence narrow in frequency. A narrow analysis window creates a modulation spectrum that has a high
resolution in acoustic frequency, but limited extent in modulation frequency. It can be used to analyze the lowest modulation frequencies present in a signal. As the narrowband modulation analysis in figure 4.2 shows, the harmonics of the speech signal are resolved in acoustic frequency, and the modulation frequency dimension is limited to 20 Hz. Wideband modulation frequency analysis, on the other hand, uses an analysis window that is short in time and wide in frequency. A wide analysis window creates a modulation spectrum that has low resolution in acoustic frequency, but a large extent in modulation frequency. This type of analysis is not useful for all signals, but in the case of speech it can be used to resolve the fundamental frequency and analyze the formant frequencies in the modulation frequency dimension. As the wideband modulation analysis in figure 4.2 shows, the fundamental frequency of this talker varies between 105 and 115 Hz and the energy of the formants is strongest at 600 and 1300 Hz.

The incoherent modulation transform can also be used to effortlessly reduce or remove cyclostationary noise, such as fan noise and machine noise, in a signal. Figure 4.3 illustrates this for a speech signal that is corrupted by engine noise. The incoherent modulation spectrogram representation of the signal shows a distinct strip of energy around a modulation frequency of 26 Hz, and to a lesser degree at the first modulation harmonic at 52 Hz. The engine noise modulation frequency content is clearly separated from the modulation content.
of the speech signal between 0 and 10 Hz. This separation can be exploited to reduce the engine noise without affecting the speech signal by applying a modulation notch filter to the signal. The output of the modulation notch filter in the right panel of figure 4.3 shows that the modulated or component of the engine noise has been effectively removed. What remains of the engine noise is its stationary component at 0 modulation frequency, which can be further reduced with a conventional noise suppression technique.

4.3 **Incoherent vs. coherent detection**

4.3.1 **Limitations of incoherent envelope detectors**

Incoherent envelope detectors have three important limitations: (1) the subband magnitude and subband phase signals generally exceed the bandwidth of the subband signal; (2) incoherent detectors force a conjugate symmetric spectrum on the modulator, which is an unrealistic assumption for most natural signals; (3) the modulator domain of incoherent detectors is not closed under convolution.

That incoherent envelope detectors do not preserve the subband bandwidth was shown by Dugundji [30], as follows. Let \( x(t) \) be an narrowband analytic signal with bandwidth \( B \), i.e.,

\[
X(\Omega) = 0 \text{ for } |\Omega - \Omega_0| \geq \frac{1}{2}B.
\]  

(4.11)
Consider the spectrum of the square of the envelope, i.e.,

\[ K(\Omega) = \int |x(t)|^2 e^{-j\Omega t} dt \]  \hspace{1cm} (4.12a)

\[ = \int x(t)x^*(t)e^{-j\Omega t} dt \]  \hspace{1cm} (4.12b)

\[ = X(\Omega) * X^*(-\Omega) \]  \hspace{1cm} (4.12c)

\[ = \int_{-\frac{1}{2}B}^{B} X(\xi)X^*(\xi - \Omega)d\xi. \]  \hspace{1cm} (4.12d)

Equation (4.12c) shows that the square of the magnitude envelope of a signal \( x(t) \) can be considered to be the signal demodulated in frequency by itself. It readily follows from equation (4.11) and (4.12d) that the spectrum of the square of the envelope \( K(\Omega) \) vanishes outside \(|\Omega| \leq B\). Therefore, the bandwidth of the square of the magnitude envelope or the Hilbert envelope is of the same order of the subband. However, as Dugundji remarks: “...(these hypotheses) do not allow one to draw the conclusion that the envelope itself is bandlimited; indeed, there seems to be no physical reason that it should be so.”

To illustrate this limitation of the magnitude detector, consider the signal

\[ x(t) = e^{j\Omega_c t} \cos(\Omega_m t), \quad \Omega_c > \Omega_m. \]  \hspace{1cm} (4.13)

Clearly, \( X(\Omega) = 0 \) for \( |\Omega - \Omega_c| > \Omega_m \). Furthermore, its envelope squared is given by

\[ k(t) = |x(t)|^2 = x(t)x^*(t) \]  \hspace{1cm} (4.14a)

\[ = [e^{j\Omega_c t}\cos(\Omega_m t)][e^{j\Omega_c t}\cos(\Omega_m t)]^* \]  \hspace{1cm} (4.14b)

\[ = \cos^2(\Omega_m t) \]  \hspace{1cm} (4.14c)

\[ = \frac{1}{2}\cos(2\Omega_m t) + \frac{1}{2}, \]  \hspace{1cm} (4.14d)

hence \( K(\Omega) = 0 \) for \(|\Omega| > 2\Omega_m\). However, the envelope of the signal is given by

\[ m(t) = |x(t)| \]  \hspace{1cm} (4.15a)

\[ = |e^{j\Omega_c t}\cos(\Omega_m t)| \]  \hspace{1cm} (4.15b)
\[ = |\cos(\Omega_m t)|, \quad (4.15c) \]

which has discontinuities at \( t = \frac{1}{2} \frac{\pi}{\Omega_m} + k \frac{\pi}{\Omega_m} \) for integer \( k \). Its spectrum is given by

\[
M(\Omega) = \sum_k \frac{4}{-(-1)^k ((2k)^2 - 1)} \delta(\Omega - 2k\Omega_m). \tag{4.16} \]

The discontinuities in the envelope introduce harmonics in the envelope spectrum at twice the modulator frequency \( \Omega_m \). The amplitude of the harmonics falls off at a rate of \( \Omega^{-2} \).

For the signal \( x(t) \), the magnitude envelope \( m(t) \) is evidently not limited to the signal’s bandwidth \( B = 2\Omega_m \).

There are signals whose magnitude envelope is bandlimited to their bandwidth. A sufficient condition on a signal to have a bandlimited envelope is that the signal \( x(t) \) should be the product of a monochromatic carrier \( c(t) = e^{j\Omega_0 t} \) and a bandlimited, non-negative real envelope \( a(t) = |a(t)|, A(\Omega) = 0 \) for \( |\Omega| > B_a \). It is easy to verify that the bandwidth of the modulator \( B_m = B_a \) in that case:

\[
x(t) = a(t)c(t) = e^{j\Omega_0 t}a(t) \tag{4.17}
\]

\[
m(t) = |x(t)| = |e^{j\Omega_0 t}a(t)| = |a(t)| = a(t) \tag{4.18}
\]

\[
\Rightarrow M(\Omega) = A(\Omega) \tag{4.19}
\]

\[
\Rightarrow M(\Omega) = 0 \text{ for } |\Omega| > B_a. \tag{4.20}
\]

From empirical observation, we postulate that it is also a necessary condition that a signal is the product of a monochromatic carrier and a bandlimited, non-negative real envelope for the bandwidth of the magnitude envelope to be bandlimited to the bandwidth of the subband. Hence, a signal has a bandlimited envelope if and only if it is of the prescribed form.

If the modulator of a bandlimited signal is not bandlimited, then its carrier must also be not bandlimited for their product to be bandlimited. For example, for the signal in equation (4.13) and its envelope spectrum in equation (4.16), we find that the carrier spectrum
When convolved, the infinite number of terms in the modulator and carrier spectrum exactly cancel each other in their convolution to yield the bandlimited signal \(x(t)\). However, when the modulator is filtered this cancelation will be imperfect, because it depends on the presence of all the terms in the modulator in unmodified form.

The excessive bandwidth of an incoherent modulator manifests itself as broadband modulation frequency artifacts during incoherent modulation frequency analysis. Furthermore, uncanceled terms of the carrier after modulation filtering cause audible distortion in modulation filtered signals, and reduce the effectiveness of a modulation filter.

The second limitation of incoherent detectors is that the spectrum of the magnitude envelope is conjugate symmetric, since the magnitude envelope is real valued. A conjugate symmetric envelope is a valid assumption for a subband signal that can be modeled as a monochromatic carrier and a non-negative real modulator. However, as Atlas, Li and Thompson [5] argued, it is unrealistic to assume that natural signals such as speech and music exhibit subbands that are conjugate symmetric with respect to some fixed frequency \(\Omega_{\text{sym}}\). It requires a high degree of spectral symmetry in the signal, which additionally must be synchronous with the location and shape of the filterbank’s subband filters. We have only observed conjugate symmetric subband spectra in synthetic signals; all subband spectra of natural signals that we have analyzed were asymmetric.

Finally, the third limitation is that the modulator domain of incoherent detectors is not closed under convolution. By definition, an incoherent envelope is non-negative and real. However, after filtering the modulator with an arbitrary modulation filter, \(\tilde{m}(t) = m(t) * g(t)\), the modified modulator \(\tilde{m}(t)\) is no longer guaranteed to be non-negative and real. For example, it is very likely that \(\tilde{m}(t) < 0\) for some \(t\) when \(g(t)\) is a high-pass modulation filter that attenuates or removes the DC component of \(m(t)\). Since the modified modulator \(\tilde{m}(t)\) can be negative for some values of \(t\), it is no longer a true modulator with respect to an incoherent envelope detector, which yields modulators that are non-negative for all \(t\). The modified modulator is therefore not guaranteed to belong to the modulator domain after
convolution, hence we say that the modulator domain is not closed under convolution.

4.3.2 Motivation for complex-valued modulators

How can the problems associated with incoherent envelope detectors be remedied? To answer this question, we must revisit the product model for carriers and modulators of equation (3.14),

\[ x(t) = m(t)c(t), \]

(4.22)

or in polar form,

\[ a_x(t)e^{j\phi_x(t)} = \left[ a_m(t)e^{j\phi_m(t)} \right] \left[ a_c(t)e^{j\phi_c(t)} \right]. \]

(4.23)

The polar notation allows us to cast the envelope detection or carrier estimation problem in a multiplicative amplitude decomposition and an additive phase decomposition problem,

\[ a_x(t) = a_m(t)a_c(t) \]  \hspace{1cm} (4.24a)

\[ \phi_x(t) = \phi_m(t) + \phi_c(t). \]  \hspace{1cm} (4.24b)

The objective of an envelope detector is to determine \( a_m(t) \) and \( \phi_m(t) \) given \( a_x(t) \) and \( \phi_x(t) \). Without additional restrictions, the amplitude and phase decompositions are obviously ambiguous (see also [15]). Incoherent detectors resolve this ambiguity by making the additional assumption that \( \phi_m(t) = 0 \) and \( a_c(t) = 1 \) for all \( t \), such that \( a_m(t) = a_x(t) \) and \( \phi_c(t) = \phi_x(t) \). However, this assumption causes the problems associated with incoherent detectors that we discussed in section 4.3.1.

The behavior of the envelope detector improves if we assume instead that \( \phi_m(t) \neq 0 \) for all \( t \), which means that the modulator is complex-valued. This assumption avoids that the envelope detector forces a symmetric spectrum onto the modulator. It also guarantees that the modulator domain is closed under convolution, since the space of complex signals is closed under convolution. However, the assumption \( \phi_m(t) \neq 0 \) for all \( t \) also renders the phase decomposition of equation (4.24b) ambiguous. We propose to solve this ambiguity by imposing a bandwidth constraint on the carrier. We assign only those phase variations to the carrier phase \( \phi_c(t) \) that are slow relative to the signal’s general phase trend. The slowly
varying phase $\phi_c(t)$, combined with a unit magnitude assumption $a_c(t) = 1$, ensures that the bandwidth of the estimated carrier and modulator are of the same order as the bandwidth of the signal, thereby satisfying the bandwidth preservation requirement of envelope detectors.

Given a signal’s phase signal $\phi_x(t)$, an appropriate slowly varying carrier phase signal $\phi_c(t)$ can be estimated by lowpass filtering an estimate of the signal’s instantaneous frequency,

$$\alpha_c(t) = \text{IF}\{\phi_x(t)\} \ast h_{lp}(t), \quad (4.25)$$

which is defined as the derivative of the signal’s phase,

$$\text{IF}\{\phi_x(t)\} = \frac{d\phi_x(t)}{dt}, \quad (4.26)$$

and subsequently integrating the instantaneous frequency signal $\alpha_c(t)$ over time,

$$\phi_c(t) = \int_0^t \alpha_c(\tau) d\tau. \quad (4.27)$$

The carrier signal $c(t)$ is then given by

$$c(t) = e^{j\phi_c(t)}. \quad (4.28)$$

Since the modulator signal $m(t)$ is defined using coherent demodulation,

$$m(t) = \frac{x(t)}{e^{j\phi_c(t)}} = \frac{x(t)}{e^{j\phi_c(t)}} = x(t)e^{-j\phi_c(t)}, \quad (4.29)$$

we refer to this type of envelope detectors and carrier estimators as coherent envelope detectors and coherent carrier estimators.

### 4.4 Coherent carrier estimation

In this section we introduce three coherent carrier estimators that we have developed in the course of our research into distortion-free modulation analysis and filtering systems. Section 4.4.1 describes the smoothed Hilbert carrier estimator, which defines a signal’s carrier as a bandlimited version of its Hilbert carrier. In section 4.4.2 we discuss a carrier
estimator based on an instantaneous frequency estimator that was originally proposed by Atlas and Janssen [4], and that we adopted in our research. We subsequently improved this carrier estimator by replacing the instantaneous frequency estimator with the frequency reassignment operator [6], which we describe in section 4.4.4. We illustrate the use of the coherent detectors with some examples in section 4.4.5.

4.4.1 Smoothed Hilbert carrier estimator ¹

Consider an analytic subband signal in polar form, \( x_k^+(t) = a_k(t)e^{j\phi_k(t)} \). The subband’s phase signal, \( \phi_k(t) \), can be expressed as the sum of a linear phase term \( \Omega_{k,m}t \), an initial phase \( \phi_{k,0} \) and a phase deviation term \( \theta_k(t) \), as follows

\[
\phi_k(t) = \Omega_{k,m}t + \phi_{k,0} + \theta_k(t). \tag{4.30}
\]

The parameters \( \Omega_{k,m} \) and \( \phi_{k,0} \) can be chosen such that the phase deviation term \( \theta_k(t) \) has zero mean. It is easy to verify that for every phase signal \( \phi_k(t) \) there exists a unique pair \( (\Omega_{k,m}, \phi_{k,0}) \) that satisfies this condition. The frequency \( \Omega_{k,m} \) is in a sense the “average” frequency of the subband \( x_k(t) \). We refer to this frequency as the midband frequency, a concept and term first used by Rice [112, p. 81] to describe a specific frequency in a subband that is not necessarily at the center of the subband.

We propose a carrier estimator that defines a subband’s carrier to be a smoothed version of equation (4.30), as follows

\[
\tilde{\phi}_k(t) = \Omega_{k,m}t + \phi_{k,0} + [\theta_k(t) \ast h_{lp}(t)]. \tag{4.31}
\]

Given the smoothed subband phase signal \( \tilde{\phi}_k(t) \), the smoothed Hilbert carrier estimator is defined by

\[
c_k(t) = D^c_{SH}\{x_k(t)\} \tag{4.32a}
\]

\[ \triangleq e^{j\tilde{\phi}_k(t)}, \quad (4.32b) \]

and the coherent envelope by

\[ m_k(t) = \mathcal{D}_{\text{SH}} \{ x_k(t) \} \quad (4.33a) \]
\[ \triangleq x_k^+(t)e^{-j\tilde{\phi}_k(t)}. \quad (4.33b) \]

By smoothing (i.e. band limiting) the phase signal \( \theta_k(t) \), the slow rate of change of the carrier estimate can be controlled. With no smoothing of \( \theta_k(t) \), i.e., \( h_{lp}(t) = \delta(t) \), the detected envelope equals the magnitude or Hilbert envelope, and with full smoothing of \( \theta_k(t) \), i.e., \( h_{lp}(t) = 1 \), the detector returns the complex envelope corresponding to the monochromatic carrier at the midband frequency.

The smoothed Hilbert carrier estimator can be viewed as a rudimentary instantaneous frequency estimator. Instantaneous frequency is defined as the derivative of instantaneous phase, \( \Omega(t) = \frac{d\phi(t)}{dt} \). By subtracting the midband frequency from the instantaneous phase, and band-limiting the result, the smoothed Hilbert carrier estimator effectively limits the instantaneous frequency to a small frequency range around the midband frequency.

### 4.4.2 Instantaneous frequency carrier estimator

A coherent carrier estimator that is based explicitly on an instantaneous frequency (IF) estimator was proposed by Atlas and Janssen [4]. The IF estimator in their carrier estimator is a substantial modification of the differential FM detector by Glas [36]. They modified this detector to obtain a carrier estimator that is asymptotically unbiased and insensitive to amplitude modulations. Given a signal \( x(t) \) and its discrete short-time Fourier transform, \( X_k(m) \), the IF carrier estimator estimates a subband’s carrier as follows. Let

\[ I_k(m) = \text{Re} \{ X_k(m) \} \quad (4.34) \]

and

\[ Q_k(m) = \text{Im} \{ X_k(m) \} \quad (4.35) \]
be the real and imaginary part of the complex-valued subband signals. The estimator derives an unnormalized IF estimate,

$$Z_k(m) = Z_{i,k}(m) + jZ_{q,k}(m), \quad (4.36)$$

from the real and imaginary parts of a subband. The in-phase part is defined by

$$Z_{i,k}(m) = I_k(m-1)I_k(m+1) + Q_k(m-1)Q_k(m+1), \quad (4.37)$$

and the quadrature part is defined by

$$Z_{q,k}(m) = I_k(m-1)Q_k(m+1) - Q_k(m-1)I_k(m+1). \quad (4.38)$$

A unimodular phase-only IF estimate $\alpha_k(m)$ is obtained from $Z_k(m)$ by normalizing $Z_k(m)$ by its magnitude,

$$\alpha_k(m) = \begin{cases} \left(\frac{Z_k(m)}{|Z_k(m)|}\right)^{\frac{1}{2}} & |Z_k(m)| > \epsilon \\ \alpha_k(m-1) & |Z_k(m)| \leq \epsilon \end{cases}, \quad (4.39)$$

for arbitrarily small $\epsilon$. To avoid numerical precision problems, the subband’s previous IF estimate is used for small magnitudes of $Z_k(m)$. The IF estimate $\alpha_k(m)$ is smoothed with a lowpass filter to reduce its variance. An instantaneous phase estimate is recursively constructed from the smoothed IF estimate according to

$$W_k(-1) = 1, W_k(0) = \alpha_k(0) \quad (4.40a)$$

$$W_k(m) = W_k(m-1)\alpha_k(m). \quad (4.40b)$$

Given the instantaneous phase estimate $W_k(m)$ of a subband $X_k(m)$, the IF carrier estimator $\mathfrak{D}_{IF}^c$ is defined by

$$C_k(m) = \mathfrak{D}_{IF}^c\{X_k(m)\} \quad (4.41a)$$

$$\triangleq W_k(m), \quad (4.41b)$$
and the coherent envelope is found by coherent demodulation of the subband signal,

\[ M_k(m) = \mathcal{D}_{\text{IF}}\{X_k(m)\} \]
\[ \triangleq X_k(m)C_k^*(m). \]  

(4.42a)  

(4.42b)

### 4.4.3 Analysis of the instantaneous frequency carrier estimator

The instantaneous frequency (IF) estimator that is used in the carrier estimator described in section 4.4.2 is equivalent to the first difference of subband phase, which can be shown as follows. If we express the complex-valued discrete short-time Fourier transform subbands \( X_k(m) \) in polar form, i.e.,

\[ X_k(m) = A_k(m)e^{j\phi_k(m)}, \]  

(4.43)

then we can rewrite (4.34) and (4.35) as

\[ I_k(m) = \text{Re}\{X_k(m)\} = A_k(m)\cos[\phi_k(m)] \]  

(4.44)

and

\[ Q_k(m) = \text{Im}\{X_k(m)\} = A_k(m)\sin[\phi_k(m)]. \]  

(4.45)

Plugging these values for \( I_k(m) \) and \( Q_k(m) \) in (4.37) and (4.38) and applying standard trigonometric identities yields

\[ Z_{q,k}(m) = I_k(m-1)Q_k(m+1) - Q_k(m-1)I_k(m+1) \]
\[ = 2A_k(m-1)A_k(m+1)\{\sin[\phi_k(m+1) - \phi_k(m-1)]\} \]  

(4.46a)  

(4.46b)

and

\[ Z_{t,k}(m) = I_k(m-1)I_k(m+1) + Q_k(m-1)Q_k(m+1) \]
\[ = 2A_k(m-1)A_k(m+1)\{\cos[\phi_k(m+1) - \phi_k(m-1)]\}. \]  

(4.47a)  

(4.47b)

---

The unnormalized IF estimate $Z_k(m)$ can then be expressed as

$$Z_k(m) = Z_{i,k}(m) + jZ_{q,k}(m)$$

$$= A_k(m - 1)A_k(m + 1) \exp\{j[\phi_k(m + 1) - \phi_k(m - 1)]\},$$

such that the phase-only IF estimate $\alpha_k(m)$ can be written as

$$\alpha_k(m) = \left(\frac{Z_k(m)}{|Z_k(m)|}\right)^{\frac{1}{2}}$$

$$= \exp\{j\frac{1}{2}[\phi_k(m + 1) - \phi_k(m - 1)]\}.$$  

In other words, the IF estimator defined in [4] is based on the finite central phase difference at each point in the discrete short-time Fourier transform. However, as Kodera et al. [69] showed, the true instantaneous frequency of a subband is defined by the time derivative of the subband phase, i.e.,

$$\alpha_k(t) = \frac{d\phi_k(t)}{dt}.$$  

Therefore, the IF estimator in equation (4.49) is a linear approximation to the true subband instantaneous frequency. The accuracy of the linear approximation decreases with increasing decimation factor of the discrete short-time Fourier transform. In addition, it also requires special care to unwrap the STFT phase.

### 4.4.4 Frequency reassignment carrier estimator

The linear approximation and phase unwrapping of the IF carrier estimator can be avoided by using the frequency reassignment operator from time-frequency reassignment [6]. This operator computes the true derivative of subband phase without resorting to finite approximations, and it avoids the phase unwrapping problem.

The definition of the frequency reassignment operator is based on the continuous-time moving window transform. The definition of the discrete-time moving window transform

---

Footnote: 3

was given in equation (3.81). The continuous-time moving window transform is defined as

$$X(\tau, \Omega) = \text{MWT}\{x(t)\}$$

$$\triangleq \mathcal{F}\{x(t + \tau)w(-t)\}$$

$$= \int x(t + \tau)w(-t)e^{-j\Omega t}dt,$$

where $w(t)$ is an analysis window. A time-domain signal $x(t)$ can be reconstructed from the coefficients $X(\tau, \Omega) = A(\tau, \Omega)e^{j\psi(\tau, \Omega)}$ by

$$x(t) = \int\int X(\tau, \Omega)h(\tau - t)e^{-j\Omega[\tau - t]}d\Omega d\tau$$

$$= \int\int A(\tau, \Omega)h(\tau - t)e^{j[\psi(\tau, \Omega)-\Omega\tau+\Omega t]}d\Omega d\tau,$$

where $h(t)$ is a (real-valued) lowpass kernel function, or analysis window. For signals having magnitude spectra $A(\tau, \Omega)$, whose time variation is slow relative to the phase variation, the maximum contribution to the reconstruction integral comes from the vicinity of the point $\tau, \Omega$ satisfying the phase stationarity condition

$$\frac{\partial}{\partial \Omega}[\psi(\tau, \Omega) - \Omega \tau + \Omega t] = 0$$

$$\frac{\partial}{\partial \tau}[\psi(\tau, \Omega) - \Omega \tau + \Omega t] = 0,$$

or equivalently, around the point $\hat{\tau}, \hat{\Omega}$ defined by

$$\hat{\tau}(\tau, \Omega) = \tau - \frac{\partial \psi(\tau, \Omega)}{\partial \Omega}$$

$$\hat{\Omega}(\tau, \Omega) = \frac{\partial \psi(\tau, \Omega)}{\partial \tau}.$$

This phenomenon has long been known in such fields as optics as the principle of stationary phase (see for example [101]). The method of time-frequency reassignment changes, or reassigns, the point of attribution of $X(\tau, \Omega)$ from $(\tau, \Omega)$ to this point of maximum contribution $(\hat{\tau}(\tau, \Omega), \hat{\Omega}(\tau, \Omega))$ sometimes called the “center of gravity” of the distribution, by way of analogy to a mass distribution.
The reassigned coordinates in (4.55) and (4.56) can be computed by

\[ 
\hat{\tau}(\tau, \Omega) = \tau - \text{Re} \left\{ \frac{X_{\text{Th}}(\tau, \Omega) \cdot X^*(\tau, \Omega)}{|X(\tau, \Omega)|^2} \right\}, 
\]

\[ 
\hat{\Omega}(\tau, \Omega) = \Omega + \text{Im} \left\{ \frac{X_{\text{Dh}}(\tau, \Omega) \cdot X^*(\tau, \Omega)}{|X(\tau, \Omega)|^2} \right\}, 
\]

where \( X(\tau, \Omega), X_{\text{Th}}(\tau, \Omega) \) and \( X_{\text{Dh}}(\tau, \Omega) \) are short-time Fourier transforms of the signal \( x(t) \) that are computed using an analysis window \( h(t) \), a time-weighted analysis window \( h_T(t) = t \cdot h(t) \), and a time-derivative analysis window \( h_D(t) = \frac{d}{dt} h(t) \), respectively.

Using the auxiliary window functions \( h_T(t) \) and \( h_D(t) \), the reassignment operations can be computed at any time-frequency coordinate \( \tau, \Omega \) from an algebraic combination of the values of three Fourier transforms evaluated at \( \tau, \Omega \), without directly evaluating or approximating the partial derivatives of phase. Since this algorithm operates only on spectral data evaluated at a single time and frequency, and does not explicitly compute any derivatives, it can easily be implemented in digital systems using discrete times and frequencies.

The time-weighted window function, \( h_T(t) \), is trivially computed by point-wise multiplication of the original window function \( h(t) \) by a time ramp; in discrete time, \( h_T(n) = nh(n) \). If the derivative of the window function is unknown, then \( h_D(t) \) can also be computed numerically. The derivative theorem for Fourier transforms states that if \( X(\Omega) = \mathcal{F}\{x(t)\} \) then \( j\Omega X(\Omega) = \mathcal{F}\{\frac{d}{dt} x(t)\} \). We can therefore construct the time-derivative window used in the evaluation of the frequency reassignment operator by computing the Fourier transform of \( h(t) \), multiplying by \( j\Omega \), and inverting the Fourier transform. That is,

\[ 
\frac{d}{dt} h(t) = \mathcal{F}^{-1}\{j\Omega H(\Omega)\} = \text{Im} \left\{ \mathcal{F}^{-1}\{\Omega H(\Omega)\} \right\}, 
\]

and so, in discrete time,

\[ 
h_D(n) = -\text{Im} \left\{ \mathcal{F}^{-1}\left\{ \frac{2\pi k}{N} H(k) \right\} \right\}. 
\]

An example of the auxiliary short-time analysis windows employed in our computation of
Figure 4.4: Representative analysis windows employed in the three short-time transforms used to compute reassigned times and frequencies. (left) The original window function $h(n)$ (a Kaiser window with shaping parameter $\beta = 9$); (center) The time-weighted window function $h_T(n)$; (right) The time-derivative window function $h_D(n)$.

The reassignment operations is shown in figure 4.4. To adapt the frequency reassignment operator for use as a carrier estimator, it is converted to a carrier signal, such that

$$C_\Omega(\tau) = \mathcal{D}_{FR}\{X_\Omega(\tau)\}$$

$$\triangleq \exp \left[ j \int_0^\tau \tilde{\Omega}(\xi, \Omega) d\xi \right], \quad (4.61a)$$

and

$$M_\Omega(\tau) = \mathcal{D}_{FR}\{X_\Omega(\tau)\}$$

$$\triangleq X_\Omega(\tau) C^*_\Omega(\tau). \quad (4.62b)$$

4.4.5 Examples

Figure 4.5 illustrates the use of each coherent modulation transform on the same signals that were used in the demonstration of the incoherent transforms in section 4.2.3. For simplicity, the waveforms and spectrograms of the signals are not repeated; they are the same as in figure 4.1. The most striking difference between the incoherent and coherent modulation analysis examples is that the coherent modulation spectrum is shown for both positive and negative modulation frequencies, whereas the incoherent modulation spectrum is shown.
Figure 4.5: Examples of coherent modulation frequency analysis, using each of the coherent modulation analysis techniques described in section 4.4, for three signal classes. (left) Amplitude modulated monochromatic carriers; (center) Modulated white noise; (right) Speech.
for only positive modulation frequencies. Showing only positive frequencies for incoherent modulators suffices, because incoherent modulators are real-valued and have conjugate symmetric spectra. Coherent modulators, however, are complex-valued and have no symmetry in their spectrum. Therefore, both positive and negative modulation frequencies must be shown.

A prominent feature of the modulation spectra in figure 4.5 is that none of the coherent detectors is able to properly estimate the modulators of the modulated white noise signal. This is a direct consequence of the noisy subband phase of the modulated white noise signal. The coherent detectors attempt to estimate a smoothly varying carrier in the noisy subband phase, but fail to detect one because there is no structure in the subband phase.

The instantaneous frequency carrier estimator is imprecise on the amplitude modulated monochromatic carrier signal, especially in subbands that are off-center from a modulated carrier. In those subbands, the IF carrier estimator is biased towards the upper or lower modulator sideband, instead of correctly estimating the carrier frequency. The smoothed phase and frequency reassignment carrier estimators appear to be far less sensitive to off-center biases. Finally, figure 4.5 also shows that the frequency reassignment carrier estimator performs the best on the speech signal. It is able to estimate the carrier frequencies with high accuracy, and produces modulators that are very well limited to ±10 Hz in modulation frequency.

In summary, there are no great differences between the coherent carrier estimators defined in section 4.4. The smoothed Hilbert carrier estimator and frequency reassignment estimator seem to be more accurate than the instantaneous frequency carrier estimator in most cases. The coherent detectors perform best on natural signals such as speech and music. For example, both the instantaneous frequency carrier estimator and the frequency reassignment estimator have been successfully applied to a musical instrument separation task [4, 122]. However, comparing figures 4.1 and 4.5 shows that incoherent detectors perform superior on artificial signals with strictly positive envelopes, such as the amplitude modulated monochromatic carrier or the modulated white noise.

A full comparison of the incoherent and coherent modulation transforms defined in this chapter is presented in section 4.6. There, the performance of each modulation transform
under two signal reconstruction techniques is evaluated on several criteria. However, we first define a method to measure the effectiveness of a modulation filter, which is one of the performance criteria.

4.5 Modulation filter effectiveness

We propose a method to measure the effectiveness of modulation filtering approaches with respect to their filtering performance. This method was developed to help quantify the discussion in literature about modulation filter effectiveness (e.g., [26, 27, 35, 120, 121]), and as a performance benchmark for modulation transforms.

Our method to measure modulation filtering effectiveness is similar to Viemeister’s temporal modulation transfer function [152] and the modulation transfer functions of Houtgast and Steeneken [54] and Schroeder [129]. We determine the incoherent short-time modulation spectrogram of a test signal before and after modulation filtering, and compute their ratio. We then average the resulting modulation frequency response across acoustic frequency, to summarize it into a single curve describing the filter’s frequency response as a function of modulation frequency. To make the result less dependent on a particular test signal, we additionally average the modulation frequency responses of a set of \( N \) test signals, as well as over time. For a formal definition of our technique, let \( X_l(i, k) \) denote the (incoherent) discrete short-time modulation transform (DSTMT) of a signal \( x(n) \), according to definition 3.7.4, and let \( \tilde{X}_l(i, k) \) denote the DSTMT of the modulation filtered signal \( \tilde{x}(n) \). Given these definitions and the description above, our method to measure modulation filtering effectiveness can be expressed as

\[
X_{\text{eff}}(i) = \frac{1}{NK} \sum_{n=1}^{N} \sum_{l} \sum_{k=0}^{K-1} \frac{|\tilde{X}_l(i, k)|}{|X_l(i, k)|}.
\] (4.63)

Figure 4.6 compares the effective modulation frequency response of the modulation filtering systems based on the envelope/carrier detectors discussed in sections 4.2 and 4.4.

Figure 4.6: Effective modulation frequency response of a severe lowpass modulation filter (1 Hz cutoff) for the four modulation transforms discussed in section 4.2 and 4.4 (the number in brackets indicates the lowpass cutoff frequency for the carrier smoother employed in the method): • Incoherent; □ SH[9]; ▽ SH[4]. ◇ SH[0]. + IF[9]; × IF[4]; ■ FR[9]; ● FR[9].

Several properties of the effective responses in figure 4.6 are worth mentioning. The peak around 39 Hz in the ideal response and in the coherent and smoothed carrier responses is caused by the spacing of the subbands in the modulation filtering system that was used. Such peaks are unavoidable in modulation transforms with uniformly spaced subbands. Furthermore, all effective responses have the tendency to follow the ideal response until they reach their maximum amount of suppression at a modulation frequency of around 5-10 Hz. The maximum amount of suppression that the incoherent modulation transform achieves is fixed (at approximately -15 dB), but can be controlled for the other methods by varying the cut-off frequency of the IF or carrier lowpass filter. We also observe that modulation transforms based on the IF carrier estimator and the smoothed Hilbert carrier estimator have a similar effective modulation frequency response for equal amounts of IF and phase smoothing. However, none of the modulation filtering systems achieve the suppression of the design specifications. The smoothed phase carrier approach with 0 Hz smoothing performs the best in this respect, but the resulting signals contain a distinctly audible “buzz”, which lowers the attractiveness of this technique.
4.6 Signal reconstruction

In section 3.5.1 we introduced the concept that a time-frequency function \( X(m, k) \) is a “valid short-time Fourier transform” if and only if there exists a time-domain signal \( x(n) \) for which \( \text{DSTFT}\{x(n)\} = X(m, k) \). We also argued that filtering the modulators \( M(m, k) \) of a valid STFT \( X(m, k) \), and recombining them with their carriers,

\[
\tilde{X}(m, k) = [M(m, k) * g_k(m)]C(m, k),
\]

(4.64)
does not guarantee that the modulation filtered STFT \( \tilde{X}(m, k) \) is a valid STFT. The reason for this is that a signal’s STFT, when not critically sampled, is a redundant representation of that signal. Arbitrary modification of the STFT does not preserve that redundancy, and generally yields a STFT for which there does not exist a time-domain signal that has that STFT. In section 3.5.5 we showed that a modulation filtered time-domain signal \( \tilde{x}(n) \) can be reconstructed from \( \tilde{X}(m, k) \) even when it is invalid, because the inverse STFT acts as a pseudo-inverse on invalid data. However, in that case the STFT of the modulation filtered signal is different from the desired modulation filtered STFT, i.e., \( \text{DSTFT}\{\tilde{x}(n)\} \neq \tilde{X}(m, k) \).

In section 3.9 we extended the concepts of valid time-frequency functions and pseudo-inverses to the generalized modulation framework.

In this section, we evaluate the performance of the modulation transforms defined in 4.2 and 4.4 under two reconstruction techniques on two criteria: (1) reconstruction error, measured with a signal-to-error ratio; and (2) effective modulation frequency response. Furthermore, we evaluate each technique in an informal subjective listening test. This section is organized as follows. In section 4.6.1, we review the signal reconstruction techniques that we consider in the evaluation. In section 4.6.2 we describe the signals and the evaluation procedure. In section 4.6.3 we present the results, followed by conclusions and a discussion in section 4.6.4.

---

4.6.1 Reconstruction methods

A classic solution to signal reconstruction from the MSTFT is offered by Griffin and Lim [39]. They propose a method of reconstruction that estimates a time-domain signal with an STFT closest to the desired STFT in a least squared-error (LSE) sense. They refer to this estimator as the LSEE-MSTFT. Under some mild restrictions on the synthesis window, the LSEE-MSTFT technique reduces to the computationally more efficient weighted overlap-add (WOLA) reconstruction technique [16].

The LSEE-MSTFT or WOLA technique reconstructs a time-domain signal from the complex-valued STFT \( X(m,k) \). In the case of incoherent modulation transforms, however, we are mostly interested in reconstructing a time-domain signal that has the desired STFT magnitude (STFTM), i.e. \( |\text{DSTFT}\{\tilde{x}(n)\}| = |\tilde{X}(m,k)| \), because incoherent modulation filtering leaves the STFT phase (STFTP) untouched. There exists an alternative method of reconstruction from modified STFT magnitude (MSTFTM) alone. This technique, also defined in [39], is referred to as LSEE-MSTFTM. It is an iterative procedure that estimates a time-domain signal that has an STFTM that is closest to the desired MSTFTM in a least squared-error sense.

The original definitions of the LSEE-MSTFT and LSEE-MSTFTM methods from [39], which we repeat next for convenience, are based on the continuous Fourier transform. Shpiro and Malah [132], however, showed that those definitions can be extended to the discrete Fourier transform. Therefore, we present definitions of the reconstruction techniques here that are based on the discrete STFT.

**LSEE-MSTFT or weighted overlap-add (WOLA)**

To estimate a signal \( x(n) \) whose STFT \( X(m,k) \) is closest to a given modified STFT \( Y(m,k) \) in the squared error sense, consider the distance measure

\[
D[x(n), Y(m,k)] = \frac{1}{K} \sum_{m} \sum_{k=0}^{K-1} |X(m,k) - Y(m,k)|^2.
\] (4.65)
Furthermore, let \( y(m, n) \) be given by
\[
y(m, n) = \frac{1}{K} \sum_{k=0}^{K-1} Y(m, k)e^{j2\pi kn/K}.
\] (4.66)

Griffin and Lim [39] and Shpiro and Malah [132] show that the following signal reconstruction from the MSTFT minimizes the distance measure \( D[x(n), Y(m, k)] \):
\[
x(n) = \sum_m w(mR - n)y(m, n) \sum_m w^2(mR - n). \] (4.67)

This solution coincides with weighted overlap-add (WOLA) reconstruction,
\[
x(n) = \sum_m f(mR - n)y(m, n),
\] (4.68)
when the WOLA synthesis window \( f(n) \) is equal to the STFT analysis window \( w(n) \), i.e. \( f(n) = w(n) \), and has the property
\[
\sum_m f^2(mR - n) = 1.
\] (4.69)

**LSEE-MSTFTM**

Griffin and Lim [39] also define an iterative procedure called LSEE-MSTFTM that estimates a signal from a MSTFTM that is closest to the MSTFTM according to the least squared error distance measure
\[
D_M[x(n),|Y(m, k)|| = \frac{1}{K} \sum_m \sum_{k=0}^{K-1} |X(m, k)| - |Y(m, k)|^2.
\] (4.70)

The LSEE-MSTFTM algorithm works as follows. Let \( x^i(n) \) denote the estimated signal \( x(n) \) after the \( i \)th iteration, and let \( X^i(m, k) \) denote its STFT. At the \( i \)th iteration, the desired STFTM \( |Y(m, k)| \) is substituted into \( X^i(m, k) \):
\[
\hat{X}^i(m, k) = |Y(m, k)| \frac{X^i(m, k)}{|X^i(m, k)|}
\] (4.71)
The next estimate $x^{i+1}(n)$ is obtained from $\hat{X}^i(m,k)$ using equation (4.67). For values of $(m,k)$ where $|X^i(m,k)| = 0$ in the equation above, $\hat{X}^i(m,k)$ is set to $|Y(m,k)|$. Each iteration of the LSEE-MSTFTM algorithm is guaranteed to decrease the distance metric $D_M[x(n), |Y(m,k)|]$.

### 4.6.2 Data and procedure

The input signals used in our evaluation are shown in figure 4.7. Each signal consists of a letter or digit spoken by a male speaker. The signals are sampled at 12800 Hz and are all 625 ms in length. These signals were selected because they are varied in their modulation content, yet they are short enough to be processed efficiently by our modulation filtering algorithms.

The signals are modulation filtered using the four filters defined in table 4.1. All stop-band ripple tolerances are set to 40 dB. Modulation filtering is done with four modulation transforms that are based on the incoherent (magnitude) detector, the smoothed Hilbert carrier estimator, the instantaneous frequency carrier estimator, and the frequency reas-
Table 4.1: Pass-band and stop-band frequencies of the modulation filters used in the analysis of signal reconstruction after modulation filtering.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Type</th>
<th>$F_{\text{pass}}$ (Hz)</th>
<th>$F_{\text{stop}}$ (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP1</td>
<td>low-pass</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LP2</td>
<td>low-pass</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>BP</td>
<td>band-pass</td>
<td>4; 12</td>
<td>2; 14</td>
</tr>
<tr>
<td>HP</td>
<td>high-pass</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

assignment carrier estimator, respectively. Signal reconstruction from the modified STFT is done with either the WOLA or the LSEE-MSTFTM reconstruction algorithm. The LSEE-MSTFTM algorithm requires an initial signal estimate $x^0(n)$, or an initial STFTP estimate $X^0(m,k)$. In our analysis we consider three initial phase estimates: the original STFTP, random phase, and zero phase. Although the coherent modulation filtering approach does not deal with the actual STFTM, it nonetheless uses WOLA for reconstruction. Therefore, it is valid to consider LSEE-MSTFTM as an alternative method of reconstruction for the coherent approach. In all instances, the LSEE-MSTFTM algorithm is run for 100 iterations.

4.6.3 Results

Reconstruction error

To measure the reconstruction error, we express the distance between the reconstructed signal and the desired STFTM with the signal-to-error ratio (SER) defined in [40] as follows:

$$\text{SER} [x(n), |Y(m,k)|] = 10 \log \left( \frac{\sum_m \sum_{k=0}^{K-1} |Y(m,k)|^2}{\sum_m \sum_{k=0}^{K-1} [||X(m,k)|| - |Y(m,k)||]^2} \right).$$

(4.72)

The advantage of the SER over the distance measures defined in section 4.6.1 is that the SER is independent of signal energy. It is guaranteed to increase in successive iterations of the LSEE-MSTFTM algorithm.

Figure 4.8 on page 92 shows the signal-to-error ratios as a function of iteration number for all the LSEE-MSTFTM signal reconstructions, averaged over the input signals. Each panel
in figure 4.8 corresponds to a specific combination of a modulation filter and a modulation transform. The SER of the WOLA reconstruction technique (which is identical to a single iteration of the LSEE-MSTFTM algorithm initialized with the original STFTP) is indicated in each panel by a black horizontal line.

The figure demonstrates that the performance of the modulation filters is fairly consistent regardless of the type of filter or the modulation transform used. Among the iterative techniques, the LSEE-MSTFTM algorithm initialized with the original STFTP yields the highest SER in most cases. Next best is random phase initialization, followed by initialization with zero phase. The SER of the LSEE-MSTFTM algorithm initialized with random phase and zero phase start out below the SER of the WOLA reconstruction technique, but reaches the WOLA level in the first 20–40 iterations. These two techniques could be interesting reconstruction alternatives when original phase is not available or can otherwise not be used. It seems that in that case the random phase initialization provides faster convergence and a higher SER. In general, the iterative reconstruction techniques yield a 3–5 dB improvement in SER over the WOLA reconstruction technique after 100 iterations.

Effective modulation frequency response

Figure 4.9 on page 93 shows the effective modulation frequency responses for the four modulation transforms and the four modulation filters, computed over the input signals according to the definition in equation (4.63). Each panel in figure 4.8 corresponds to a specific combination of a modulation filter and a modulation transform. It shows the effective modulation frequency response after WOLA reconstruction, and after LSEE-MSTFTM reconstruction initialized with original phase, random phase, and zero phase. The frequency response of the designed modulation filter is shown in each panel in gray.

It is evident from the figure that none of the effective modulation frequency responses matches the designed modulation filter response. This phenomena was previously reported for the incoherent modulation transform in [35, 120], but appears to hold true for the coherent modulation transforms as well. However, there are a few notable differences. First, the effective frequency response of coherent modulation transforms on the modulation bandpass
Figure 4.8: Average signal-to-error ratio (SER) of four modulation filters for incoherent and coherent modulation transforms, as a function of the number of iterations in signal reconstruction. Columns correspond with modulation filters, rows correspond with modulation transforms. Each panel shows the SER for WOLA reconstruction (black), and the SER per iteration of LSEE-MSTFTM initialized with original phase (blue), random STFTP (red), and zero phase (green).
Figure 4.9: Effective modulation frequency responses of four modulation filters for incoherent and coherent modulation transforms after different types of signal reconstruction. Columns correspond with modulation filters, rows correspond with modulation transforms. Each panel shows the effective modulation frequency response for WOLA reconstruction (black), and for LSEE-MSTFTM reconstruction initialized with original phase (blue), random phase (red), and zero phase (green). The frequency response of the designed modulation filter is shown in each panel in gray.
filter seems to be “out-of-sync” with the modulation filter design. Moreover, the coherent modulation transforms appear to be ineffective on the modulation highpass filter. These results, which were unexpected and somewhat disappointing, are not completely understood, but several explanations are plausible. One theory is based on the fact that the modulators of the incoherent modulation transform are “fixed” using half-wave rectification after modulation bandpass and highpass filtering. Such a fix is required to make the filtered modulators proper magnitudes again, since bandpass filtering or highpass filtering may have caused them to go negative. No such fix is possible or necessary for the modulators of the coherent transforms, since they are complex-valued. This difference between incoherent and coherent modulation transforms may be the cause of the ineffectiveness of the coherent transforms, or rather the effectiveness of the incoherent transform, on modulation bandpass and highpass filters. Another theory is that the mismatch between the modulation transform technique that is used to compute the effective modulation frequency response (incoherent) and the modulation transform technique that is used to filter the signals (coherent) may contribute to the apparent ineffectiveness of the coherent modulation transforms on modulation bandpass and highpass filtering.

Other notable differences between the effective modulation frequency responses of the incoherent modulation transform and the coherent modulation transforms are that the WOLA reconstruction technique seems to be more effective on the extreme lowpass filter LP1 with coherent processing than with incoherent processing. On the moderate lowpass filter LP2, however, the performance of the incoherent and coherent transforms in terms of their effective modulation frequency response is similar. Finally, the iterative LSEE-MSTFTM reconstruction technique improves the effective modulation frequency response mostly over the 30–60 Hz modulation frequency range, by about 10–15 dB. Remarkably, it does not make a real difference for the effective modulation frequency response whether the LSEE-MSTFTM algorithm is initialized with original phase, random phase, or zero phase, regardless of the modulation transform or modulation filter that was used.
Informal listening test

An informal listening test of all processed signals revealed that all methods produced signals of similar quality. The signals generated with the LSEE-MSTFTM algorithm were slightly more “buzzy”, which was most noticeable after zero phase initialization. This effect was also somewhat enhanced by the extreme low-pass filter LP1. The coherently filtered signals sounded a little less reverberated and seemed to preserve a little more of the natural quality of the original speech. Based on the informal listening test, we would prefer the WOLA signals of either transform, but would not reject the LSEE-MSTFTM signals.

4.6.4 Conclusions and discussion

Comparing the WOLA and LSEE-MSTFTM signal reconstruction techniques, we conclude that the WOLA technique is a good choice for signal reconstruction after modulation filtering. Overall, it achieves reasonable stopband suppression and good signal quality. The reconstruction error of WOLA is usually small and hard to improve with LSEE-MSTFTM without using considerably more processing time. In summary, we think that the LSEE-MSTFTM reconstruction technique does not outperform the WOLA technique in terms of signal-to-error ratio, modulation filtering effectiveness or signal quality to such an extent that its higher computational demand is justified.

Nevertheless, the LSEE-MSTFTM algorithm with random phase initialization could still be useful in cases where the original STFTP cannot be used. For example, if the subband envelopes are filtered by an IIR filter with non-linear phase response, it may be hard or impractical to “line-up” the original STFTP with the modified STFTM. Then, signal reconstruction from MSTFTM alone is unavoidable, and the LSEE-MSTFTM signal reconstruction technique with random phase initialization would be a good choice.

Comparing the different modulation transforms, we conclude that the coherent transforms estimator performed below our expectations with respect to the effective modulation frequency response. From the arguments given in [78, 120, 122], and considering the positive results obtained with coherent transforms on music signals in [4, 122], we believe that coherent modulation transforms are in principal better than incoherent modulation transforms.
for modulation frequency analysis and filtering of natural signals such as speech and music. However, their ineffectiveness with modulation bandpass and highpass filtering warrants further study.

4.7 Summary

In this chapter we introduced incoherent envelope detectors, namely the Hilbert envelope detector and the magnitude detector, and showed that they are equivalent to each other. We demonstrated that the non-negative real modulators of these incoherent detectors have three important limitations: they violate the bandwidth constraint of envelope detectors, they force a symmetric spectrum onto the modulator, and they are not closed under convolution. We demonstrated that these problems can be repaired by using coherent envelope detectors that yield complex-valued modulators. We defined three types of coherent detectors, and compared their performance on three broad signal classes. We introduced a method to measure the effectiveness of modulation filters, similar to the (temporal) modulation transfer functions defined in the literature. Finally, we analyzed two techniques for signal reconstruction from modified short-time Fourier transform, and determined that the current approach of weighted overlap-add reconstruction is a suitable technique for inverting our modulation transforms.
Chapter 5

APPLICATION

5.1 Introduction

In the chapters 3 and 4 we have described the theoretical and practical aspects of modulation filtering systems. In this chapter and the next, we focus on the second objective of this dissertation: to design a signal processing technique based on modulation analysis and modulation filtering that improves the intelligibility of a target talker in the presence of interfering talkers in a single channel (monophonic) recording.

We choose this problem because it is a highly relevant problem for users of hearing aids and cochlear implants for which no satisfactory solution exists yet, as we mentioned in the introduction of this dissertation. Additionally, we discovered while analyzing short-time modulation spectrograms of speech mixtures that individual talkers are naturally separated in the wideband modulation frequency domain. This prompted us to investigate the full potential of modulation filtering systems to automatically enhance the intelligibility of a target talker in favor of interfering talkers.

We limit ourselves to algorithms that are designed to enhance only a single talker, whose parameters can be trained or adjusted off-line to a particular target talker. This restriction is not unpractical if we assume that users of hearing devices are primarily interested in the enhancement of one specific talker, for example his or her spouse. To ensure that our enhancement techniques are useful in hearing devices under cocktail-party conditions, they should be able to identify the target talker in a mixture of an unknown number of unknown interfering talkers. Furthermore, they should be able to identify the talker based solely on information about the target talker that can be acquired during fitting of the hearing device in a clinical setting; for example, information about the timbre of the target talker’s voice, or the target talker’s habitual fundamental frequency. The algorithms should also be able to operate under a wide range of signal to noise ratios. Finally, they should produce high
quality output that enhances the target talker but does not introduce distortion or artifacts.

In this chapter, we present two preliminary experiments that motivated our choice to apply modulation filtering to the problem of target talker enhancement. Then we describe our initial efforts to automate this target talker enhancement experiment using incoherent modulation frequency analysis. Finally, we discuss our latest coherent modulation filtering technique for high-quality intelligibility enhancement of a target talker.

5.2 Preliminary experiments

Our motivation to apply modulation analysis and filtering to the problem of target talker enhancements comes from the successful separation of two co-channel talkers, by filtering each talker in the modulation frequency domain using manually designed modulation masks. This experiment is described in section 5.2.1. We established the potential performance of automated modulation filtering techniques to the talker enhancement problem in an optimal modulation filtering experiment, which is described in section 5.2.1.

5.2.1 Manual modulation masking experiment

Modulation spectrograms of speech from multiple talkers exhibit several prominent features that could possibly be exploited to separate talkers. For example, figure 5.1a shows the wideband incoherent modulation spectrogram of the sum of two talkers, DSTMT\{x_a(n) + x_b(n)\}. The speech signals \(x_a(n)\) and \(x_b(n)\) are taken from Te-Won Lee’s “a real cocktail party effect” dataset [76], and are a recording of a male talker saying “one, two, three,” and a different male talker saying “uno, dos, tres,” respectively. The figure shows that the modulation spectral energy of the two talkers is concentrated at different modulation frequencies due to their different fundamental frequencies. Moreover, the energy at each talker’s fundamental frequency in the modulation frequency dimension is localized in acoustic frequency, and peaks at major acoustic features of its source such as formants, as is illustrated by figure 5.1b. The separation of the two talkers in modulation frequency, together with the

Figure 5.1: Modulation spectrogram of the sum of two talkers, and localization of their fundamental frequency energy across acoustic frequency. (a) Modulation spectrogram. Highlighted features are the fundamental frequency energy of talker A (red) and talker B (black). (b) Localization of the fundamental frequency energy of talker A (red) and B (black) across acoustic frequency.

acoustic frequency localization of fundamental frequency energy, suggests that the talkers are separable in the modulation domain.

Method

To test this hypothesis, we analyzed the modulation spectral content of 262 ms long frames of both $x_a(n)$ and $x_b(n)$, as well as their 0 dB mix $x(n) = x_a(n) + x_b(n)$, at 64 ms intervals. Based on the apparent separation of the talkers in the modulation domain, we manually constructed two binary modulation spectral masks, $M_a(i, k)$ and $M_b(i, k)$, such that

$$\tilde{x}_a(n) = \text{IDSTMT} \left\{ M_a(i, k) \text{DSTMT}\{ x(n) \} \right\} \approx x_a(n) \quad (5.1)$$

and

$$\tilde{x}_b(n) = \text{IDSTMT} \left\{ M_b(i, k) \text{DSTMT}\{ x(n) \} \right\} \approx x_b(n). \quad (5.2)$$

Figure 5.2 shows an example of both masks for the frame at time $t = 1.216$ ($l = 19$) of the signal $x(n)$. In this example, the mask for each talker suppresses the acoustic frequencies
Figure 5.2: Modulation spectrogram of the speech mixture, and manually designed modulation masks. (a) Modulation spectrogram of the speech mixture, DSTMT\{x_a(n) + x_b(n)\}; (b) Modulation mask $M_{19}^a(i,k)$, superimposed in black on the modulation spectrogram of (a); (c) Mask $M_{19}^b(i,k)$, superimposed in black on the modulation spectrogram of (a). Black indicates regions of the modulation spectrogram that will be masked out (i.e. set to zero) before signal reconstruction.

associated with the other talker over the entire modulation frequency range, except for modulation frequencies at the target’s fundamental frequency.

**Results**

The separation performance of the manual modulation masks was measured with the signal-to-distortion ratio (SDR) that is commonly used to evaluate blind source separation algorithms. The SDR is defined in [153] as

$$\text{SDR} = 10 \log_{10} \frac{\sum_n x^2(n)}{\sum_n [x(n) - \hat{x}(n)]^2},$$

where the difference between the original signal $x(n)$ and the separated signal $\hat{x}(n)$ in the denominator represents the separation distortion. The SDR of the first signal was $\text{SDR}_a = 5.34 \text{ dB}$, and of the second signal $\text{SDR}_b = 5.20$ dB. This indicates that the manual modulation masks achieve a little more than 5 dB separation between the two signals. The result of masking the short-time modulation transform of $x(n)$ with the masks $M_{i}^{a}(i,k)$
and $M^b(i,k)$ is illustrated by the spectrograms in figure 5.3. As these spectrograms show, formants are well preserved and assigned to the correct speaker, and most voiceless sounds are also separated well. The loss of low frequencies in the voiced fricative /s/ in “dos” and “tres” is audible as slight amplitude pumping. Although the spectrograms do not show it, there is also some audible crosstalk between the talkers at word onsets and offsets. The separation of the speakers is good, but the non-linearity of the modulation domain signal processing has added distortion to the signals, giving them a slight metallic or tinny quality.

5.2.2 Optimal coherent filtering experiment

We evaluated the performance of an “optimal” coherent modulation filtering technique in a subjective listening test, to determine how successful coherent modulation filtering
techniques could potentially be at improving the intelligibility of a target talker in the presence of other talkers. Our optimal coherent filtering technique is similar to the Wiener-like RASTA filtering technique of Hermansky et al. [50]. It attempts to minimize the squared error between a known target signal and the same target signal corrupted by interfering talkers noise, by using a time-varying modulation filter.

Methods

In the subjective listening test, stimuli were presented in three processing conditions: (1) original, unprocessed stimuli; (2) linear, time invariant bandpass coherent modulation filter; and (3) least-squares optimal time-varying coherent modulation filter.

The unprocessed stimuli are used as the reference condition to which the other two conditions are compared. The bandpass modulation filter is included as a second control condition, to evaluate the benefit of optimal and time-varying modulation filtering over a regular modulation filter. The passband of the bandpass filter extended from 4–16 Hz in modulation frequency. The hypothesis supporting this choice is that corruption of a speech signal by other speech signals will introduce high modulation frequencies in the corrupted signal. By eliminating modulation frequencies outside the range of frequencies that are important for single talker speech intelligibility, we expect to reduce the influence of the interfering talkers.

The optimal coherent modulation filtering algorithm uses the short-time Fourier transform to decompose a signal into subbands, and the instantaneous frequency carrier estimator defined in section 4.4.2 to separate the subbands into carriers and modulators,

\[ X_k(m) = \text{DSTFT}\{x(n)\} \]  \hspace{1cm} (5.4)
\[ M_k(m) = \mathcal{D}_{\text{IF}}\{X_k(m)\} \]  \hspace{1cm} (5.5)
\[ C_k(m) = \mathcal{D}^\circ_{\text{IF}}\{X_k(m)\}. \]  \hspace{1cm} (5.6)

Given the target talker’s signal, \(x(n)\), and the target signal corrupted with noise from other talkers, \(\bar{x}(n) = x(n) + b(n)\), the algorithm computes \(X_k(m) = \text{DSTFT}\{x(n)\}\) as well as the carriers \(\bar{C}_k(m)\) and modulators \(\bar{M}_k(m)\) of the corrupted signal \(\bar{x}(n)\). It demodulates the...
target signal’s STFT by the corrupted modulators,

\[ M_k(m) = X_k(m)\bar{C}_k^*(m). \]  

Next it determines a modulation filter \( \hat{g}_k(m) \) for each subband that minimizes the squared error between the desired modulator and the filtered corrupted modulator,

\[ \hat{g}_k(m) = \arg\min_{g(m)} \sum_m |M_k(m) - \bar{M}_k(m) * g(m)|^2. \]  

Each of the modulation filters \( \hat{g}_k(m) \) is limited to 24 ms in length to have sufficient filtering effectiveness with an acceptable amount of latency without over-fitting the problem. The filter coefficients are synchronously updated every 250 ms for all subbands.

To validate the optimal filtering approach, the quality of speech processed by the optimal coherent modulation filter was informally compared to the output of a similar least-squares optimal filtering algorithm that operated entirely in the time-domain. The time-domain algorithm designs an LTI filter \( \hat{h}(n) \) that minimizes the squared error between the desired signal and the filtered corrupted signal,

\[ \hat{h}(n) = \arg\min_{h(n)} \sum_n |x(n) - \bar{x}(n) * h(n)|^2. \]  

The time-domain filter was limited to 20 ms in length and updated every 25 ms to have the same degrees of freedom as the optimal modulation filters. An informal subjective evaluation showed that the modulation domain approach enhanced the intelligibility of the target talker more, and yielded speech signals with less artifacts and better suppression of the interfering talkers than the time-domain algorithm.

**Stimuli**

We used the same twelve target speech signals as in the experiments by Turner et al. [144]. Each signal is a spondee (i.e., a bisyllabic word with equal stress on each syllable, such as birthday, northwest, or woodwork) spoken by a female talker. The fundamental frequency
of the spondees ranges from 212–250 Hz, and they range in duration from 1.12 to 1.63 seconds [144]. The twelve spondees are taken from a commercial recording by Harris [45], and are of equal difficulty for this type of subjective test.

The interfering talkers noise signal in our experiment is the same as the competing-talker condition of Turner et al. [144]. It consists of a sentence spoken by a male talker saying “Name two uses for ice” (fundamental frequency range 81–106 Hz) and a sentence spoken by a female talker, different from the target female talker, saying “Bill might discuss the foam” (fundamental frequency range 149-277 Hz). These sentences were originally recorded for the SPIN test [10]. The sentences were mixed at equal root-mean-squared (RMS) amplitudes, and the mixed signal was normalized to the same RMS amplitudes as the spondees. The duration of the interfering talkers noise signal is 1.96 seconds.

Each target signal was mixed with the interfering talkers noise signal at a signal-to-noise ratio (SNR) ranging from -50 dB to +30 dB in steps of 2 dB. The RMS amplitude of the target signals was kept constant in all mixtures, and the RMS amplitude of the noise signal was scaled to the desired SNR. The same interfering talkers noise signal was used in all mixtures. The onset of the spondee is 500 ms after the onset of the interfering talkers noise.

**Subjects and presentation**

Five bilateral hearing loss patients (ages 40-83) and six normal hearing subjects (ages 26-38) participated in the test. Subjects were seated in a double-walled, sound attenuating booth. All sounds were presented binaurally to the subjects through TDH-50P headphones. The loudness of the target talker spondees was calibrated to 65 dBA prior to the experiment, and was held at that level throughout the test.

For the normal hearing subjects, all stimuli were processed with a 6-channel cochlear implant simulator before presentation. The simulator divides an input signal into subbands using a six-channel FIR filterbank whose subband frequency ranges are 80–308, 308–788, 788–1794, 1794–3906, 3960–8338, and 8338–17640 Hz. It combines the Hilbert envelope of each subband with a (broadband) white noise carrier, filters the resulting noisy envelope with the corresponding subband filter, and sums all subband outputs together to form a
signal that simulates the output of a cochlear implant.

**Procedure**

The subjective listening test consisted of a training phase and a testing phase. During training, subjects were allowed to listen to each spondee (without interfering talkers noise) until they had familiarized themselves with the sounds. The testing phase of the listening test was setup as an adaptive speech reception threshold (SRT), twelve alternative forced-choice test (12AFC) using a simple 1-up, 1-down method [77]. In each trial, a spondee was randomly selected out of the twelve spondees, and was presented in noise at a certain SNR. The subject selected the spondee they heard from twelve buttons on a monitor which were labeled with the spondee words. The subject was required to respond in each trial. The SNR was decreased by 2 dB on a positive response, and increased by 2 dB on a negative response. This procedure was repeated until 14 reversals in SNR were completed. The mean of the SNR at the last 10 reversals was taken as the estimate of the 50% correct SRT on the psychometric curve [77].

The 50% correct SRT levels of all three processing conditions were measured in random order. The measurements were repeated three times for each subject, and the order of the three processing conditions was randomized in each repetition. The adaptive procedure and user interface for the test were implemented in MATLAB.

**Results and discussion**

The results of the listening test are shown on page 106 for the hearing impaired subjects, and on page 107 for the normal hearing subjects. In figure 5.4, the estimated speech reception threshold (SRT) of each hearing impaired subject is plotted against repetition number for all processing conditions. Note that the estimated SRT of some of the hearing impaired subjects is as good as -30 dB SNR because the subjective test was a closed-set recognition task. The second hearing impaired subject did not complete all 3 repetitions for the bandpass and optimal modulation filtering conditions, and all data from this subject was excluded from further analysis.
Figure 5.4: Estimated speech reception thresholds for the hearing impaired subjects of the listening test on optimal modulation filtered speech. Speech reception threshold (SRT), in dB, plotted against repetition number for each subject, and for three processing methods: → unprocessed; ➤ bandpass coherent modulation filtering; ➦ optimal coherent modulation filtering.

Table 5.1: Table of means for the hearing impaired subjects of the listening test on optimal modulation filtered speech. Average speech reception thresholds over subjects 1, 3, 4, and 5, by processing method and repetition number. Difference between methods can be seen by comparing rows, learning effect can be seen by comparing columns.

<table>
<thead>
<tr>
<th>processing</th>
<th>repetition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td></td>
<td>-19.6</td>
<td>-23.0</td>
<td>-23.6</td>
<td>-22.1</td>
</tr>
<tr>
<td>bandpass</td>
<td></td>
<td>-14.3</td>
<td>-15.0</td>
<td>-14.7</td>
<td>-14.7</td>
</tr>
<tr>
<td>optimal</td>
<td></td>
<td>-23.5</td>
<td>-27.8</td>
<td>-26.2</td>
<td>-25.8</td>
</tr>
<tr>
<td>overall</td>
<td></td>
<td>-19.1</td>
<td>-21.9</td>
<td>-21.5</td>
<td>-20.9</td>
</tr>
</tbody>
</table>

Table 5.2: Repeated measures ANOVA for the hearing impaired subjects of the listening test on optimal modulation filtered speech. F-statistic and p-value of the within-subject factors processing and repetition, and their interaction. (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>processing</td>
<td>$F_{2,6} = 13.73$</td>
<td>0.006 (**)</td>
</tr>
<tr>
<td>repetition</td>
<td>$F_{2,6} = 1.96$</td>
<td>0.22</td>
</tr>
<tr>
<td>interaction</td>
<td>$F_{4,12} = 0.42$</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Figure 5.5: Estimated speech reception thresholds for the normal hearing subjects of the listening test on optimal modulation filtered speech. Speech reception threshold (SRT), in dB, plotted against repetition number for each subject, and for three processing methods:

- unprocessed;
- bandpass coherent modulation filtering;
- optimal coherent modulation filtering.

Table 5.3: Table of means for the normal hearing subjects of the listening test on optimal modulation filtered speech. Average speech reception thresholds over all subjects by processing method and repetition number. Difference between methods can be seen by comparing rows, learning effect can be seen by comparing columns.

<table>
<thead>
<tr>
<th>processing</th>
<th>repetition</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>unprocessed</td>
<td>-9.4</td>
<td>-10.8</td>
</tr>
<tr>
<td>bandpass</td>
<td>-7.7</td>
<td>-9.6</td>
</tr>
<tr>
<td>optimal</td>
<td>-28.2</td>
<td>-30.5</td>
</tr>
<tr>
<td>overall</td>
<td>-15.1</td>
<td>-17.0</td>
</tr>
</tbody>
</table>

Table 5.4: Repeated measures ANOVA for the normal hearing subjects of the listening test on optimal modulation filtered speech. F-statistic and p-value of the within-subject factors processing and repetition, and their interaction. (*) $p < 0.05$, (**) $p < 0.01$, (***$) p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>processing</td>
<td>$F_{2,10} = 181.57$</td>
<td>$&lt; 0.001$ (***$)</td>
</tr>
<tr>
<td>repetition</td>
<td>$F_{2,10} = 6.65$</td>
<td>$0.015$ (*)</td>
</tr>
<tr>
<td>interaction</td>
<td>$F_{4,20} = 0.92$</td>
<td>$0.47$</td>
</tr>
</tbody>
</table>
The average estimated SRT per processing condition and repetition over all hearing impaired subjects are presented in table 5.1. Overall, the hearing impaired subjects did not benefit from the bandpass coherent modulation filter (an increase in SRT by 7.4 dB SNR), but did benefit from processing with the optimal coherent modulation filter (a decrease in SRT by 3.7 dB SNR). The results of a repeated measures ANOVA that was performed on this data is shown in table 5.2. The analysis revealed that the effect of the processing method was significant, but that the learning effect and the interaction between processing method and repetition number were not significant.

In figure 5.5, the estimated speech reception threshold (SRT) of each normal hearing subject is plotted against repetition number for all processing conditions. Note that the estimated SRT of some of the normal hearing subjects is as good as -45 dB SNR because the subjective test was a closed-set recognition task.

The average estimated SRT per processing condition and repetition over all normal hearing subjects are presented in table 5.3. Overall, the normal hearing subjects did not benefit from the bandpass coherent modulation filter (an increase in SRT by 0.05 dB SNR), but did benefit greatly from processing with the optimal coherent modulation filter (a decrease in SRT by 21.4 dB SNR). The results of a repeated measures ANOVA that was performed on this data is shown in table 5.4. The analysis revealed that the effect of the processing method was very significant, and that there was a significant learning effect. The interaction between processing method and repetition number was not significant.

The results show that speech intelligibility improves consistently with the optimal coherent modulation filter. However, most of the subjects do not benefit from the bandpass modulation filter. Our hypothesis that reducing modulation frequencies above 16 Hz increases single talker speech intelligibility appears to be false.

This experiment was not intended as a final solution to the talker enhancement problem, but just as a “proof of concept” for the use of modulation filtering to improve speech intelligibility. The subjective test shows that optimal coherent modulation filtering can improve the intelligibility of speech considerably.
5.3 Talker enhancement I: Incoherent frequency weighting\(^3\)

The optimal filtering experiment in section 5.2.2 showed that it is possible to achieve a considerable intelligibility improvement using a coherent modulation filter. This improvement was only obtained, however, with full knowledge of the desired output signal, which is unavailable in realistic circumstances. Therefore, our next goal is to design a modulation filtering scheme that improves the intelligibility of a target talker while relying only on information about the target that could be available \textit{a priori} in realistic circumstances.

To translate our experience from the manual modulation masking experiment into an automated talker separation method, we analyzed the design process of the manual modulation masks. We concluded that the observations about the distribution of energy in the joint-frequency domain, as shown in figure 5.1, could be summarized into the following two principles: (1) talkers can be identified and detected in the modulation frequency dimension based on their fundamental frequency range; (2) localization of a talker’s fundamental frequency energy in acoustic frequency determines the acoustic frequencies that are relevant for the intelligibility of a talker. Given these principles, a human operator judiciously assigns those regions of the joint-frequency plane to the modulation mask for a talker that contribute most to the perception of that talker, and removes those regions from the mask that contribute most to the perception of the other talker.

Note that applying these principles requires wideband incoherent modulation analysis of a signal, since the fundamental frequency of a talker must be resolved in modulation frequency rather than in acoustic frequency. Neither coherent nor narrowband incoherent modulation analysis have the ability to resolve fundamental frequency in the modulation dimension. Hence, the use of incoherent modulation analysis is essential to transform the manual modulation masking experiment into an automated talker separation method. It is however equally important that the separated signals are of high quality and free of artifacts for the intended application in hearing devices. Therefore, we opted to use only the forward modulation transform to analyze signals, to avoid all distortion, although small,\(^3\)

from reconstruction from the modified short-time Fourier transform (see section 3.5 and 4.6). Instead, we propose to design the automated separation algorithm such that the output of the modulation frequency analysis stage are the coefficients to a more conventional linear time-varying FIR filter that separates the talkers.

The proposed separation technique is parameterized by the fundamental frequency ranges of the talkers in the signal. The fundamental frequency range of a talker is relatively easy to acquire from just a few recording of its voice, which makes it an attractive parameter for a separation technique. However, using the fundamental frequency range has several weaknesses too. It is only useful for discriminating voiced speech segments, such as vowels, and does not provide any information to identify unvoiced speech, such as consonants. Furthermore, the fundamental frequency of speech from a single talker changes over time, as is illustrated by the example of a speech signal and its time-varying fundamental frequency in figure 5.6. The fundamental frequency of interfering talkers is similarly time-varying, and will occasionally enter the target’s fundamental frequency range, causing false detections.
Such false detections can be mitigated with a more sophisticated talker identification algorithm, in combination with an algorithm that tracks the target’s fundamental frequency over time. In the initial design of our separation algorithm, we avoid these weaknesses by assuming that the fundamental frequency range of the target and the interfering talker are sufficiently non-overlapping.

5.3.1 Method

To demonstrate the feasibility of using a speech signal’s modulation spectral representation to automatically separate two talkers, we devised the following target enhancement algorithm. Given a signal sampled at $F_s$ Hz that is the sum of a target talker and an interfering talker, i.e., $x(n) = x_t(n) + x_i(n)$. Assume that the target talker’s fundamental frequency is within the frequency range $F_t = [F_{t,\text{low}}, F_{t,\text{high}}]$ and that the interfering talker’s fundamental frequency is within the range $F_i = [F_{i,\text{low}}, F_{i,\text{high}}]$. The fundamental frequency ranges can be broad, but should be sufficiently non-overlapping. Let $X_t(i, k) =$ DSTMT$_{R,S}$($x(n)$) be the multi-rate discrete short-time modulation transform of the sum signal $x(n)$, with DSTFT downsampling factors $R$ and $S$. Define

$$ Q = \left\{ i : \frac{i F_s}{R} \in F \right\} \quad (5.10) $$

as the set of modulation frequency indexes $i$ that fall in the fundamental frequency range $F$. Consider the modulation spectral energy as a function of acoustic frequency index over the target’s fundamental frequency range,

$$ E_t^i(k) = \sum_{i \in Q_t} |X_t(i, k)|^2, \quad (5.11) $$

as well as over the interfering talker’s fundamental frequency range,

$$ E_i^i(k) = \sum_{i \in Q_i} |X_i(i, k)|^2. \quad (5.12) $$
For each frame $l$, i.e., the time instance $n = lRS$, the ratio of target energy and interferer energy yields a frequency masking function

$$F_l(k) = \frac{E^t_l(k)}{E^t_l(k) + E^i_l(k)}.$$  \hfill (5.13)

Each frame’s frequency masking function is transformed to an impulse response by combining it with the appropriate linear phase response $\phi(k)$ and taking the inverse DFT,

$$f_l(n) = \frac{1}{K} \sum_{k=0}^{K-1} F_l(k) e^{j\phi(k)} e^{j2\pi nk/K}.$$  \hfill (5.14)

A time-varying filter $h_k(n)$ is constructed for all times $k$ by linear combination of the two nearest impulse responses, according to

$$h_k(n) = (1 - \alpha_k)f_{\beta_k}(n) + \alpha_k f_{1+\beta_k}(n),$$  \hfill (5.15)

where $\alpha_k = k/RS - \beta_k$, $\beta_k = \lfloor k/RS \rfloor$ and $\lfloor x \rfloor$ indicates the largest integer smaller than or equal to $x$. The time-varying filter is then used to separate the target talker from the interfering talker, as follows

$$\bar{x}_t(n) = \sum_k x(k)h_k(n).$$  \hfill (5.16)

5.3.2 Signals and processing

We evaluated the incoherent frequency weighting algorithm on two speech signals from the TIMIT data set [34]. The first signal, $x_c(n)$, is a sentence spoken by a male talker saying “Will Robin wear a yellow lily?” (fundamental frequency range 100–124 Hz), and the second signal, $x_d(n)$, is a sentence spoken by a different male talker saying “Do they allow atheists in church?” (fundamental frequency range 125–164 Hz). The second signal was truncated to the duration of the first signal of 1.78 seconds, and the signals were mixed at equal RMS amplitudes, $x(n) = x_c(n) + x_d(n)$.

We applied the incoherent frequency weighting algorithm described in section 5.3.1 to the mixture signal $x(n)$, once with the first talker as the target talker and once with the
second talker as the target talker. The algorithm parameters were set to $R = 16$, $K = 512$, $S = 38$, $I = 512$ and the analysis windows $w(n)$ and $v(m)$ were a 48-point and 78-point Hanning window. The sampling frequency of the signals was 16 kHz.

### Results

The performance of the algorithm was measured with the signal-to-distortion ratio (SDR) specified in equation (5.3). The SDR of the individual talkers was $\text{SDR}_c = 4.13$ dB and $\text{SDR}_d = 3.71$ dB. This indicates that the incoherent frequency weighting algorithm achieves about 4 dB improvement in SNR. A qualitative evaluation of the algorithm is provided by the spectrograms in figure 5.7. As these spectrograms show, the incoherent frequency weighting algorithm allocates the formants (and hence the vowels and other sonorant sounds) to the correct speaker. There is some crosstalk between the speakers, most noticeably the voiceless fricative /s/ in “atheist” at $t = 1.3$ and the voiceless affricate /ch/ in “church” at $t = 1.7$. The stop /t/ in “atheist” is diminished in the output of the second talker, and the glide

![Spectrograms](image-url)
/w/ in “wear” is lost in the output of the second talker. Besides these issues with unvoiced sounds, the separated output sounds natural and undistorted, and with reasonably good separation between the speakers.

5.3.4 Discussion

Although the results obtained with the incoherent frequency weighting algorithm are encouraging, we found that the method did not generalize well to mixtures of other talkers or to lower talker-to-interference ratios (TIR).

The most prohibitive factor for generalization to other mixtures is the overlap between the fundamental frequency ranges of the talkers. When the fundamental frequency ranges of the target talker and the interfering talker overlap more, the method’s ability to distinguish between the talkers reduces, as expected. One way to enhance this ability would seem to shorten the analysis frames of the modulation transforms, for the following reasons. Short-time modulation transforms analyze signals in frames that are usually 200–500 ms in duration. The fundamental frequency of a speech signal typically varies considerably over such a time frame. Therefore, the fundamental frequency energy extends over a wide range of modulation frequencies in the incoherent modulation spectrogram. Shortening the modulation analysis frame reduces this spread, as the fundamental frequency varies less over the duration of the shorter frame. However, the modulation spectrogram’s resolution also decreases proportionally to the length of the analysis frame, negating the effect of the reduced spread. Lengthening the analysis frame would increase the modulation spectrogram’s resolution, but also averages more of the speech signal resulting in more spread. The trade-off between spread and resolution imposes an upper bound on the ability of the wideband incoherent modulation spectrogram to resolve the time-varying fundamental frequency of a talker.

The performance of the incoherent frequency weighting algorithm is also negatively affected by FM to AM transduction of the time-varying fundamental frequency and its harmonics. This can be understood as follows. The fundamental frequency and harmonics of a speech signal vary in frequency over time. Filterbanks that are used in modulation trans-
forms have subbands that have fixed locations in frequency. Consequently, the fundamental frequency and its harmonics move in and out of the subbands over time. When a harmonic falls in a subband, the envelope of that subband corresponds to the amplitude modulation of the harmonic. However, when the harmonic moves out of the subband, the envelope of the subband will drop to zero, even when the amplitude of the harmonic is non-zero. Hence, the frequency modulation of the harmonic is incorrectly perceived as amplitude modulation in the subband. The FM-to-AM transduction effect, which can actually be exploited to detect amplitude modulation by cooperation between neighboring subbands [107], causes modulation detection errors in the current definition of modulation transforms.

The incoherent frequency weighting algorithm could possibly be improved by incorporating modulation spectral phase in the analysis. Currently, the algorithm uses modulation spectral magnitude only. Since modulation spectral magnitude is time-shift invariant, it does not convey any information about one important cue for talker separation: common onset and offset of harmonics [8]. We have attempted to use the initial phase as well as the group delay of subband envelopes to cluster subbands into target and interfering talker clusters. Unfortunately, we were unsuccessful in improving the talker separation algorithm by incorporating modulation spectral phase. It seems that time-related features such as common onset and offset of harmonics are better detected in the time-frequency domain than in the complex modulation spectral domain.

Based on our analysis of the separation results of the incoherent frequency weighting algorithm, we expect that small improvements of the incoherent algorithm are possible. However, we also expect that its two fundamental weaknesses, i.e., the modulation frequency resolution upper bound and the FM-to-AM transduction errors, prevent great improvements in the robustness of the algorithm. We therefore conclude that a new modulation filtering paradigm is needed to develop a robust talker separation algorithm.

5.4 Talker enhancement II: Coherent modulation filter

To enhance a target talker in the presence of interfering talkers we propose a novel coherent modulation filtering technique. The technique can be considered to be a coherent modulation filtering technique, since its main processing components are a carrier estimator and a
(lowpass) modulation filter. However, each component is implemented differently than in a typical modulation filtering system. These custom implementations are necessary in order to meet the requirements we impose on the talker enhancement technique. Specifically, the carrier estimator must be robust to noise, since it must be able to detect the target talker in the presence of multiple interfering talkers, and at target-to-interference ratios well below 0 dB. Furthermore, the talker enhancement technique must achieve high-quality output that is free of processing artifacts, because it must be acceptable in hearing aid applications. These requirement put high demands on the carrier estimator and the modulation filter and warrant their novel implementation. Section 5.4.1 and 5.4.2 describe each component of the proposed talker enhancement technique in detail. Another novelty of the proposed technique is that it adaptively mixes the output of the coherent modulation filtering with a filtered version of the input signal to avoid high-frequency artifacts. Details of the mixing procedure are discussed in section 5.4.3.

5.4.1 Carrier estimator

The coherent carrier estimators discussed in section 4.4 are designed to operate on noise-free data. In the presence of an interfering talker, they are unable to detect the carriers of only the target talker, and return subband carriers that are a mix of the target talker’s carriers and the interfering talker’s carriers. The novel implementation of the carrier estimator detects the carriers of the target talker regardless of the presence of one or more interfering talkers. It consists of a target fundamental frequency detector and a carrier refinement stage.

Target fundamental frequency estimator

The target fundamental frequency detector detects harmonic structures in the input signal at the fundamental frequency range of the target talker. It is based on a harmonic model of voiced speech similar to the one used in the STRAIGHT technique [62],

\[
x(t) = \sum_k A_k(t) \sin \left( \int_{t_0}^t k[\Omega(\tau) + \Omega_k(\tau)]d\tau + \phi_k \right),
\]

(5.17)
where \( A_k(t) \) is the time-varying amplitude, \( \Omega_k(t) \) is the time-varying angular frequency, and \( \phi_k \) is the phase at time \( t_0 \) of the \( k \)-th harmonic component. According to this model, voiced speech can be accurately described as a nearly harmonic sum of amplitude modulated sinusoids that are frequency modulated by an individual FM-component \( \Omega_k(t) \) around a common FM-component \( \Omega(t) \). The objective of the fundamental frequency detector is to robustly determine this common FM-component \( \Omega(t) \) for a target talker.

After reviewing and evaluating the fundamental frequency estimators defined in [2, 67, 68, 91, 96, 99, 127, 143, 155, 157], we found that a modified version of the harmonic product spectrum method by Schroeder [127] worked best for our purposes. To estimate \( f(m) \), a time-sampled version of \( \Omega(t) \), our fundamental frequency detector first computes the discrete short-time Fourier transform \( X(m, \omega) \) of the input signal \( x(n) \) sampled at \( F_s \) Hz. It then sums the log-magnitude-squared of \( X(m, \omega) \) over the first \( P \) harmonic multiples of the target talker’s fundamental frequency range \( F = [F_l, F_h] \),

\[
D(m, \omega) = \sum_{p=1}^{P} \log |X(m, p\omega)|^2, \quad 2\pi \frac{F_l}{F_s} \leq \omega \leq 2\pi \frac{F_h}{F_s}. \quad (5.18)
\]

The harmonic product spectrum frame \( D_m(\omega) = D(m, \omega) \) represents the total amount of energy that is present in the first \( P \) harmonics of the fundamental frequency \( \omega \) (in radians/sample) at time \( m \). Signal components that are harmonically related in the target talker’s fundamental frequency range add up constructively and create a maximum in the frame \( D_m(\omega) \).

Using the harmonic product spectrum \( D(m, \omega) \) in this form has several benefits. \( D(m, \omega) \) is evaluated only over the target talker’s fundamental frequency range, \( F = [F_l, F_h] \), so that octave errors in the fundamental frequency estimate do not occur because typically \( F_h < 2F_l \). Furthermore, the harmonic product spectrum is robust to noise because it uses all of a speech signal’s strongest harmonics to estimate its fundamental frequency. Finally, it allows a simple and efficient implementation, for example via quadratic interpolation of frequency samples computed with the fast Fourier transform [1, 108].

The harmonic product spectrum \( D(m, \omega) \) is used to compute the common fundamental frequency component \( \omega(m) \) as follows. First, two empirical measures of the peakedness of
a frame $D_m(\omega) = D(m, \omega)$ define a target voice activity detector $v(m)$,

\[
v(m) = \begin{cases} 
1 & v(m-1) = 0, \ p(m) > p_1, \ q(m) > q_1 \\
0 & v(m-1) = 1, \ p(m) < p_0, \ q(m) < q_0 \\
v(m-1) & \text{otherwise}
\end{cases}
\]  
(5.19)

where $p(m)$ and $q(m)$ express the peakedness of the maximum of $D_m(\omega)$ with respect to the (other) local maxima of $D_m(\omega)$ and to the entire frame $D_m(\omega)$, respectively. They are computed according to

\[
p(m) = \frac{\max[D_m(\omega)] - \mu_l(m)}{\sigma_l(m)} 
\]  
(5.20)

\[
q(m) = \frac{\max[D_m(\omega)] - \mu(m)}{\sigma(m)},
\]  
(5.21)

where $\mu_l(m)$ and $\sigma^2_l(m)$ are the mean and variance of the local maxima of the frame $D_m(\omega)$, not including the global maximum itself, and $\mu(m)$ and $\sigma^2(m)$ are the mean and variance of the entire frame $D_m(\omega)$. The parameters $p_0$, $p_1$, $q_0$, and $q_1$ in equation (5.19) are empirically determined minimum and maximum thresholds on the peakedness of the frame.

The fundamental frequency $f(m)$ is determined in tandem with the target voice activity detector, and feeds back to it. It is defined as

\[
f(m) = \begin{cases} 
\text{argmax}_{\omega} D_m(\omega) & v(m) = 1 \\
0 & v(m) = 0
\end{cases}
\]  
(5.22)

where $f(m) = 0$ indicates that the target voice was not detected and no fundamental frequency estimate could be made.

To satisfy the physical continuity and smoothness constraints of the fundamental frequency, the target fundamental frequency detector employs the following heuristics. First, it removes consecutive positive detections in the target voice activity detector that last less than a preset short duration, $\Delta_v$. This improves the quality of the enhanced signal by avoiding sudden bursts of energy in the output. Furthermore, the detector assumes that
the “initial” fundamental frequency estimate $f(m)$ for which $v(m) = 1$ and $v(m - 1) = 0$ is accurate. This is a reasonable assumption since the peakedness of the harmonic product spectrum at those instances must exceed the maximum peakedness thresholds. Hence, the $D_m(\omega)$ is strongly peaked at the initial fundamental frequency estimate, and $f(m)$ is a good estimate of the fundamental frequency. Given this assumption, the fundamental frequency detector recursively detects jumps larger than $\Delta_f > 0$ in $f(m)$,

$$j(m) = \begin{cases} 
1 & |f(m) - f(m - 1)| > \Delta_f \\
0 & \text{otherwise}
\end{cases}$$

(5.23)

When $j(m) = 1$ and $j(m - 1) = 0$ for $m = m_0$, the fundamental frequency detector tracks the duration of the jump. If the duration of the jump exceeds a predetermined threshold $\Delta_m > 0$, i.e.,

$$|f(m) - f(m_0)| > \Delta_f, \quad \text{for } m = m_0 + 1, \ldots, m_0 + \Delta_m,$$

(5.24)

the fundamental frequency detector considers the jump a loss of continuity and resets the target voice activity detector at $m = m_0$, i.e. $v(m_0) = 0$. If the fundamental frequency estimate $f(m)$ returns to within $\Delta_f$ of $f(m_0)$ at a time $m_1 < m_0 + \Delta_m$, the fundamental frequency detector considers the jump a temporary loss of smoothness and bridges the gap with a linearly interpolated fundamental frequency, i.e.,

$$f(m) = (1 - \lambda)f(m_0) + \lambda f(m_1), \quad m = m_0, \ldots, m_1,$$

(5.25)

where

$$\lambda = \frac{m - m_0}{m_1 - m_0}.$$  

(5.26)

The operation of the target fundamental frequency detector is illustrated by an example of its components in figure 5.8.
Figure 5.8: Components of the target fundamental frequency detector. From top to bottom: Sampled harmonic product spectrum $D(m, \omega)$; peakedness measures $p(m)$ and $q(m)$, with minimum and maximum threshold shown in red and green; target voice activity detection $v(m)$, shown before (gray) and after (blue) suppressing short bursts; fundamental frequency estimate $f(m)$, shown as $\operatorname{argmax}_\omega D_m(\omega)$ (gray) and final estimate (blue).
Carrier refinement

In practice, the target fundamental frequency \( f(m) \) is based on \( D(m, \omega_k) \), a version of \( D(m, \omega) \) that is sampled in frequency on a fixed grid,

\[
\omega_k = 2\pi \frac{F_h - F_l}{F_s} \cdot \frac{k}{K} + 2\pi \frac{F_l}{F_s}, \quad \text{for } k = 0, \ldots, K.
\] (5.27)

The carrier refinement stage increases the precision of the fundamental frequency estimates by reevaluating the harmonic product spectrum on a denser grid around the estimated fundamental frequency, \( \hat{D}[m, \omega_l(m)] \),

\[
\omega_l(m) = 2\pi \frac{F_h - F_l}{KF_s} \cdot \frac{2l - L}{L} + f(m), \quad \text{for } l = 0, \ldots, L.
\] (5.28)

The refined fundamental frequency estimate \( \hat{f}(m) \) is then the location of the maximum in the \( m \)-th frame \( \hat{D}_m[\omega_l(m)] = \hat{D}[m, \omega_l(m)] \), i.e.,

\[
\hat{f}(m) = \arg\max_{\omega_l(m)} \hat{D}_m[\omega_l(m)].
\] (5.29)

The carrier refinement stage also estimates the frequencies of the harmonics of the target’s speech. According to the STRAIGHT model of voiced speech in equation (5.17), the frequency of the \( k \)-th harmonic is defined as the harmonic multiple of the common fundamental frequency plus an individual harmonic deviation. Presently, we ignore the individual harmonic deviation and use only the harmonic multiple of the fundamental frequency as the harmonic frequency estimate, since this proved to be a more robust estimate of the harmonics in low target-to-interference ratios. Hence, the frequencies of the harmonics are estimated as

\[
\hat{f}_k(m) = k\hat{f}(m).
\] (5.30)

5.4.2 Modulation filter

The modulation filter of our approach to target talker enhancement is implemented as a bank of time-varying second order IIR resonator filters. Each resonator filter can be interpreted
as a lowpass modulation filter in a time-varying subband, where each subband is centered on a harmonic of the speech signal.

The idea of time-varying resonator filters is similar to the dynamic tracking filter originally propose by Roberts [115] for satellite communication systems, which was later redefined and extended by Kumaresan, Ramalingam and Rao [72] and Rao and Kumaresan [110, 111] for speech signals. The dynamic tracking filter is a complex-valued, single-pole filter, with system function

\[ H_n(z) = \frac{1 - re^{j\omega(n)}z^{-1}}{1 - r}. \]  

(5.31)

The resonance frequency \( \omega(n) \) of the dynamic tracking filter is dynamically updated with the first-order phase difference of the filter’s output, which causes the filter to follow the dominant component in its passband. To apply the dynamic tracking filter to multicomponent signals such as speech, Rao and Kumaresan [110] propose to use one dynamic tracking filter per component, and to precede each of them with a time-varying all-zeros filter that suppresses the other components.

Our implementation of the bank of resonator filters is different from the multicomponent dynamic tracking filter in two important aspects. First, we separate the tracking of a component from the filtering of the component. This allows us to use the target fundamental frequency estimator described in section 5.4.1. This estimator achieves greater noise robustness by exploiting the harmonic structure of voiced speech to track the fundamental frequency, rather than tracking each harmonic independently. Second, the bandwidth of our resonator filter increases proportionally to the resonating frequency, to accommodate the increasing bandwidth of the higher harmonics of speech.

Each resonator in the bank of filters is implemented as a second-order IIR filter [100]. The \( k \)-th resonator is defined by the time-varying difference equation

\[
y_k(n) = 2\gamma_k \cos(f_k(n))y_k(n - 1) - (2\gamma_k - 1)y_k(n - 2)
+ (1 - \gamma_k)[x(n) - x(n - 2)].
\]  

(5.32)

The gain term, \( \gamma_k = \frac{1}{1 + \beta_k} \), ensures that the filter has unit response at the resonance fre-
quency. It is defined in terms of the bandwidth factor, \( \beta_k = \frac{1}{3} \sqrt{3} \tan(B_k/2) \), which depends on the resonator bandwidth \( B_k \). The time-varying resonance frequency \( f_k(n) \) is obtained from the refined harmonic estimate \( \hat{f}_k(m) \). In practice, the refined estimates are often sampled at a lower rate than the signal’s time variable \( n \). In that case, \( \hat{f}_k(m) \) is upsampled to \( f_k(n) \) through simple linear interpolation. However, if \( \hat{f}_k(m) = 0 \), i.e., when no target voice activity was detected at time \( m \), the resonator frequencies \( f_k(n) \) are held at their last known good value until the next activity from the target voice is detected. That way, the resonator filters can adapt faster to a new resonance frequency in the target’s fundamental frequency range. The output of the filters at the times that no activity from the target voice was detected is suppressed in the mixing stage, which is described in section 5.4.3.

Given the output of each resonator filter, as defined by equation 5.32, the output of the multiresonator filterbank is defined as the sum of the individual resonators

\[
y(n) = \sum_{k=1}^{K} y_k(n).
\]  

(5.33)

The number of resonators in the filterbank is determined by the lower limit, \( F_l \), of the target talker’s fundamental frequency range, and by the cutoff frequency, \( F_{lp} \), of the lowpass filter used in the mixing stage. The number \( K \) should be large enough such that \( KF_l > F_{lp} \).

A time-varying multiresonator filterbank is distinctly different from an adaptive comb filter (for example [79, 96, 148]). The resonance frequencies of the multiresonator filterbank are free to vary independently, whereas the resonating frequencies of a comb filter are locked to the same fundamental frequency. Moreover, the multiresonator frequencies can take on any value, but the comb filter frequencies can only be integer subharmonics of the sampling frequency, i.e.,

\[
F_{\text{comb}} = k \frac{F_s}{K}, \quad k = 0, \ldots, K,
\]  

(5.34)

for integer \( K > 0 \). Finally, the bandwidths of the multiresonator filters are independent of each other, unlike the equal bandwidth of the comb filter’s “combs”.
5.4.3 Mixing

In the mixing stage of our novel target talker enhancement algorithm, the target voice activity detection, \( v(m) \), is combined with the output of the multiresonator filterbank, \( y(n) \), and mixed with the original signal, \( x(n) \), as follows. The input \( x(n) \) and the multiresonator filterbank output \( y(n) \) are filtered with a lowpass filter \( h_{lp}(n) \), resulting in the low-frequency signals

\[
x_{lp}(n) = x(n) * h_{lp}(n) \\
y_{lp}(n) = y(n) * h_{lp}(n),
\]

and their high-frequency counterparts

\[
x_{hp}(n) = x(n) - x_{lp}(n) \\
y_{hp}(n) = y(n) - y_{lp}(n),
\]

where we have assumed for convenience that \( h_{lp}(n) \) has zero phase. The lowpass filter’s cutoff frequency \( F_{lp} \) is chosen such that the higher harmonics of the target’s speech signal, which tend to drift away from an exact integer multiple of the target’s fundamental frequency, are suppressed. A typically value for \( F_{lp} \) is in the range of 1500–2000 Hz.

Furthermore, the voice activity signal \( v(m) \) is upscaled to the input signal’s sampling rate, if it is not already at that rate. The sudden onsets and offsets of the binary voice activity signal are replaced by short cosine-shaped ramps. The upsampled and smoothed voice activity signal, \( v(n) \), is then used to modulate the lowpass and highpass filtered input and filterbank signals, as follows:

\[
\tilde{x}(n) = \beta_1 v(n)y_{lp}(n) + \frac{1}{\beta_2} v(n)x_{hp}(n) + \frac{1}{\beta_3} (1 - v(n))x(n).
\]

The mixing constant \( \beta_1 > 1 \) amplifies the low-pass filtered output of the multiresonator filterbank when voiced speech from the target talker is detected. This enhances the harmonic structure of voiced speech from the target talker in low frequencies, while avoiding the
“metallic” artifacts commonly associated with an overly forced harmonic structure in high frequencies.

The second term of equation (5.39) controls the amount of high frequencies that are passed from the input $x(n)$ to the output $\tilde{x}(n)$ at times when voiced speech from the target talker is detected. At those times, the signal component $x_{hp}(n)$ contains the higher harmonics of the target talker’s speech, and potentially contains high frequencies from the interfering talkers. The mixing constant $\beta_2 > 1$ that attenuates this component is a compromise between attenuating the interfering talker to an acceptable level, while maintaining enough of the target talker’s higher harmonics. By passing the higher harmonics of the target talker unfiltered, we have found that much of the naturalness of the target’s voiced speech is preserved.

The mixing constant $\beta_3 > 1$ attenuates the input signal when no voiced speech from the target talker is detected. It is important for the intelligibility of the target talker to pass this component at a moderate level, because it contains the unvoiced parts of the target’s speech signal. It also helps to maintain the overall quality of the signal to pass this component to the output in attenuated form.

5.5 Experiment

The performance of the coherent modulation filtering approach to target talker enhancement described in section 5.4.1 was evaluated with a subjective listening on hearing impaired subjects and normal hearing subjects. This test was similar to the listening test that was used to evaluate the optimal coherent modulation filter approach described in section 5.2.2. The objective of the listening test was to measure the speech reception threshold of a target talker in two-talker interference under various processing conditions. Details of the test are given in section 5.5.1 through 5.5.4. The results and a discussion of the results are presented in section 5.5.5.

5.5.1 Stimuli

The spondee signals and the interfering talkers noise signal from the listening test described in section 5.2.2 were also used in this test; see that section for a complete description of
the signals. Each spondee signal was mixed with the interfering talkers noise signal at a signal-to-noise ratio (SNR) ranging from -50 dB to +20 dB in steps of 2 dB. The RMS amplitude of the spondee signals was kept constant in all mixtures, and the RMS amplitude of the noise signal was scaled to the desired SNR. The same interfering talkers noise signal was used in all mixtures.

5.5.2 Methods

In the subjective listening test, stimuli were presented in three processing conditions: (1) original, unprocessed stimuli; (2) target talker enhancement using coherent modulation filtering; and (3) target talker enhancement using coherent modulation filtering, with target talker detection and estimation done on the spondee without interfering talkers. In case of the hearing impaired subjects, presentation of the original stimuli over a hearing aid with noise reduction enabled was added as a fourth processing condition.

The original, unprocessed stimuli are used as the reference condition to which the other conditions are compared. Henceforth, we will refer to this condition as the “unprocessed” condition. The second condition, target talker enhancement using coherent modulation filtering, is the novel target talker enhancement approach described in section 5.4. The values of the parameters of this algorithm as they were used in the processing for the listening test are listed in table 5.5. We will refer to this processing condition as the “coherent” condition.

The third processing condition is the same as the second condition, except that voice activity of the target talker is detected and the frequencies of the harmonics of the target’s speech are estimated on the spondee signal alone, instead of on the mixture of spondee and interfering talkers noise signal. This condition was included as a control condition, to evaluate the performance of the novel coherent modulation filter independent from the target talker detector/estimator. We wished to evaluate the novel modulation filter in isolation, because we anticipated that the target detector/estimator would have to perform at very low signal-to-noise ratios in the subjective test, given the results of the previous spondee listening test. It can be expected that the target detector/estimator does not perform well
Table 5.5: Parameter values of the coherent modulation filtering algorithm as used in the subjective listening test.

<table>
<thead>
<tr>
<th>description</th>
<th>parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>target detector/estimator parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>target fundamental frequency range</td>
<td>$F = [F_l, F_h]$</td>
<td>[212, 250] Hz</td>
</tr>
<tr>
<td>number of harmonics</td>
<td>$P$</td>
<td>14</td>
</tr>
<tr>
<td>peakedness thresholds</td>
<td>${p_0, p_1, q_0, q_1}$</td>
<td>{2, 5, 2.5, 5}</td>
</tr>
<tr>
<td>voice activity smoothness</td>
<td>$\Delta_v$</td>
<td>20 ms</td>
</tr>
<tr>
<td>fundamental frequency smoothness</td>
<td>${\Delta_f, \Delta_m}$</td>
<td>{5 Hz, 20 ms}</td>
</tr>
<tr>
<td><strong>modulation filter parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of resonators</td>
<td>$K$</td>
<td>14</td>
</tr>
<tr>
<td>resonator bandwidths</td>
<td>$B_k$</td>
<td>$\frac{1}{2}(3k + 23)$ Hz</td>
</tr>
<tr>
<td><strong>mixer parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lowpass cutoff</td>
<td>$F_{lp}$</td>
<td>2000 Hz</td>
</tr>
<tr>
<td>mixing constants</td>
<td>${\beta_1, \beta_2, \beta_3}$</td>
<td>{2, 2, 2}</td>
</tr>
</tbody>
</table>

under those circumstances, since no fundamental frequency estimator has been shown to work well in SNRs below -20 dB. By including this condition in the test, we will be able to determine whether potential poor performance of the “coherent” condition is due to the extreme SNRs encountered in the listening test, or whether it is caused by the coherent modulation filter. We will refer to this third condition as the “coherent in quiet” condition.

For hearing impaired subjects, the above three conditions were presented over a hearing aid with noise reduction and directionality disabled. To compare our novel target talker enhancement approach to the noise reduction of the hearing aid, a fourth condition was included in the listening test. This condition, which we will refer to as “noise reduction,” uses the original, unprocessed stimuli, and presents them to the hearing impaired subject over a hearing aid with noise reduction enabled.

5.5.3 Subjects and presentation

Three bilateral hearing loss patients and six normal hearing subjects participated in the test. The normal hearing subjects volunteered their time, and the hearing impaired subjects were compensated for their time. Four normal hearing subjects were non-native English speakers.
Their understanding of the English language was more than sufficient to participate in this simple word-recognition test. All other subjects were native speakers of American English. Five normal hearing subjects had previous experience with auditory experiments, and two had prior exposure to the test material. All subjects were seated in a double-walled, sound attenuating booth.

The sounds were presented binaurally to the hearing impaired subjects through a loudspeaker. Each hearing impaired subject was using a single hearing aid that was programmed for their hearing loss. Directionality and noise reduction were turned off for all processing conditions, but noise reduction was turned on for the “unprocessed with noise reduction” test condition. The non-test ear of hearing impaired subjects was plugged off.

For the normal hearing subjects, all sounds were presented binaurally through a loudspeaker. The stimuli were processed with a 6-channel cochlear implant simulator before presentation. The simulator divides an input signal into subbands using a six-channel FIR filterbank whose subband frequency ranges are 80–308, 308–788, 788–1794, 1794–3906, 3960–8338, and 8338–17640 Hz. It combines the Hilbert envelope of each subband with a (broadband) white noise carrier, filters the resulting noisy envelope with the corresponding subband filter, and sums all subband outputs together to form a signal that simulates the output of a cochlear implant.

The loudness of the target talker spondees was calibrated to 55 dBA prior to the experiment, and was held at that level throughout the test.

5.5.4 Procedure

The subjective listening test consisted of a training phase and a testing phase. The training phase consisted of two training rounds to familiarize the subject with the signals. During a training round, each spondee (without interfering talkers noise) was presented twice to the subject, while a button labeled with the spondee word was highlighted on a monitor. The testing phase of the listening test was setup as an adaptive speech reception threshold (SRT), twelve alternative forced-choice test (12AFC) using a simple 1-up, 1-down method [77]. In each trial, a spondee was randomly selected out of the twelve spondees, and was presented in
noise at a certain SNR. The subject selected the spondee they heard from twelve buttons on a monitor which were labeled with the spondee words. The subject was required to respond in each trial. The SNR was decreased by 2 dB on a positive response, and increased by 2 dB on a negative response. This procedure was repeated until 14 reversals in SNR were completed. The mean of the SNR at the last 10 reversals was taken as the estimate of the 50% correct SRT on the psychometric curve [77].

The 50% correct SRT levels of all processing conditions were measured in random order. In case of hearing impaired subjects, the program controlling the subjective test prompted the operator to switch on the hearing aid’s noise reduction just prior to the “unprocessed with hearing aid noise reduction” processing condition, and again prompted the operator to switch it off immediately after it. The measurements were repeated six times for each subject, and the order of the processing conditions was randomized in each repetition. The adaptive procedure and user interface for the test were implemented in MATLAB.

5.5.5 Results

The results of the listening test are shown on page 130 for the hearing impaired subjects, and on page 131 and 132 for the normal hearing subjects. In figure 5.9, the estimated speech reception threshold (SRT) of each hearing impaired subject is plotted against repetition number for all processing conditions. There seems to be a downward trend in the SRT with increasing repetition number for all subjects and all processing methods, suggesting the presence of a learning effect. The first subject’s SRT for the coherent processing condition varies greatly (between -4 and -24 dB SNR), for an unknown reason. Furthermore, this subject appears to have no benefit from any type of processing and performs best in the unprocessed condition. The second subject’s SRTs are in the 0 to -20 dB SNR range. This subject appears to benefit from the coherent and the coherent in quiet processing, but has little or no benefit from the hearing aid noise reduction. The third subject’s SRTs are in the -20 to -40 dB SNR range. This subject appears to benefit from coherent in quiet processing, but not from hearing aid noise reduction or coherent processing.

The average estimated SRT per processing condition and repetition over all hearing
Figure 5.9: Estimated speech reception thresholds for the hearing impaired subjects of the listening test on coherent talker enhanced speech. Speech reception threshold (SRT), in dB, plotted against repetition number for each subject, and for four processing methods: unprocessed; noise reduction; coherent; coherent in quiet.

Table 5.6: Table of means for the hearing impaired subjects of the listening test on coherent talker enhanced speech. Average speech reception thresholds over all subjects by processing method and repetition number. Difference between methods can be seen by comparing rows, learning effect can be seen by comparing columns.

<table>
<thead>
<tr>
<th>processing</th>
<th>repetition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td>-15.7</td>
<td>-21.3</td>
<td>-23.5</td>
<td>-23.5</td>
<td>-22.7</td>
<td>-22.6</td>
<td>-21.6</td>
<td></td>
</tr>
<tr>
<td>noise reduction</td>
<td>-18.0</td>
<td>-18.2</td>
<td>-20.6</td>
<td>-20.8</td>
<td>-22.8</td>
<td>-23.0</td>
<td>-20.6</td>
<td></td>
</tr>
<tr>
<td>coherent</td>
<td>-8.8</td>
<td>-19.7</td>
<td>-17.3</td>
<td>-22.2</td>
<td>-19.9</td>
<td>-20.8</td>
<td>-18.1</td>
<td></td>
</tr>
<tr>
<td>overall</td>
<td>-16.7</td>
<td>-21.4</td>
<td>-20.7</td>
<td>-22.6</td>
<td>-22.7</td>
<td>-23.1</td>
<td>-21.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7: Repeated measures ANOVA for the hearing impaired subjects of the listening test on coherent talker enhanced speech. F-statistic and p-value of the within-subject factors processing and repetition, and their interaction. (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>processing</td>
<td>$F_{3,6} = 1.79$</td>
<td>0.25</td>
</tr>
<tr>
<td>repetition</td>
<td>$F_{5,10} = 7.45$</td>
<td>0.004 (**)</td>
</tr>
<tr>
<td>interaction</td>
<td>$F_{15,30} = 2.35$</td>
<td>0.023 (*)</td>
</tr>
</tbody>
</table>
Figure 5.10: Estimated speech reception thresholds for the normal hearing subjects of the listening test on coherent talker enhanced speech. Speech reception threshold (SRT), in dB, plotted against repetition number for each subject, and for three processing methods: unprocessed; coherent; coherent in quiet.

Table 5.8: Table of means for the normal hearing subjects of the listening test on coherent talker enhanced speech. Average speech reception thresholds over all subjects by processing method and repetition number. Difference between methods can be seen by comparing rows, learning effect can be seen by comparing columns.

<table>
<thead>
<tr>
<th>processing</th>
<th>repetition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td></td>
<td>-11.5</td>
<td>-14.3</td>
<td>-14.7</td>
<td>-14.2</td>
<td>-15.2</td>
<td>-18.1</td>
<td>-14.6</td>
</tr>
<tr>
<td>coherent</td>
<td></td>
<td>-11.2</td>
<td>-13.6</td>
<td>-16.7</td>
<td>-17.8</td>
<td>-17.9</td>
<td>-16.4</td>
<td>-15.6</td>
</tr>
<tr>
<td>coherent in quiet</td>
<td></td>
<td>-12.0</td>
<td>-17.7</td>
<td>-17.6</td>
<td>-18.8</td>
<td>-20.0</td>
<td>-21.0</td>
<td>-17.8</td>
</tr>
<tr>
<td>overall</td>
<td></td>
<td>-11.6</td>
<td>-15.2</td>
<td>-16.3</td>
<td>-16.9</td>
<td>-17.7</td>
<td>-18.5</td>
<td>-16.0</td>
</tr>
</tbody>
</table>
Table 5.9: Repeated measures ANOVA for the normal hearing subjects of the listening test on coherent talker enhanced speech. F-statistic and p-value of the within-subject factors processing and repetition, and their interaction. (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>processing</td>
<td>$F_{2,10} = 17.19$</td>
<td>$&lt; 0.001 (***)$</td>
</tr>
<tr>
<td>repetition</td>
<td>$F_{5,25} = 10.96$</td>
<td>$&lt; 0.001 (***)$</td>
</tr>
<tr>
<td>interaction</td>
<td>$F_{10,50} = 1.44$</td>
<td>0.19</td>
</tr>
</tbody>
</table>

impaired subjects are presented in table 5.6. Overall, the hearing impaired subjects received no benefit from hearing aid noise reduction (an increase in SRT by 1.0 dB SNR), and no benefit from coherent processing (an increase in SRT by 3.5 dB SNR). They did however benefit from coherent in quiet processing (a decrease in SRT by 2.9 dB SNR). The results of a repeated measures ANOVA that was performed on this data is shown in table 5.7. The analysis revealed that the effect of the processing method was not significant. Furthermore, the learning effect was very significant and the interaction between processing method and repetition number was significant.

In figure 5.10, the estimated speech reception threshold (SRT) of each normal hearing subject is plotted against repetition number for all processing conditions. Similar to the hearing impaired subjects, there seems to be a downward trend in the SRT with increasing repetition number for all subjects and all processing methods, suggesting the presence of a learning effect. The relative differences between the three processing conditions appears to be fairly consistent between subjects.

The average estimated SRT per processing condition and repetition over all normal hearing subjects are presented in table 5.8. Overall, the normal hearing subjects received benefit from the coherent processing (a decrease in SRT by 1.0 dB SNR) and from the coherent in quiet processing (a decrease in SRT by 3.2 dB SNR). The data also clearly shows the learning effect, as average SRT decreases monotonically with increasing repetition. The results of a repeated measures ANOVA that was performed on this data is shown in table 5.9. The analysis revealed that the effect of the processing method as well as the learning effect
were very significant, and that there was no significant interaction between them.

The results of the subjective listening test indicate that the novel coherent modulation filtering target talker enhancement technique may improve speech intelligibility. However, the number of participants in both subject groups is low, so firm conclusions should not be drawn based on this data. They merely suggest that the coherent modulation filtering approach may be the right direction of research, and that further testing is required. We plan to conduct a larger subjective test in the near future, and intend to include cochlear implant users in that test. We provide some suggestions in section 6.2 to change the test procedure that could improve the accuracy of the speech reception threshold estimates.

5.6 Summary

In this chapter we presented our research into target talker enhancement for hearing devices based on modulation analysis and filtering. We described two preliminary experiments that motivated us to pursue this direction of research. The first experiment, separation of two talkers with manually designed modulation masks, demonstrated to us that different talkers are separable in the modulation frequency domain. The second experiment, optimal coherent modulation filtering of speech, indicated that the potential for automatic talker separation and enhancement in the modulation frequency domain was large.

Next, we introduced our target talker enhancement technique based on incoherent modulation analysis and a time-varying (acoustic) filter. We used this technique to separate two talkers with non-overlapping fundamental frequency ranges, and presented the separation results. Based on these results, we discussed two important drawbacks of incoherent modulation analysis for talker separation, namely its limited ability to analyze time-varying signals such as speech with sufficient resolution, and the modulation detection errors due to FM-to-AM transduction in the filterbank.

Finally, we introduced our novel coherent modulation filtering approach to talker enhancement. The new technique is based on a bank of time-varying resonator filters, whose resonant frequencies are driven by a robustly estimated fundamental frequency signal.
Chapter 6

CONCLUSION

This chapter provides a summary of our research contributions in section 6.1, and presents some suggestions for future work in section 6.2.

6.1 Contributions

Our research contributions on the topic of modulation analysis and filtering can be divided into three categories: theory, practical aspects, and application to talker enhancement. In this section, we list our contributions to each of these categories. To the theory of modulation analysis and filtering systems, we have contributed the following items.

- We have developed a principled modulation analysis and filtering framework based on a modulation transform and an envelope detector or carrier estimator. Our modulation transform is completely defined in terms of existing Fourier-based transforms.

- We have defined the signal representations that can be obtained with the modulation transform, and declared consistent names for the representations and their magnitude and phase components.

- We have given the conditions that envelope detectors, carrier estimators and modulation filters must satisfy for proper invertibility of the modulation transform. We have demonstrated that practical modulation filters generally violate these conditions, and depend on the pseudo-invertibility of the modulation transform for signal reconstruction. These pseudo-inverses of the modulation transform introduce approximation errors in the reconstructed signal. We have further analyzed these signal reconstruction approximation errors in the evaluation of our proposed modulation transforms.
• We have shown that sampled versions of the modulation transforms can be generalized into a modulation filtering framework that is based on arbitrary filterbanks rather than a Fourier-based frequency analyzer. This generalized modulation filtering framework encompasses all modulation filtering systems defined in the literature, yet is subject to the theory of modulation transforms.

With respect to the practical aspects of modulation analysis and filtering, we have made the following contributions.

• We have critically analyzed commonly used incoherent envelope detectors such as the magnitude detector and the Hilbert envelope detector. We have motivated that, for artifact-free modulation filtering, the bandwidth of modulators and carriers must be of the same order as the bandwidth of the subband they are derived from. We have demonstrated that the incoherent envelope detectors do not satisfy this bandwidth constraint. Furthermore, we have shown additional weaknesses of the incoherent envelope detector, such as the unrealistic assumption that modulators have conjugate symmetric spectra.

• We have argued to overcome the limitations of the incoherent envelope detectors with coherent carrier estimators. By definition, these carrier estimators produce bandlimited modulators and carriers. They yield complex-valued modulators that can have arbitrary spectra, which better fits the modulators of natural signals such as speech and music.

• We have described three coherent carrier estimators: (1) our proposed smoothed Hilbert carrier estimator, which lowpass filters the Hilbert phase to satisfy the bandwidth constraint; (2) the instantaneous frequency carrier estimator proposed in [4], which defines the subband carrier as its bandlimited instantaneous frequency; and (3) our proposed frequency reassignment carrier estimator, which improves on the instantaneous frequency carrier estimator by eliminating the need for linear approximations and unwrapping of subband phase.
We have demonstrated the use of the coherent carrier estimators on modulation frequency analysis of amplitude-modulated sinusoids, amplitude-modulated white noise, and a speech signal. We have concluded that the smoothed Hilbert and the frequency reassignment techniques estimate the carriers and modulators more accurately on the sinusoids than the instantaneous frequency carrier estimator. None of the coherent methods is able to detect the modulators of the modulated white noise signal, which is a logical consequence of that signal’s noisy phase. Of the coherent estimators, the frequency reassignment technique estimates the carriers of the speech signal with the highest accuracy, and produces modulators that easily satisfy the bandwidth constraint.

We have evaluated and compared the incoherent and coherent modulation transforms on lowpass, bandpass and highpass modulation filters, using the weighted overlap-add (WOLA) and the least squared-error estimation on modified short-time Fourier transform magnitude (LSEE-MSTFTM) signal reconstruction techniques. Based on this evaluation, we have concluded that WOLA is a reasonable choice for the signal reconstruction technique after modulation filtering. Although the LSEE-MSTFTM technique is able to improve the signal-to-error ratio of the reconstructed signal by about 5 dB, it does so at a significant computational cost while also giving the reconstructed signal a “buzzy” sound quality. We have also demonstrated that the maximum effective suppression that any modulation filter achieves is about -20 dB for a modulation filter that was designed with -40 dB suppression in the stopband. We have determined that the proposed coherent carrier estimators are in general only more effective than the incoherent envelope detectors on lowpass modulation filters, and that they do not perform well on bandpass and highpass modulation filters. This was unexpected, as the coherent estimators satisfy the bandwidth constraint and other desirable properties of carrier estimators that we have identified.

Our research into practical modulation filtering systems has resulted in the modulation toolbox [119]. The modulation toolbox is set of MATLAB functions that implement the incoherent modulation transform. The toolbox contains a graphical user interface that is
capable of modulation frequency analysis of signals, as well as signal modification through modulation masking. A coherent version of the modulation toolbox is in preparation and will be made available to the academic community soon.

We have applied our theory and experience of modulation analysis and filtering to the problem of target talker enhancement in the presence of co-channel interfering talkers, which resulted in the following contributions.

- We have shown that the energy distributions of two co-channel talkers in the modulation frequency domain are largely non-overlapping. We were able to exploit this feature to achieve a 5 dB separation between the two talkers using an incoherent modulation masking technique with manually designed modulation masks.

- We have demonstrated the potential of modulation filtering approaches to talker separation using an optimal coherent modulation filtering technique. We have evaluated this technique in a subjective listening test on normal hearing and hearing impaired subjects, where it achieved a maximum improvement in speech reception threshold of 20 dB SNR. The approach, however, requires full knowledge of the desired signal, which is unavailable in practice.

- We have automated the manual masking modulation experiment into an incoherent modulation analysis approach to target talker enhancement, which achieves 4 dB separation between two co-channel talkers with sufficiently non-overlapping fundamental frequencies.

- We have motivated that the performance of the incoherent talker enhancement algorithm is bounded by a fundamental limitation of the wideband incoherent modulation spectrogram to resolve the time-varying fundamental frequency of a target talker. Moreover, FM-to-AM transduction in the filterbank causes subband modulation detection errors, which negatively affects the performance of the incoherent technique.

- We have determined that the existing envelope detectors and carrier estimators are
not sufficiently robust to noise from interfering talkers to be successful in the target talker enhancement problem at low target-to-interference ratios.

- We have proposed a novel coherent modulation filtering approach to the talker enhancement problem that overcomes the problems associated with the incoherent approach. Its carrier estimator is more robust to noise because it assumes a harmonic model of voiced speech and jointly estimates the carrier frequencies of the low-frequency harmonics. Moreover, its modulation filter can resolve carriers and modulators with high resolution in both time and frequency, and does not suffer from the same fundamental limitation in resolution as the incoherent modulation spectrogram.

- We have evaluated the novel coherent talker enhancement algorithm in a preliminary subjective listening test on normal hearing and hearing impaired subjects. We have tested both the originally proposed algorithm as well as a version of the algorithm where the carrier estimates were replaced by the true carriers, in order to evaluate the performance of the novel modulation filter in isolation. The results of the listening test showed that the coherent algorithm did not improve speech reception for the hearing impaired (an increase in SRT by 3.5 dB SNR), but that the modulation filter alone did improve speech reception by 2.9 dB SNR. A repeated measures ANOVA, however, indicated that these differences were not statistically significant, in part due to the small number of hearing impaired subjects that were tested. Additional testing with hearing impaired subjects is required before meaningful conclusions can be drawn.

Both versions of the algorithm did improve speech reception for the normal hearing listeners, who were tested over a cochlear implant simulation. The improvements were 1.0 dB and 3.2 dB SNR, respectively, and the repeated measures ANOVA on this data indicated that the signals processed with these algorithm were significantly different from the unprocessed signals.
6.2 Future work

The research on modulation analysis and filtering is not concluded. Based on our experience with modulation analysis and filtering techniques, as described in this dissertation, we envision the following directions for future research.

- The effective modulation frequency response of the currently defined incoherent and coherent detectors does not meet the designed frequency response of modulation filters. In our research, we continue to analyze this issue and strive to develop coherent envelope detectors with a more effective modulation frequency response.

- The ineffectiveness of modulation filters is caused in part by the reconstruction from a modified short-time Fourier transform, which is not well-defined for arbitrary modifications as was discussed in section 3.5.5. An interesting area of research would be to study the use of time-frequency transforms with better reconstruction properties than the short-time Fourier transform as a front-end to modulation filters. For example, we believe that modulated complex lapped transforms (see for example [82, 83, 158]) permit arbitrary modification of a critically sampled time-frequency representation of a signal without causing blocking artifacts upon reconstruction. Furthermore, the modulated complex lapped transform may also have better reconstruction properties after modulation filtering than the short-time Fourier transform for oversampled time-frequency representations of a signal, but additional research is required to confirm this statement.

- Coherent detectors use subband phase and its derivative to estimate carriers. From a practical standpoint, working with subband phase and derivatives is difficult. The use of subband phase could possibly be circumvented in a modulation filtering framework based on quadrature carriers. In such a framework, a complex-valued subband would be split into its real and imaginary components, and a different carrier and modulator would be estimated for each component, thus avoiding the need to work directly with subband phase. This could result in a more robust way to estimate carriers in practice.
A key problem that we observe with the current approaches to modulation analysis and filtering is that separating a broadband signal into narrowband frequency sub-bands with an LTI filterbank causes FM-to-AM transduction. When the subbands are processed independently, the FM-to-AM transduction effect results in errors in modulation detection, as was explained in section 5.3.4. A possible solution could be to jointly estimate carriers and modulators of neighboring subbands, or to jointly estimate them on the broadband signal as a whole. This is essentially the approach we have taken to carrier estimation in the target talker enhancement application, with positive results.

The higher harmonics of speech signals are usually less well defined and have more bandwidth than its lower harmonics. This makes it harder for carrier estimators to correctly estimate carriers in high-frequency subbands of speech signals. It could be worthwhile to investigate a hybrid modulation analysis and filtering system that uses coherent carrier estimation for the low-frequency subbands and incoherent envelope detection for the high-frequency subbands, or a scheme that integrates information from both incoherent detection and coherent detection over the entire frequency range.

To improve the performance of the target talker enhancement technique in the very low target-to-inference ratios that are encountered in the subjective listening test, it may be useful to estimate the parameters of the (stronger) harmonics of the interferer first, and to estimate the carriers and modulators of the target talker on the residual signal. Moreover, an even more robust approach would be to not apply a “winner takes all” strategy that estimates the carriers and modulators of each talker in order of their strength, but to estimate the carriers and modulators of all talkers simultaneously.

The novel coherent target talker enhancement technique has been demonstrated to work on spondee words from a female talker in the subjective listening experiment. However, the target detection and estimation algorithm should also be tested on and possibly extended to signals with more varying fundamental frequency, as spondee words tend to be spoken rather monotone. The target detection and estimation al-
algorithm should also be made more robust to interfering talkers with a fundamental frequencies range that overlaps more with the target’s fundamental frequency range. A possible approach could be to use dynamic programming strategies to better exploit the temporal smoothness of the target’s fundamental frequency trajectory and distinguish it from interfering talkers.

- The implementation of IIR filters, such as the resonator filters of the novel coherent target enhancement algorithm, typically requires special care to avoid unstable output signals. Furthermore, there do not always exist fast implementations of IIR filters, which means that they could place a high computational demand on the hearing device on which they are implemented. As an alternative, the resonator filters could perhaps be replaced by a large number of narrowband FIR subbands that are densely sampled in frequency. Such implementations are common in audio coding to find narrowband regions of high energy in time-frequency, similar to the function of the resonator filters in our application, while avoiding the practical complexities of IIR filters.

- For maximum benefit from our proposed target talker enhancement algorithm for cochlear implants, it can be expected that the target talker detector and the fundamental frequency estimator must be integrated with cochlear implant processing, rather than added to a cochlear implant as a pre-processing step. An interesting direction of research would be to determine how information from the coherently estimated carriers can be used effectively in the stimulation of electrodes in cochlear implants.

In addition to the suggestions for future research, we have several recommendations for a follow-up subjective listening test.

- According to Levitt [77], the initial SNR of the speech reception threshold (SRT) measurements should be set to the level of the expected SRT. This concentrates the subsequent trails around the expected SRT, and gives the most robust estimate of SRT. In our set-up of the listening test, the initial SNR was set to +10 dB, which is above the expected SRT for all subjects tested. Thus, several measurements and
some time were wasted on irrelevant SNRs. The impact on the results is negligible, as the first four reversals are discarded in the analysis. However, it may be a better practice to assess a subject's expected SRTs, and to initialize SNRs at that level.

- Some learning effect was observed in the results of our listening tests. This learning effect could possibly be reduced by randomly mixing trials from different processing techniques. In the current setup, all 14 reversals from one processing condition are finished before another processing condition is started. Within a processing condition, the subject receives subtle feedback about his performance by the decrease in SNR on a correct response and the increase in SNR on an incorrect response. By mixing trials from several processing conditions, the response in one trial affects the SNR of the next trial of the same processing condition, which may be several trials removed from it. This randomization probably decouples the subject’s response from the feedback sufficiently to reduce the learning effect significantly. Unfortunately, such a mixing scheme would be cumbersome when testing hearing impaired subjects on more than one hearing device, because it would require the subject to switch hearing devices between trials rather than between processing conditions.

- Some subjects were able to distinguish the target at very low and somewhat unrealistic signal-to-noise ratios, mainly because the listening test was setup as a closed-set word recognition task. To make the subjective test more representative of the real-world problem that we attempt to solve, the test could be converted to an open-set word recognition task, or a speech reception threshold higher than 50% correct could be measured. For example, by measuring the 75% correct threshold, more measurements are taken at higher SNRs, which are more realistic to be encountered in environments that the novel coherent modulation filtering target talker enhancement algorithm was designed to operate in.

- In our listening test, we observed differences between processing conditions in the order of the 2 dB step-size between successive trials. It may therefore be advisable to
use a smaller step size towards the end of each SRT measurement, in order to get a more accurate estimates of the average SRT of each processing condition.

With this dissertation, it is our hope that we have shed some light on the intricacies of modulation analysis and filtering systems, and that we have provided a useful theoretical framework and practical signal processing tools that inspire others to consider modulation frequency analysis and modulation frequency signal representations for their research problems. Furthermore, we hope that we have contributed a fresh look at the cocktail party problem from a modulation filtering perspective.
BIBLIOGRAPHY


[40] D. W. Griffin, D. S. Deadrick, and J. S. Lim, “Speech synthesis from short-time Fourier transform magnitude and its application to speech processing,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 8, 1984, pp. 2.4.1–2.4.4. 4.6.3


[107] ———, “AM-FM separation using auditory-motivated filters,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 465–480, 1997. 2.3.1, 5.3.4


[124] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in International Conference on Spoken Language Processing (INTERSPEECH), 2006. 2.4.1


Z. Shpiro and D. Malah, “An algebraic approach to discrete short-time Fourier transform analysis and synthesis,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1984, pp. 2.3.1–2.3.4. 4.6.1, 4.6.1

P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007. 2.4.1


VITA

Steven Marco Schimmel was born in Apeldoorn, The Netherlands, on July 27, 1974, the son of Ria and Berd Schimmel. After completing his VWO education at the Christelijk Lyceum Apeldoorn in 1992, he attended Delft University of Technology, Delft, The Netherlands, from which he graduated in 2001 with a Masters of Science degree in Computer Science. From 2000 until 2002 he worked at Philips Research, Eindhoven, The Netherlands, first on his Master’s thesis project on motion-sensitive video watermarking, later on the Philips audio fingerprinting project. In 2002, he moved to Seattle, Washington to enter graduate school at the University of Washington. He graduated in 2007 with a Doctor of Philosophy degree in Electrical Engineering. He will continue his research into speech and music signal processing with application to hearing devices in a post-doc position at the University of Zürich, Switzerland, in the group of Prof. Dillier. He will move to Zürich with his partner, Marloes Maathuis.