Harmonic Encoding in Cochlear Implants

Tyler Ganter

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Electrical Engineering

University of Washington

2016

Reading Committee:

Les Atlas, Chair

Jay Rubinstein

Kaibao Nie

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Harmonic Encoding in Cochlear Implants

Tyler Ganter

Chair of the Supervisory Committee:
Professor Les Atlas
Electrical Engineering

This document is about extracting harmonic envelopes, what matters, what doesn't and how to design your system accordingly. It is broken into three parts:

- envelope extraction techniques and their relationships
- phase preservation
- system design (filter and downshift)

Many strategies consider the effects of leaving modulations in the signal, but nothing really talks about what the envelope should be, independent of the modulations. If we do this first, we can than think about how the modulations affect this envelope as a separate modulation component.

If explicitly inducing modulations it is important to remove any other modulations, and this is how.

TABLE OF CONTENTS

Page

List of Figures . . . iii

Chapter 1: Introduction . . . 1
- 1.1 Overview . . . 1
- 1.2 Survey of Literature . . . 1
- 1.3 Contents of Thesis . . . 3
- 1.4 Notational Conventions . . . 3

Chapter 2: Cochlea Implant Processing . . . 4
- 2.1 DSP Algorithms . . . 10

Chapter 3: Envelope Extraction Methods . . . 16
- 3.1 Incoherent Methods . . . 16
- 3.2 Coherent Methods . . . 20
- 3.3 Coherent Angle Encoding . . . 22
- 3.4 The Relationships . . . 26
- 3.5 Summary . . . 36

Chapter 4: Harmonic Envelopes . . . 37
- 4.1 Steady-State Analysis . . . 38
- 4.2 Steady-State Metrics . . . 46
- 4.3 Induced VS Explicit Temporal Modulation . . . 47
- 4.4 Time-Varying F0 . . . 51
- 4.5 Transients . . . 51
- 4.6 Evaluation of Strategies . . . 52
- 4.7 Non-ideal Pitch Estimators . . . 67

Chapter 5: Conclusion . . . 75
5.1 Mapping and Selection . . . 75
5.2 Implementation Considerations . . . 76
5.3 Summary . . . 78
5.4 Future Work . . . 78

Appendix A: Derivations . . . 79

LIST OF FIGURES

Figure Number Page

2.1 CA vs CIS 5

2.2 Channel Vocoder Processing 7

2.3 Signal Flow in CI 9

2.4 Tranformation from DSP output to Electrical Signal 9

2.5 ACE Flow Diagram 11

2.6 condensed ACE Flow Diagram 11

2.7 F0mod Flow Diagram 12

2.8 Induced vs Explicit Temporal Modulations 13

2.9 HSSE Flow Diagram 14

2.10 HSSE (with phase) Flow Diagram 15

3.1 Incoherent vs Coherent Processing 17

3.2 Magnitude Only VS Coherent Angle Encoding Block Diagrams 22

3.3 Cello Example 25

3.4 Method Comparison: magnitude spectrum at each step 35

4.1 Magnitude of spectrum for equations 4.5 - 4.8 39

4.2 $(a)|\widehat{X}[n,f)|$ $(b)|\widehat{X}[n,f-2F_0)|$ $(c)|\widehat{X}[n,f-2F_0)||H_1(f)|$ $(d)|\widehat{X}^*[n,-f+2F_0)||H_1(-f)|$ $(e)|\mathcal{F}\{\tilde{m}_1^2[n]\}|$ (f) contributions of separate components of (e) 42

4.3 Envelope Estimate $-2F_0$ Component 44

4.4 Envelope Estimate $-F_0$ Component 44

4.5 Envelope Estimate Baseband Component 44

4.6 Comparison of ideal and non-ideal downshift frequencies, faded lines represent ideal $(a)|\widehat{X}[n,f)|$ $(b)|\widehat{X}[n,f-F_{DS})|$ $(c)|H(f)|$ $(d)|\widehat{X}[n,f-F_{DS})||H(f)|$ $(d)|\widehat{X}^*[n,-f+F_{DS})||H(-f)|$ $(e)|\mathcal{F}\{\tilde{m}_0^2[n]\}|$ 45

4.7 Clarinet vs Saxophone Harmonic Components 49

4.8 Clarinet vs Saxophone Envelope Estimates 50

4.9 G_k Downshift Quatization 53

4.10 G_k filter order 54
4.11 G_k filter design 55
4.12 G_k variation across harmonics 56
4.13 SIR_k filter order and quantization 57
4.14 SIR_k filter design 58
4.15 SIR_k variation across harmonics with hanning filter 59
4.16 SIR_k variation across harmonics with rectangular filter 60
4.17 $D_{k,i}$ filter design and order 61
4.18 $D_{k,i}$ downshift quantization 62
4.19 $D_{k,i}$ at rate of iF_0 62
4.20 $D_{k,1}$ across harmonics 63
4.21 Transient Step Response, order = 128, 256, 512 (increasing order corresponds to longer reponse time) 64
4.22 Transient Input/Output Change 65
4.23 Transient Distortion for Common Instruments 66
4.24 SIR_k, hanning filter and pitch estimate quantization 70
4.25 SIR_k, adaptive filter and pitch estimate quantization 70
4.26 SIR_k, hanning filter, pitch estimate quantization and ±5Hz estimation error 71
4.27 SIR_k, adaptive filter, pitch estimate quantization and ±5Hz estimation error 71
4.28 $D_{k,1}$, hanning filter and pitch estimate quantization 72
4.29 $D_{k,1}$, hanning filter and pitch estimate quantizationr 73
4.30 $D_{k,1}$, hanning filter, pitch estimate quantization and ±5Hz estimation error . 73
4.31 $D_{k,1}$, hanning filter, pitch estimate quantization and ±5Hz estimation error . 74

ACKNOWLEDGMENTS

I to express sincere appreciation to Les Atlas, Kaibao Nie and Jay Rubinstein for their guidance. I would also like to thank Brad Ekin, Scott Wisdom, Tommy Powers, Eldridge Alcantara and Dave Dolengewicz for the insightful discussions. I couldn't have accomplished any of this without the professors and staff at the University of Washington that have supported me all along

.

DEDICATION

to G-ma

Chapter 1

INTRODUCTION

Goal: How do we encode harmonics in CIs?

this is the introduction

Why harmonic encoding? Help differentiate signals, (timbre), improve SIN performance, free up channels for other information

1.1 Overview

we are considering what is the ideal matched filter, and how close of an approximation do we need?

"By definition, timbre is the perceptual attribute that distinguishes two sounds that have the same pitch, loudness, and duration (American National Standards Institute, 1973)."

need motivation for why encoding harmonics:

1) more truthful/natural/data-efficient representation with limited tools

2) mitigate artifacts

The human ear has orders of magnitude more filters than ACE, (roughly 1500/22 I think).

HSSE could simulate this higher resolution by choosing different filter center frequencies based on the input signal

1.2 Survey of Literature

Equivalent noise bandwidth (ENBW) considers BW of noise if squished into a box of gain 1 around the downshift frequency. [windows for harmonic analysis] This isn't entirely applicable since our harmonic has BW ¿ epsilon, and since for any window most of the energy is close to 0, most of the so-called noise is actually desired harmonic signal. If this were not the case, (I think) rectangular window would be the best, but since it distributes the noise more heavily to higher frequencies away from zero, it is actually worse (higher sidelobes)

“some windows have a high rate of sidelobe decay that allows minimizing the error due to interference. However the steeper the sidelobe decay the wider the main lobe width and then the worse the minimum resolution bandwidth.” [An Intelligent FFT-Analyzer with Harmonic Interference Effect Correction and Uncertainty Evaluation]

“For NH listeners, the timbre space was best represented in three dimensions, one correlated with the temporal envelope (log-attack time) of the stimuli, one correlated with the spectral envelope (spectral centroid), and one correlated with the spectral fine structure (spectral irregularity) of the stimuli. The timbre space from CI listeners, however, was best represented by two dimensions, one correlated with temporal envelope features and the other weakly correlated with spectral envelope features of the stimuli. “temporal envelope is dominant cue for timbre perception in CI listeners” [Temporal and Spectral Cues for Musical Timbre Perception in Electric Hearing]

Hypothesis: –temporal envelope (log-attack time) this is in some cases smeared in time (F0mod) and in other cases mixed across harmonics –one correlated with the spectral envelope (spectral centroid) this is not as clearly represented as it could be (are we talking about resonance or per-harmonic details such as clarinet?) –one correlated with the spectral fine structure (spectral irregularity) this manifests in the envelope for CI processing, the problem though is that it is blurred across harmonics so the noise-like characteristics will be smoothed.

Search this thing: modulation transfer function JH goldwyn a point process framework for modeling electrical stim of the auditory nerve f

“bowed string tones are inharmonic during both their attack and decay (Beauchamp, 1974)”

“frequencies in the range of 80-300 Hz encompassing F0 for nearly all adult males and many females and children.”

“wideband vs feature extraction” [F0F2-F0F1F2] not sure what wideband implies, alternatively use temporal envelope cues.

"Another school of thought was based on speech production and perception, in which spec- tral peaks or formants, which reflect the reso- nance properties of the vocal tract, are extracted and delivered to different electrodes according to the presumed tonotopic relationship between the place of the electrode and its evoked pitch." F0F2, F0F1F2, MPEAK

what is important? hearing for any general reason...safety, functionality speech recognition what is important and lacking? music appreciation tonal language SiN quality

1.2.1 Other Strategies

any hybrid considerations? maybe hint at hsse ace hybrid

talk about unmentioned methods (AB, MedEl)

1.3 Contents of Thesis

...

1.4 Notational Conventions

...

Chapter 2
COCHLEA IMPLANT PROCESSING

Human hearing is tonotopic, that is, starting in the cochlea and through the rest of processing in the brain, sounds far apart in frequency are processed separately. The cochlea is spatially arranged; As a sound propagates through the basilar membrane the different frequencies are amplified or suppressed such that they stimulate locations physically far apart in the cochlea.

In a cochlear implant an array of electrodes is inserted into the cochlea. This array is intentionally designed to have a tonotopic organization. When current is sent to the most deeply inserted (apical) electrodes, neurons associated with low frequency sounds are stimulated. Conversely, current at a basal electrode will stimulate neurons associated with high frequencies.

Early cochlear implant strategies, under the category compressed-analog (CA), delivered band-specific analog signals to each electrode. By using bandpass filters and an electrode array the implant emulates the tonotopic organization of acoustic hearing.

Current processing strategies use feature extraction to achieve much higher performance on speech recognition. From each bandlimited signal a slow-time-varying envelope is extracted and the extra information is discarded.

These envelopes are amplitude compressed and then used to modulate continuous bipolar pulse trains on each electrode channel.

These strategies all stem from an original parent, continuous-interleaved-sampling (CIS). CIS is a solution to the problem of electric field interaction. By interleaving pulse-trains there is minimal interaction between electrodes.

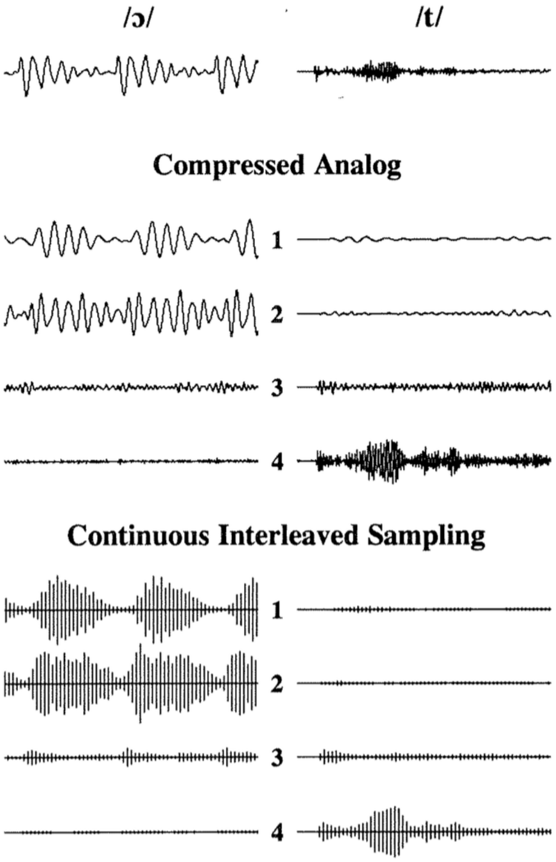

Figure 2.1: CA vs CIS

2.0.1 Sum-of-Products Model

We have now laid out enough background information to introduce a mathematical model for audio signals called the sum-of-products model.

Our digitally sampled audio signal $x[n]$, $(n \epsilon \mathbb{Z})$, is composed of bandpass components $x_k[n]$. In each bandpass component a slow-time-varying envelope $m_k[n]$ multiplies a quickly-oscillating carrier $c_k[n]$.

$$x[n] = \sum_k x_k[n] = \sum_k m_k[n] c_k[n] \tag{2.1}$$

Although there are infinite ways to decompose a signal into a sum-of-products, the model stems from real-word signals. To gain some intuition consider, for example, a voiced vowel. The vocal tract can be thought of as the carriers, $c_k[n]$. Without changing the position of the mouth, one can change the pitch of a note. The mouth then acts as the envelope, $m_k[n]$. As the mouth changes shapes it changes the formant structure structure. Equivalently, it changes the relative amplitude of each bandpass component $x_k[n]$.

As another example we may consider musical instruments. The pitch is characterized by the carriers but the timbre which is predominantly characterized by the attack time and spectral centroid [kong 2011] will be encoded by the rise time and relative amplitude of the envelopes.

2.0.2 Why Envelopes?

One of the motivations for this approach is the limited ability to perceive temporal modulations in electric hearing. In acoustic hearing modulations up to a few kHz may be perceptible, however cochlear implant envelope extraction techniques are designed to limit modulations, typically to around 160 to 320 Hz, which is closer to the range perceptible in electric hearing.

Modulation rates are also limited by pulse rate. Although there isn't a quantitative value analogous to Nyquist rate, modulations at rates higher than a certain percentage of the constant pulse rate will not be represented accurately by the modulated pulse train. That being said, cochlear implants today support modulations typically upwards of 2000pps (pulses per second), which should be sufficient provided modulations limited to 320Hz.

2.0.3 The Channel Vocoder

To gain some intuition as to how and why CIS processing works we consider a closely related system, the channel vocoder. Vocoding is a method of signal analysis and synthesis initially designed for audio data compression in telecommunication. As of the mid 70's the vocoder has gained widespread familiarity via the music industry as a funky voice effect. It is most well known for the signature robot voice heard in hits such as Kraftwerk's song "The Robots" or Styx's "Mr. Roboto". In its application to music, the vocoder extracts the bandlimited envelopes of one source (typically vocal) and applies them to each bandlimited components of a second source.

What's interesting is that this second source can be essentially any arbitrary broadband signal and yet we still understand speech from the first source. In this way the vocoder acts as a form of lossy data compression; the low data-rate envelopes are extracted and they may be later applied to, for example, white-noise.

This tells us that speech information is predominantly contained in the bandlimited envelopes, and thanks to the incredible robustness of speech to distortion, an estimated envelope is sufficient for speech comprehension.

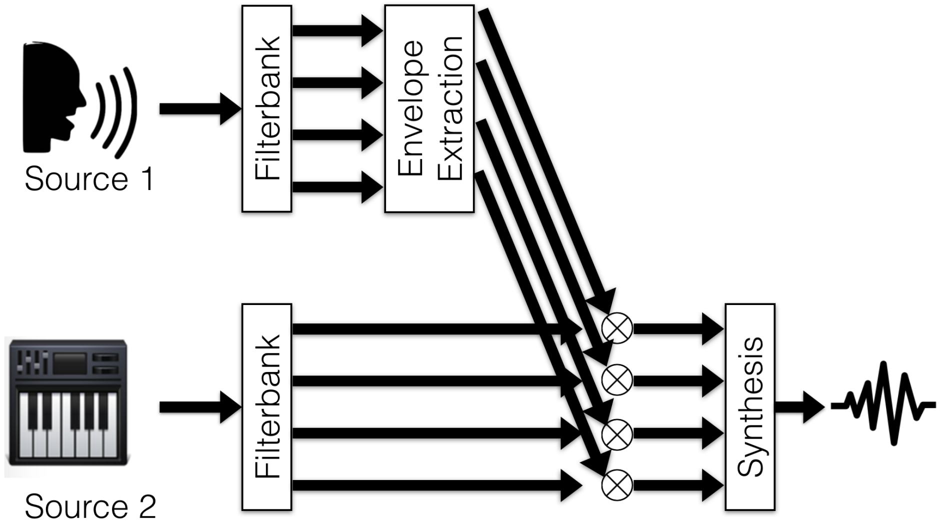

Figure 2.2: Channel Vocoder Processing

It should be noted that the second source is typically chosen to be a broadband stationary signal. If the signal is non-stationary it will have time-varying envelopes of it's own which will interact with the envelopes of the first source. Referencing back to our sum-of-products model the second source acts as a combination of carriers, $c_k[n]$, and the first source's envelopes, $m_k[n]$, are applied.

$$s^1[n] = \sum_k s_k^1[n] = \sum_k m_k^1[n]c_k^1[n] \tag{2.2}$$

$$s^2[n] = \sum_k s_k^2[n] = \sum_k m_k^2 c_k^2[n] \tag{2.3}$$

$$y[n] = \sum_k m_k^1[n] m_k^2 c_k^2[n] \tag{2.4}$$

Linking back, cochlear implant envelope extraction strategies do the same thing as vocoder signal analysis, as seen in figure 2.2, however rather than using a second source to synthesize a new sound, the envelopes directly modulate electrical pulse trains.

2.0.4 Temporal Fine Structure

The major drawback to this method of encoding is the loss of temporal fine structure. Referencing back to our sum-of-products model, we are extracting the envelopes and discarding carrier information.

When using a vocoder, vocals sung at different pitches general roughly the same output, $y[n]$. Similarly in cochlear implants temporal fine structure that encodes pitch, as well as other signal characteristics, is lost in processing.

“In most existing clinical sound processors, fine structure in the input acoustic signal is discarded, and only envelope information is preserved. ”

2.0.5 Processing Blocks

Let’s consider the processing blocks of a cochlear implant. The main stages to processing in cochlear implants are visualized in figure 2.3 below. While at every stage adjustments can be made, for the purpose of comparing DSP algorithms, all other stages will be assumed constant throughout this work unless otherwise specified.

In this document, the output of the DSP stage will be a strictly positive signal used to amplitude modulate a constant bipolar pulse train. T/C Level Mapping refers to a logarithmically-compressed mapping from amplitude to current level.

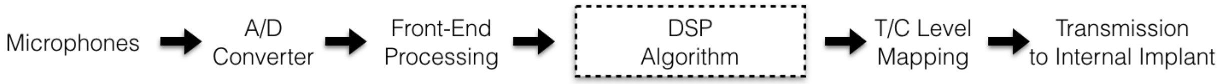

Figure 2.3: Signal Flow in CI

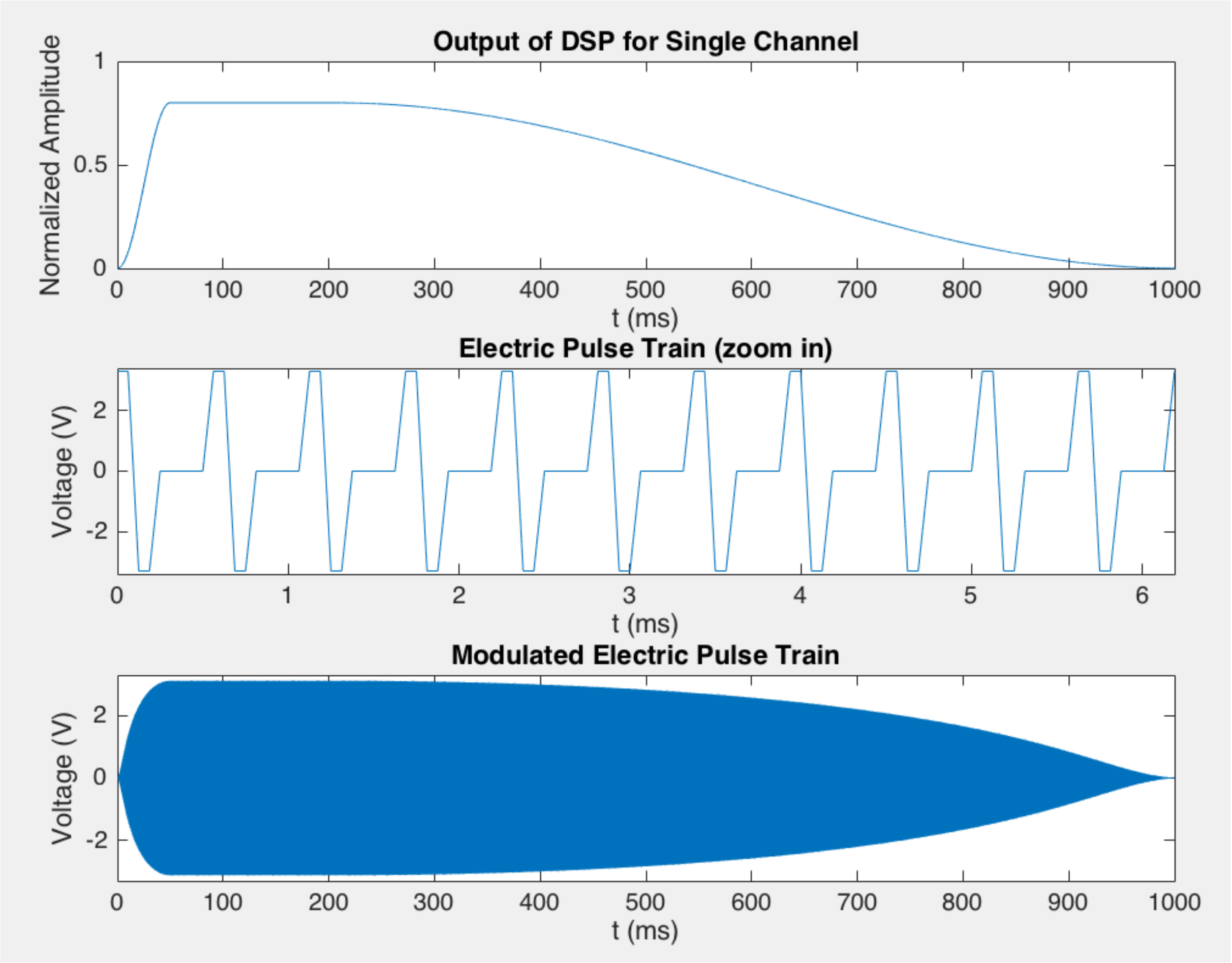

Figure 2.4: Tranformation from DSP output to Electrical Signal

2.0.6 Recap

To conclude up to this point, cochlear implants use an array of tonotopically orgranized electrodes. On each electrode a electric pulse train is transmitted and that pulse train is modulated by an envelope corresponding to a bandpass signal component

2.1 DSP Algorithms

To gain insight into how we can encode harmonic signals, in this section we will look inside the ”DSP Algorithm” box; we will compare three specific strategies, ACE, F0mod, and HSSE with the goals of evaluating the pros and cons of each and considering how to optimize performance for harmonic encoding.

2.1.1 ACE

The simplest of the considered strategies is the Advanced Combination Encoder (ACE). ACE has become a clinical standard for CI processing and is used in a vast number of users.

ACE is Cochlear Ltd’s instance of the auditory community’s generalized category of N-of-M strategies. In these strategies, K envelopes are first extracted then allocated to M channels corresponding uniquely to electrodes. During each processing frame a subset N-of-M channel envelopes is selected for stimulation on the internal implant. In the case that more than one envelope is allocated to a channel, the allocation stage must make a decision to select or combine envelopes in some way.

K - number of envelopes per frame

M - number of electrode channels

N - number of electrodes stimulated per frame

$$K \geq M \geq N$$

The following figure is simply a condensed version of the previous flow diagram. This condensed notation will be carried through to the other strategies analyzed.

While ACE does a sufficient job for many CI users in speech recognition tasks, a large gap remains between normal hearing and cochlear implants in many tasks such as pitch discrimination. This is largely attributed to the lack of temporal fine structure information

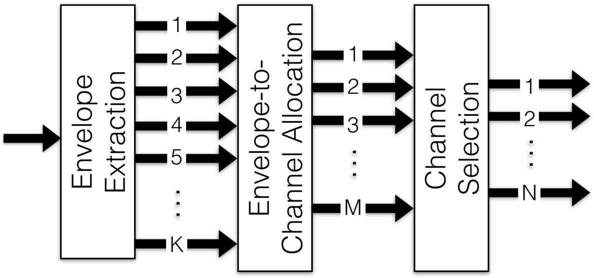

Figure 2.5: ACE Flow Diagram

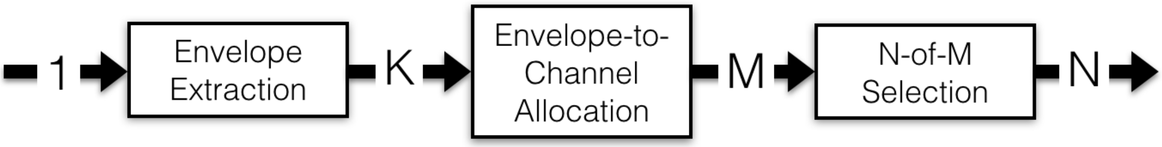

Figure 2.6: condensed ACE Flow Diagram

in this envelope encoding strategy.

ACE does, however, provide limited temporal modulations via beat frequencies. Through intentional processing artifacts, beat-frequencies will be induced in the processing of harmonic signals at a rate of the difference between the two harmonic frequencies, i.e. F_0. Typically these modulations are not full depth [ref?] and are usually limited to under 250Hz [ref?].

In this document will will refer to these artifact based modulations as induced modulations. Looking at the flow diagram of figure 2.6 it is not apparent that temporal modulations are contained in the processing path, however these modulations are encoded in the envelope itself. We can think of this as our extracted envelope containing some information about the carriers, $c_k[n]$.

Induced modulations are complementary to explicit modulation, used in F0mod and HSSE. Explicit modulations are those extracted from the signal separate from envelopes, and later applied to the final outputs.

2.1.2 F0mod

To get at the problem of pitch discrimination, (Laneau et al 2006) developed a new research strategy, F0mod. F0mod provides the same processing as ACE with one important change, explicit carrier modulation. It achieves this by adding a pitch estimator into the processing.

Once a fundamental frequency (F_0) is acquired, all output envelopes are modulated by a raised sinusoid at a rate of F_0, 2.5. F_0 is used because high modulation rates (typically above 300Hz) are not noticeable with a CI.

$$c_{ch}(t) = 0.5 + 0.5cos(2\pi F_0 t) \tag{2.5}$$

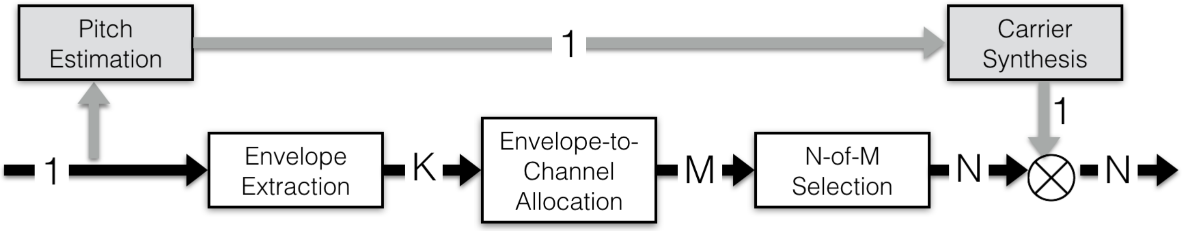

Figure 2.7: F0mod Flow Diagram

This raised sinusoid is constant modulation depth, (full dynamic range), and same across channels, (phase aligned). An example comparing this to induced modulations is shown in figure 2.8.

F0mod has shown promising results in acute tests for pitch discrimination. It has also inspired other processing strategies such as eTone, which uses a more sophisticated harmonic sieve pitch estimator as well as soft decisions to overcome the problem of encoding both harmonic and inharmonic sounds as well as those that fall somewhere in between.

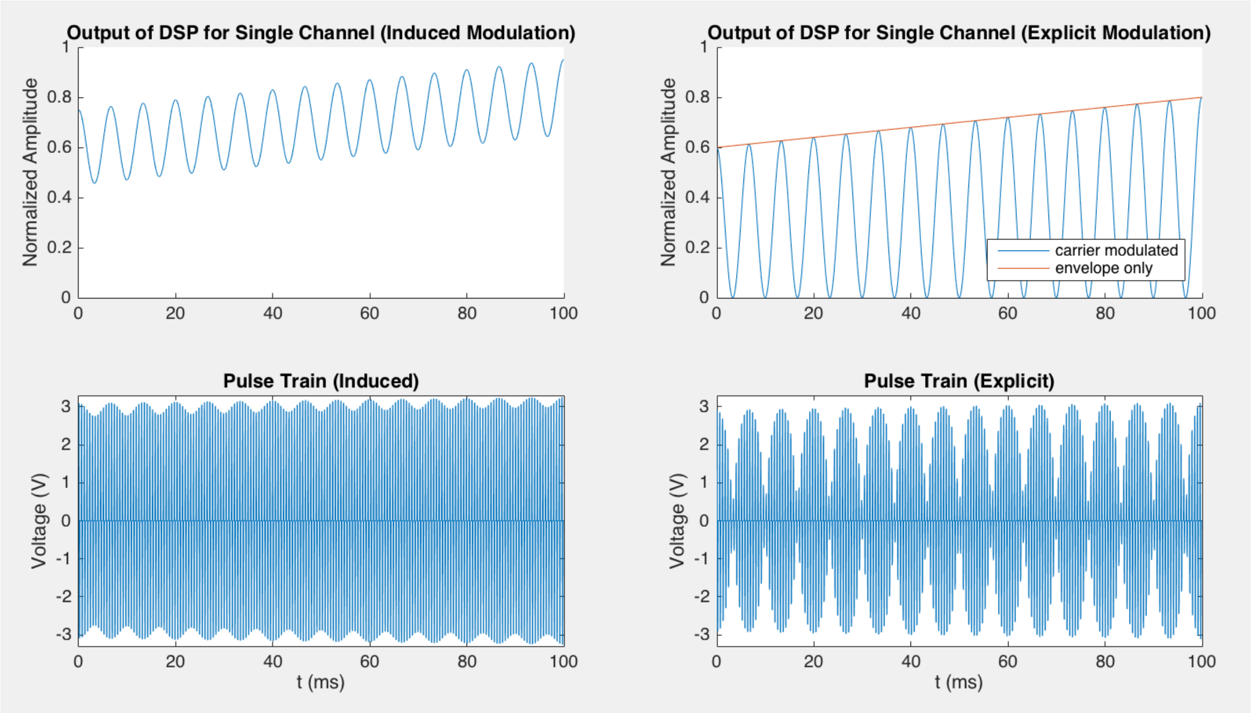

Figure 2.8: Induced vs Explicit Temporal Modulations

2.1.3 HSSE

Looking for a novel approach to improved pitch perception and more broadly music perception, (Li, Atlas, Nie) came up with Harmonic Single Sideband Encoder (HSSE).

There are two different versions of HSSE. We will start with the version most similar to F0mod.

In this version, coherent demodulation extracts harmonic envelopes. These harmonic envelopes are then combined into channels based on the harmonic index and F_0. Just as in F0mod a subset is selected for stimulation and then these envelopes are combined with carrier modulators.

$$K, M \geq N$$

The key differences between this and F0mod can be summarized quite simply: every

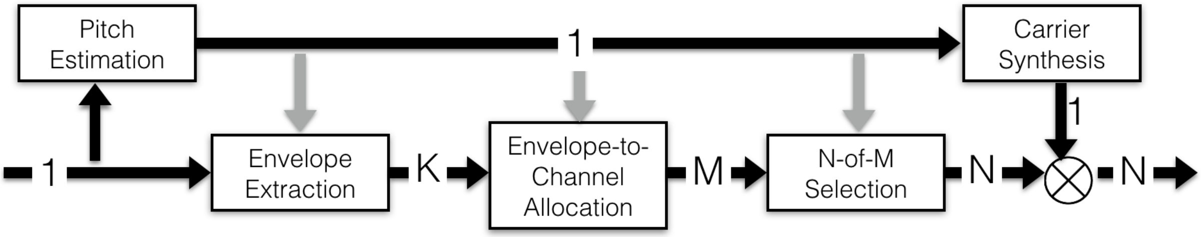

Figure 2.9: HSSE Flow Diagram

stage of typical ACE processing is now done coherently using F_0 information.

It should be noted that it is not necessarily true that $K \geq M$. In the case that no envelopes are allocated to a channel we may simply rule out that channel during the selection stage.

In the second version, more information about the carriers is retained. This put's some restrictions on the type of carrier than can be used, however it encodes time varying phase information which is unique to each envelope.

Because of the unique characteristics of each carrier, the carrier synthesis stage must be moved to an earlier point in the processing stage. First, complex envelopes containing phase information are extracted. These envelopes are then combined with a common carrier at a rate of F_0 however each output, which we will call a modulator, will be unique and time-varying in both magnitude and phase.

2.1.4 Summary

Comparing these strategies, the differences may be summarized as:

1) Envelope Extraction Method

- not discussed yet

2) Temporal Encoding Method

a) induced vs explicit

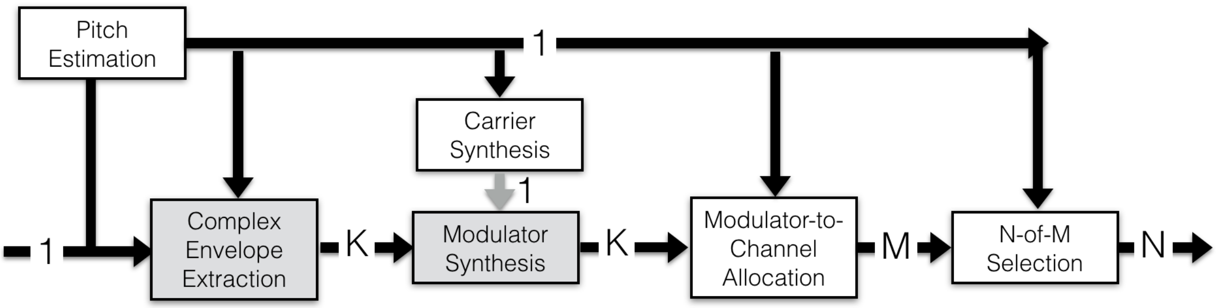

Figure 2.10: HSSE (with phase) Flow Diagram

b) phase preservation (explicit only)

c) modulation waveform (explicit only)

3) Envelope-to-Channel Allocation and Channel Selection

We will start by investigating 1 and 2(a,b). Some considerations for 2(c) and 3 will be brought up upon concluding this document, however the primary focus will be on 1 and 2(a,b).

The chapter 3 we will discuss mathematical methods to envelope extraction as well as phase preservation since phase is extracted at the same time. As a result we will generalize 1 and we will answer 2(b).

In chapter 4 we will evaluate design considerations for 1 and in doing so we will answer 2(a).

When we conclude we will briefly discuss 2(c) and 3.

Chapter 3
ENVELOPE EXTRACTION METHODS

In this chapter we will define the specific mathematical operations used to extract bandlimited time-varying envelopes. These methods fall under a general signal processing category of analysis-synthesis systems. In these methods a signal is decomposed into it's envelopes and carriers. Then the envelopes and/or carriers are manipulated individually before recombination.

One of the major focuses of research in this area is the evaluating the amount of distortion induced by the system. For example, Ghitza's test is a way of measuring the out-of-band distortion of a modulation filtering system. [REF]

Cochlear implant processing in unique in that the final output is not an audio signal. What this means is we only do the first half of the processing, the analysis step. This is critical to understand when considering methods, as all of the considerations related to synthesis or full-system distortion are no longer relevant.

This chapter is organized as follows. We will first introduce the envelope extraction methods to be considered. These methods are broken into to categories: incoherent and coherent. We will then take a quick detour to consider the efficacy of coherent angle encoding. Finally we will compare the methods and for a generalization of envelope extraction.

3.1 Incoherent Methods

The difference between incoherent and coherent is actually quite simple. Consider a system $H_k\cdot$. If this system is time-invariant then it is incoherent. If it is time-varying and the way in which it varies is a function of the input, we call this a coherent system. This is visualized in figure 3.1. In coherent methods the input not only passes through the system, it changes the system.

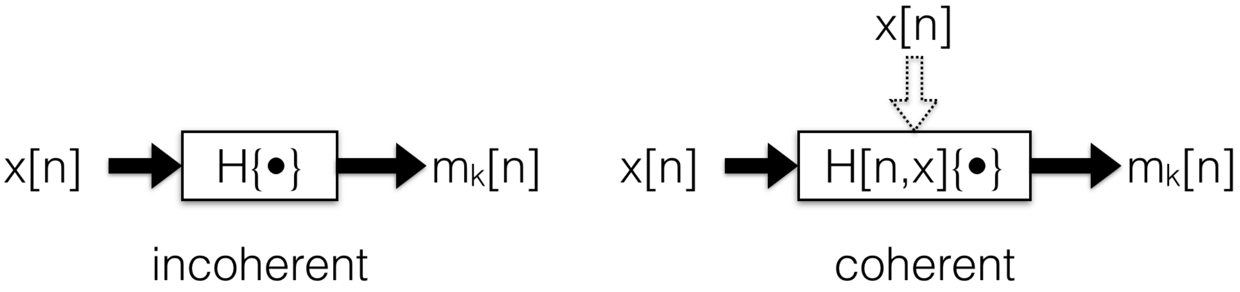

Figure 3.1: Incoherent vs Coherent Processing

In all considered methods, the input is a real digitally sampled audio waveform, $x[n]$ bounded in the normalized range of $[-1, 1]$. In the case of incoherent methods, the output will be K real digital waveforms, $m_k[n]$, in the range $[0, 1]$. All filters considered will be finite impulse response (FIR).

3.1.1 Continuous Interleaved Sampling (CIS)

This is method specifically implemented by the CIS strategy. We first bandpass filter the signal, where h_k is a bandpass filter and k has arbitrary limits. Then we full-wave rectify (take the magnitude) and finally lowpass filter that output. $h_{lp}[n]$ is a lowpass filter, typically with a cutoff around 200-400Hz.

$$m_{k,CIS}[n] = \Big| x[n] * h_k[n] \Big| * h_{lp}[n] \tag{3.1}$$

In this method the number of filters is usually the same as the number of electrodes (8-22). These filters are thus non-uniform bandwidth and center frequency, with increasing bandwidth and wider spacing at higher frequencies.

3.1.2 Hilbert Envelope

The Hilbert Envelope is method of decomposition applied far more broadly than the field of cochlear implants. Despite only retaining the envelope, we look at the carrier just to gain insight into how the signal $x[n]$ is represented in the decomposition. We first acquire the bandpass signal, $x_k[n]$. We then define our envelope as the magnitude of the analytic signal, acquire via a Hilbert transform, denoted $\mathcal{H}\{\cdot\}$.

$$\begin{aligned} x_k[n] =& x[n] * h_k[n] && (3.2)\\ \widehat{x}_k[n] =& x_k[n] + j\mathcal{H}\{x_k[n]\} && (3.3)\\ m_{k,hilbert}[n] =& \Big|\widehat{x}_k[n]\Big| && (3.4)\\ c_{k,hilbert}[n] =& cos(\angle\widehat{x}_k[n]) && (3.5) \end{aligned}$$

We can see intuitively that if the filterbank $[h_1[n], h_2[n], ...]$ has a flat total response, that all of the information of the original signal is contained in the envelopes and carriers, and thus we should be able to reconstruct the input from these components.

3.1.3 Short Time Fourier Transform (STFT)

The short-time Fourier transform (STFT) has many applications not associated with envelope extraction, however through analysis we will see that it fits the sum-of-products model we are looking for.

The STFT has two classic interpretations: a series of windowed Fourier transforms (each at a different time instant) or a collection of uniform bandpass filters (each at a different center frequency). For our purposes we will be using the later.

An STFT bin at discrete time n and discrete frequency k is defined as

$$X[n,k] = \sum_{r=-\infty}^{\infty} x[r]w[r-n]e^{-j\frac{2\pi}{N}kr}, \qquad 0 \le k < N \tag{3.6}$$

where N is the FFT order. Defining a new variable $r' = r - n$ and defining our window such that $w[n] = 0$ for $n < 0$ or $N \leq n$,

$$\begin{aligned} X[n,k] =& \sum_{r'=0}^{N-1} x[n+r']w[r']e^{-j\frac{2\pi}{N}k(n+r')} \\ =& e^{-j\frac{2\pi}{N}kn} \sum_{r'=0}^{N-1} x[n+r']w[r']e^{-j\frac{2\pi}{N}kr'} \end{aligned} \tag{3.7}$$

Let $X[n,k]$ be represented in polar form as the following

$$X[n,k] = |X[n,k]|e^{j\angle X[n,k]} \tag{3.8}$$

If we assume that the window $w[n] \neq 0$ for $0 \leq n \leq N-1$ then we have the inverse

$$\begin{aligned} x[n+r'] =& \frac{1}{Nw[r']} \sum_{k=0}^{N-1} X[n,k]e^{j\frac{2\pi}{N}k(n+r')} \\ =& \frac{1}{Nw[r']} \sum_{k=0}^{N-1} |X[n,k]|e^{j(\frac{2\pi}{N}k(n+r')+\angle X[n,k])} \end{aligned} \tag{3.9}$$

$$x[n] = \sum_{k=0}^{N-1} \frac{1}{Nw[0]} |X[n,k]|e^{j(\frac{2\pi}{N}kn+\angle X[n,k])} \tag{3.10}$$

Without loss of generality we can use a STFT hop-factor of one sample. In the case of a greater hop factor we would need to compute $x[n]$ from equation 3.9 [REF?] for some samples. Of course, if the hop factor is greater than N we cannot fully reconstruct the signal. This is especially noted because we will be recurrently the factor $w[0]$.

We can now clearly see our sum-of-products model

$$m_{k,STFT}[n] = \frac{1}{Nw[0]}|X[n,k]| \tag{3.11}$$

$$c_{k,STFT}[n] = e^{j(\frac{2\pi}{N}kn+\angle X[n,k])} \tag{3.12}$$

We can think of the STFT as a series of N LTI systems that each downshift the input signal, then lowpass filter. This can be seen mathematically if we rewrite equation 3.6 as

$$\begin{aligned} X[n,k] = &\sum_{r=-\infty}^{\infty} x[r]e^{-j\frac{2\pi}{N}kr}w[-(n-r)] \\ =&x[n]e^{-j\frac{2\pi}{N}kn} * w[-n] \end{aligned} \tag{3.13}$$

We can now look at the STFT envelope in a similar form to the other methods by plugging **??** into 3.11.

$$m_{k,STFT}[n] = \frac{1}{Nw[0]}\left|x[n]e^{-j\frac{2\pi}{N}kn} * w[-n]\right|, \qquad 0 \leq k \leq \frac{N}{2} \tag{3.14}$$

Also note that due to symmetry of the Fourier transform, envelopes are only valid for indicies between 0 and $\frac{N}{2}$.

3.2 Coherent Methods

Due to their LTI nature, incoherent methods fail to explicitly represent time varying characteristics like fundamental frequency or formant structure. [**?**] Alternatively, coherent methods will adapt to represent some specific characteristic.

3.2.1 Spectral Center-of-Gravity

One coherent method is the spectral center-of-gravity (COG). Similar to the previously described incoherent methods, spectral COG uses a fixed number of filters. The key difference lies in the center frequency of each of these filters which adapt over time as a function of the spectral distribution within predefined band limits.

Spectral COG certainly has some advantages of better representation of the signal in comparison to incoherent methods, however it still doesn't escape the limitation of fixed and pre-determined band limits that each filter operates within. We won't be investigating this method further.

3.2.2 Harmonic

To escape this, [Atlas and Others] proposed a harmonic method which uses knowledge of the structure of common audio signals to decompose the signal in a less arbitrary way. The first step is to get a pitch estimate $F_0[n]$ of the signal. We then define k complex carriers where there is a hard limit as a function of Nyquist sampling rate, $k \leq \lfloor \frac{F_s}{2F_0} \rfloor$

$$c_{k,harmonic}[n] = e^{jk\phi_0[n]} \tag{3.15}$$

where

$$\begin{aligned} \phi_0[n] =& \frac{2\pi}{F_s} \sum_{p=0}^{n} F_0[p] \\ =& \phi_0[n-1] + 2\pi \frac{F_0[n]}{F_s} \\ \phi_0[-1] =& 0 \end{aligned} \tag{3.16}$$

[modulation toolbox]

As mentioned earlier there are two versions of HSSE, the first uses a real non-negative envelope, the other uses a complex envelope.

we then define our first envelope

$$\begin{aligned} m^1_{k,harmonic}[n] =& \Big| x[n] c^*_{k,harmonic}[n] * h\big[n, F_0[n]\big] \Big| \\ =& \Big| x[n] e^{-jk\phi_0[n]} * h\big[n, F_0[n]\big] \Big| \end{aligned} \tag{3.17}$$

where $h\big[n, F_0[n]\big]$ is a lowpass filter that may vary as a function of $F_0[n]$. Note that we could have a different LPF for each k however since our carriers are linearly spaced it is natural to keep $h\big[n, F_0[n]\big]$ consistent over k.

Our second, complex envelope is the same as the first but without the final magnitude operation.

$$m^2_{k,harmonic}[n] = x[n]e^{-jk\phi_0[n]} * h\big[n, F_0[n]\big] \tag{3.18}$$

3.3 Coherent Angle Encoding

As mentioned earlier, the final DSP output is a set of real non-negative signals. We take a short aside to compare the two coherent harmonic methods, one of which, due to it's complex output, cannot be considered in a envelope-only sense.

The two alternative versions are visualized in figure 3.2. For the case of magnitude only, we can think of this as a restriction on our carrier. Since the envelope is already real non-negative the $Re\cdot$ and half-wave rectification stages don't change anything. We could pass complex exponential through these two operations before multiplying the envelope. This is equivalent to saying our carrier is a half-wave rectified sinusoid and thus we have the same general processing blocks as a single envelope of 2.9

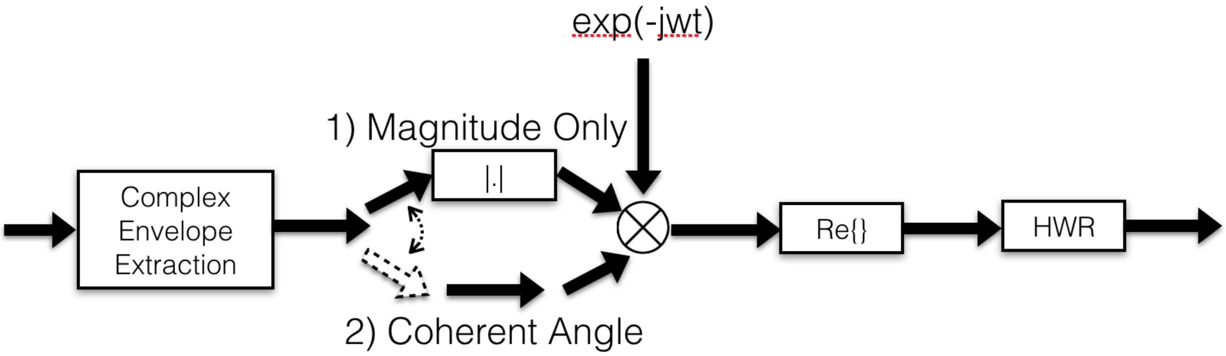

Figure 3.2: Magnitude Only VS Coherent Angle Encoding Block Diagrams

Let's consider a signal where our kth bandpass component represents the kth harmonic and is of the form

$$x_k[n] = A_k[n] cos(2\pi k F_0 n + \phi_k[n]) \tag{3.19}$$
$$BW \leq F_0$$

where $A_k[n]$ represents a real nonnegative amplitude. And BW is the bandpass signal's bandwidth. We may assume $F_0[n] = F_0$ is constant without loss of generality so long as $F_0[n]$ is roughly constant within each processing frame.

Let's assume our filter is an ideal brick-wall filter.

$$h[n] \Longleftrightarrow H(e^{j2\pi f}) \tag{3.20}$$
$$H(e^{j2\pi f}) = 1, \quad |f| < \frac{F_0}{2} \tag{3.21}$$
$$= 0, \quad else$$

Our coherent harmonic envelopes for each method will be

$$m^1_{k,harmonic}[n] = A_k[n] \tag{3.22}$$
$$m^2_{k,harmonic}[n] = A_k[n] e^{j\phi_k[n]} \tag{3.23}$$

Let us define $Rect\{y_k[n]\}$ as the half-wave rectified carrier-modulator signal which is our end goal. Using our first harmonic method

$$y^1_k[n] = m^1_{k,harmonic}[n] cos(2\pi F_0 n) \tag{3.24}$$
$$= A_k[n] cos(2\pi F_0 n)$$

Alternatively, with our second method we get

$$\begin{aligned} y_k^2[n] =& Re\{2m_{k,harmonic}^2[n]e^{j2\pi F_0 n}\} \\ =& Re\{2A_k[n]e^{j(2\pi F_0 n+\phi_k[n])}\} \\ =& A_k[n]cos(2\pi F_0 n + \phi_k[n]) \end{aligned} \tag{3.25}$$

It is clear that the difference between $y_k^1[n]$ and $y_k^2[n]$ is simply the extra term, $\phi_k[n]$. What this means may be best shown by example.

In figure 3.3 we see that when taking the magnitude, we force symmetry about 0. We see that the green much better represents the blue than the red does by preserving the spectral asymmetries that manifest themselves in the angle, not magnitude. It is unnatural and certainly won't happen in real world scenarios that a subband signal will be symmetric about the downshift frequency, however magnitude only methods force this to be true.

3.3.1 Appropriate Scaling

Despite better representing the signal, there is still an issue with $y_k^2[n]$. A more correct method is actually

$$m_{k,harmonic}^3[n] = A_k[n]e^{j\frac{1}{k}unwrap(\phi_k[n])} \tag{3.26}$$

$$\begin{aligned} y_k^3[n] =& Re\{2m_{k,harmonic}^3[n]e^{j2\pi F_0 n}\} \\ =& A_k[n]cos\Big(2\pi F_0 n + \frac{1}{k}unwrap(\phi_k[n])\Big) \end{aligned} \tag{3.27}$$

Why do we need the $\frac{1}{k}$ term? Let's consider an example where our true pitch estimate is actually $F_{0,groundtruth} = F_0 + F_{err}$. So,

$$x_k[n] = A_k[n]cos\Big(2\pi k(F_0 + F_{err})n + \phi_k[n]\Big) \tag{3.28}$$

In this case

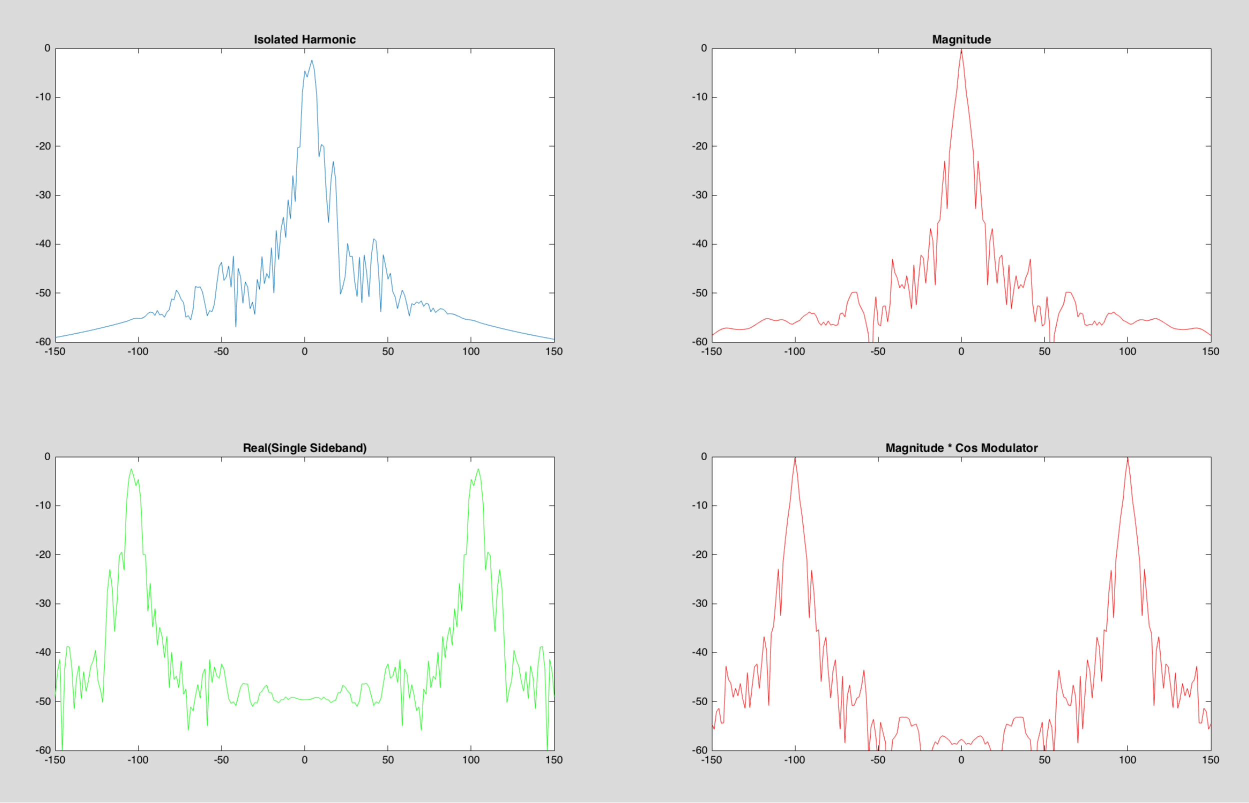

Figure 3.3: Cello Example

$$y_k^2[n] = A_k[n]cos\Big(2\pi(F_0 + kF_{err})n + \frac{1}{k}unwrap(\phi_k[n])\Big) \tag{3.29}$$

$$y_k^3[n] = A_k[n]cos\Big(2\pi(F_0 + F_{err})n + \frac{1}{k}unwrap(\phi_k[n])\Big) \tag{3.30}$$

Essentially the term $\phi_k[n]$ may be thought of as the deviation from kF_0. If we downshift the signal such that kF_0 is scaled to F_0 then it is appropriate that we scale $\phi_k[n]$ similarly.

3.3.2 Efficacy

Let us now consider the efficacy of 3.22 versus 3.26.

One hypothesis is that $\phi_k[n]$ may encode the noise-like characteristics of a signal, in which case it would remain constant for a pure sinusoid and fluctuate randomly for noise. Put to

Method	$m_k[n] =$
CIS	$\Big\lvert x[n] * h_k[n] \Big\rvert * h_{lp}[n]$
Hilbert	$\Big\lvert \widehat{x}_k[n] \Big\rvert = \Big\lvert x[n] * h_k[n] + j\mathcal{H}\{x[n] * h_k[n]\} \Big\rvert$
STFT	$\frac{1}{Nw[0]} \Big\lvert x[n]e^{-j\frac{2\pi}{N}kn} * w[-n] \Big\rvert$
Harmonic Coherent	$\Big\lvert x[n]e^{-jk\phi_0[n]} * h\big[n, F_0[n]\big] \Big\rvert, \qquad \phi_0[n] = \frac{2\pi}{F_s} \sum_{p=0}^{n} F_0[p]$

Table 3.1: Envelope Extraction Methods

test, the harmonic phase preservation did little to affect the signal and this was confirmed by testing varying filter bandwidths as well. In comparison of a toy experiment, the choice of filter bandwidth dominated noise-like qualities, with wider bandwidth capturing more of the variations.

Since the term $\phi_k[n]$ does not distinguish noise-like signals from narrowband sinusoidal signals, it is only really preserving phase alignment. But this begs the question, what does it mean to preserve the phase of a harmonic when downshifted to F_0? It is questionable as to whether this even has any logical meaning.

Furthermore, it has been suggested in [REF F0mod and kaibao??] that phase alignment is important for pitch perception in CIs. By using a magnitude-only method we guarantee alignment across channels.

Having considered this option as not a path worth further investigating, for the rest of this document we will consider envelope and carrier separately with each being a real nonnegative signal at the final output.

3.4 The Relationships

All of our methods are summarized in table 3.1. We will now consider the relationships between each of these.

3.4.1 Hilbert VS STFT

Let us start by comparing the Hilbert and STFT methods. Since "the Hilbert transform of a convolution is the convolution of the Hilbert transform on either factor" [wikipedia] we have

$$\begin{aligned}\widehat{x}_k[n] =& x_k[n] + jH\{x_k[n]\}\\ =& x[n] * h_k[n] + jH\{x[n] * h_k[n]\}\\ =& x[n] * h_k[n] + x[n] * jH\{h_k[n]\}\\ =& x[n] * [h_k[n] + jH\{h_k[n]\}]\end{aligned} \tag{3.31}$$

Now let us define our filter specifically as

$$h_k[n] = \frac{1}{Nw[0]} w[-n] cos(\frac{2\pi}{N} kn) \tag{3.32}$$

If we assume the sidelobes of $w[n]$ roll-off sufficiently fast in relation to the center-frequency $\frac{2\pi k}{N}$, we may approximate

$$\begin{aligned}\mathcal{H}\{h_k[n]\} \approx& \frac{1}{Nw[0]} w[-n] H\{cos(\frac{2\pi}{N} kn)\}\\ =& \frac{1}{Nw[0]} w[-n] sin(\frac{2\pi}{N} kn)\end{aligned} \tag{3.33}$$

Plugging our filter 3.32 into 3.31

$$\begin{aligned}
\widehat{x}_k[n] \approx& x[n] * \frac{1}{Nw[0]} w[-n] e^{j\frac{2\pi}{N}kn} \\
=& \frac{1}{Nw[0]} \sum_{r=-\infty}^{\infty} x[n-r] w[-r] e^{j\frac{2\pi}{N}kr} \\
\text{Let} \quad r' = -r & \\
=& \frac{1}{Nw[0]} \sum_{r'=0}^{N-1} x[n+r'] w[r'] e^{-j\frac{2\pi}{N}kr'} \\
=& \frac{1}{Nw[0]} \left[e^{-j\frac{2\pi}{N}kn} \sum_{r'=0}^{N-1} x[n+r'] w[r'] e^{-j\frac{2\pi}{N}kr'} \right] e^{j\frac{2\pi}{N}kn} \\
=& \frac{1}{Nw[0]} X[n,i] e^{j\frac{2\pi}{N}kn} \\
=& \left(\frac{1}{Nw[0]} x[n] e^{-j\frac{2\pi}{N}kn} * w[-n] \right) e^{j\frac{2\pi}{N}kn} \qquad (3.34) \\
& \qquad (3.35)
\end{aligned}$$

When we take the magnitude the complex exponential term goes to 1 and we are left with the STFT envelope. We come to the conclusion that under the assumption of fast sidelobe rolloff we may define a filter bank of $\frac{N}{2}+1$ filters

$$h_k[n] = w[-n] cos(\frac{2\pi}{N}kn), \qquad 0 \leq k \leq \frac{N}{2} \tag{3.36}$$

such that

$$m_{k,hilbert}[n] \approx m_{k,STFT}[n] \tag{3.37}$$

What this tells us is that the Hilbert decomposition may be viewed as a superset of the STFT method that is not constrained to uniform bandwidth linearly spaced filters.

3.4.2 STFT vs Harmonic

Let us now consider the relationship between STFT and harmonic coherent. We may choose our filter to be time-invariant and define it as

$$h\big[n, F_0[n]\big] = \frac{1}{Nw[0]} w[-n] \tag{3.38}$$

where $w[n]$ is a lowpass filter and

$$\begin{aligned} w[n] \neq & 0, \qquad 0 \leq n < N \\ = & 0, \qquad else \end{aligned} \tag{3.39}$$

In this case,

$$\begin{aligned} m_{k,harmonic}[n] =& \Big| x[n]e^{-jk\phi_0[n]} * \frac{1}{Nw[0]} w[-n] \Big| \\ =& \frac{1}{Nw[0]} \Big| x[n]e^{-jk\phi_0[n]} * w[-n] \Big| \end{aligned} \tag{3.40}$$

This bears striking resemblance to equation 3.14. We can see that in the case that $F_0[n] = \frac{F_s}{N}$,

$$m_{k,harmonic}[n] = m_{k,STFT}[n] \tag{3.41}$$

More generally, for any window of time n to $n + N - 1$ where $F_0[n]$ is constant

$$\begin{aligned} m_{k,harmonic}[n] =& \frac{1}{Nw[0]} \Big| X\Big[n, \frac{N}{1} \frac{F_0[n]}{F_s} k\Big) \Big| \\ =& \frac{1}{Nw[0]} \Big| X\Big[n, \lambda[n]k\Big) \Big| \end{aligned} \tag{3.42}$$

where $\lambda[n] = \frac{N}{1} \frac{F_0[n]}{F_s}$. The ")" is to denote that the frequency term is not necessarily an integer.

It is important to note that in practice $\lambda[n]$ is not a continuous variable. It is constrained by the quantization of the implemented pitch tracker. Provided this quantization we may compute any term $X[n, \lambda[n]k)$ by, leaving all else the same, zero-padding our FFT.

What this tells us is that in practice, we may approximate $m_{k,harmonic}[n]$ using $F_0[n]$ and a zero-padded STFT under the assumptions:

1) $F_0[n]$ is quantized

2) $F_0[n]$ is roughly constant withing a time window of $\frac{N}{Fs}$ seconds

and the restriction:

3) $h\Big[n, F_0[n]\Big]$ is time-invariant, i.e. $h\Big[n, F_0[n]\Big] = h[n]$

3.4.3 CIS VS Hilbert

Provided our envelope definitions

$$\begin{aligned} m_{k,CIS}[n] =& \Big|x_k[n]\Big| * h_{lp}[n] \\ m_{k,Hilbert}[n] =& \Big|\widehat{x}_k[n]\Big| \end{aligned}$$

We define an ideal brick-wall filter as

$$H_k(f) = \mathcal{F}\Big\{h_k[n]\Big\} \tag{3.43}$$

$$H_k(f) = 1, \quad f_k - \frac{1}{2}f_{bw} < |f| < f_k + \frac{1}{2}f_{bw} \tag{3.44}$$

$$= 0, \quad \text{else} \tag{3.45}$$

$$X(f) = \mathcal{F}\Big\{x[n]\Big\} \tag{3.46}$$

$$X_k(f) = \mathcal{F}\Big\{x_k[n]\Big\} \tag{3.47}$$

$$\widehat{X}_k(f) = \mathcal{F}\Big\{\widehat{x}_k[n]\Big\} \tag{3.48}$$

$$X_k(f) = X(f), \quad f_k - \frac{1}{2}f_{bw} < |f| < f_k + \frac{1}{2}f_{bw} \tag{3.49}$$

$$= 0, \quad \text{else} \tag{3.50}$$

$$\widehat{X}_k(f) = X(f), \quad f_k - \frac{1}{2}f_{bw} < f < f_k + \frac{1}{2}f_{bw} \tag{3.51}$$

$$= 0, \quad \text{else} \tag{3.52}$$

$$Y_k^1(f) = \mathcal{F}\Big\{\Big|\widehat{x}_k[n]\Big|^2\Big\} \tag{3.53}$$

$$Y_k^2(f) = \mathcal{F}\Big\{\Big|x_k[n]\Big|^2\Big\} \tag{3.54}$$

$$Y_k^1(f) = \widehat{X}_k(f) * \widehat{X}_k^*(-f) \tag{3.55}$$

$$= \int\limits_{-\infty}^{\infty} \widehat{X}_k(f-r)\widehat{X}_k^*(-r)dr \tag{3.56}$$

$$= \int\limits_{-\infty}^{\infty} \widehat{X}_k(r+f)\widehat{X}_k^*(r)dr \tag{3.57}$$

We can narrow the integration bounds provided the restrictions

$$\widehat{X}_k^*(r) \neq 0 \Rightarrow f_k - \frac{1}{2}f_{bw} < r < f_k + \frac{1}{2}f_{bw} \tag{3.58}$$

$$\widehat{X}_k(r+f) \neq 0 \Rightarrow f_k - \frac{1}{2}f_{bw} - f < r < f_k + \frac{1}{2}f_{bw} - f \tag{3.59}$$

$$a = max\Big(f_k - \frac{1}{2}f_{bw}, f_k - \frac{1}{2}f_{bw} - f\Big) \tag{3.60}$$

$$b = min\Big(f_k + \frac{1}{2}f_{bw}, f_k + \frac{1}{2}f_{bw} - f\Big) \tag{3.61}$$

$$Y_k^1(f) = \int\limits_{a}^{b} \widehat{X}_k(r+f)\widehat{X}_k^*(r)dr, \quad -f_{bw} < f < f_{bw} \tag{3.62}$$

$$= 0, \quad \text{else} \tag{3.63}$$

For $Y_k^2(f)$ there are actually three non-zero bands.

$$Y_k^2(f) = X_k(f) * X_k^*(-f) \tag{3.64}$$

$$= \int_{-\infty}^{\infty} X_k(r+f)X_k^*(r)dr \tag{3.65}$$

Case 1: $-2f_k - f_{bw} < f < -2f_k + f_{bw}$

$$Y_k^2(f) = \int_a^b X_k(r+f)X_k^*(r)dr \tag{3.66}$$

$$a = max\Big(f_k - \frac{1}{2}f_{bw}, f_k - \frac{1}{2}f_{bw} - f\Big) \tag{3.67}$$

$$b = min\Big(f_k + \frac{1}{2}f_{bw}, f_k + \frac{1}{2}f_{bw} - f\Big) \tag{3.68}$$

Case 2: $2f_k - f_{bw} < f < 2f_k + f_{bw}$

$$Y_k^2(f) = \int_a^b X_k(r+f)X_k^*(r)dr \tag{3.69}$$

$$a = max\Big(- f_k - \frac{1}{2}f_{bw}, -f_k - \frac{1}{2}f_{bw} - f\Big) \tag{3.70}$$

$$b = min\Big(- f_k + \frac{1}{2}f_{bw}, -f_k + \frac{1}{2}f_{bw} - f\Big) \tag{3.71}$$

Case 3: $-f_{bw} < f < f_{bw}$

This case is unique because there are two points of intersection. We can break up the integral into a sum. The first integral is exactly the same as in $Y_k^1(f)$.

$$Y_k^2(f) = \int\limits_{a_1}^{b_1} X_k(r+f)X_k^*(r)dr + \int\limits_{a_2}^{b_2} X_k(r+f)X_k^*(r)dr \tag{3.72}$$

$$a_1 = max\Big(f_k - \frac{1}{2}f_{bw}, f_k - \frac{1}{2}f_{bw} - f\Big) \tag{3.73}$$

$$b_1 = min\Big(f_k + \frac{1}{2}f_{bw}, f_k + \frac{1}{2}f_{bw} - f\Big) \tag{3.74}$$

$$a_2 = max\Big(-f_k - \frac{1}{2}f_{bw}, -f_k - \frac{1}{2}f_{bw} - f\Big) \tag{3.75}$$

$$b_2 = min\Big(-f_k + \frac{1}{2}f_{bw}, -f_k + \frac{1}{2}f_{bw} - f\Big) \tag{3.76}$$

Using the Hermitian symmetry of the real-valued $x[n]$,

$$Y_k^2(f) = \int\limits_{a_1}^{b_1} X_k(r+f)X_k^*(r)dr + \int\limits_{a_2}^{b_2} X_k^*(-r-f)X_k(-r)dr \tag{3.77}$$

$$r' = -r - f$$

$$Y_k^2(f) = \int\limits_{a_1}^{b_1} X_k(r+f)X_k^*(r)dr + \int\limits_{a_1}^{b_1} X_k^*(r')X_k(r'+f)dr' \tag{3.78}$$

$$=2\int\limits_{a_1}^{b_1} X_k(r+f)X_k^*(r)dr \tag{3.79}$$

$$=2Y_k^1(f) \tag{3.80}$$

If we lowpass filter $Y_k^2(f)$ with a filter defined

$$H_{lp}(f) = \frac{1}{2}, \quad |f| < f_{bw} \tag{3.81}$$

$$=0, \quad 2f_k - f_{bw} < |f| < 2f_k + f_{bw} \tag{3.82}$$

then

$$Y_k^2(f) = Y_k^1(f) \quad \forall f \tag{3.83}$$

We conclude that

$$\left|x_k[n]\right|^2 * h_{lp}[n] \approx \left|\widehat{x}_k[n]\right|^2 \tag{3.84}$$

Things to consider are delay and non-deal filters, however provided the distance between baseband and the $\pm 2f_k$ terms a sufficient filter is practical in practice.

Now the relationship between $m_{k,CIS}[n]$ and $m_{k,Hilbert}[n]$ is muddled by the nonlinear square root operation, however the nonlinearities induced won't be noticeably distorted by $h_{lp}[n]$. In practice, there only noticeable difference will be the added delay from the final lowpass filter in the CIS method.

3.4.4 Abstract Interpretation

One of the easier ways to interpret these methods is through a frequency domain analysis. Figure 3.4 shows an abstract view of the methods. The input is a magnitude spectrum of a signal with two harmonics. For mathematical convenience the output is actually the squared envelope. At each step a new operation is applied. This abstract analysis ignores scale factors that can always be modified by scaling filter coefficients.

First note that there are two paths for STFT. This is because there is an ambiguity in the order of operations. This can be seen mathematically in 3.85.

$$e^{-j\frac{2\pi}{N}kn}\left(x[n] * \left(w[-n]e^{j\frac{2\pi}{N}kn}\right)\right) = \left(x[n]e^{-j\frac{2\pi}{N}kn}\right) * w[-n] \tag{3.85}$$

The left side of 3.85 corresponds to the STFT 1 path. First an analytic bandpass filter centered at radian frequency $\frac{2\pi k}{N}$ is applied. The output of that is then downshifted to baseband.

The right side of 3.85 corresponds to the STFT 2 path. We first downshift by radian frequency $\frac{2\pi k}{N}$, then lowpass filter.

For both STFT 1 and STFT 2 the final operation is a magnitude.

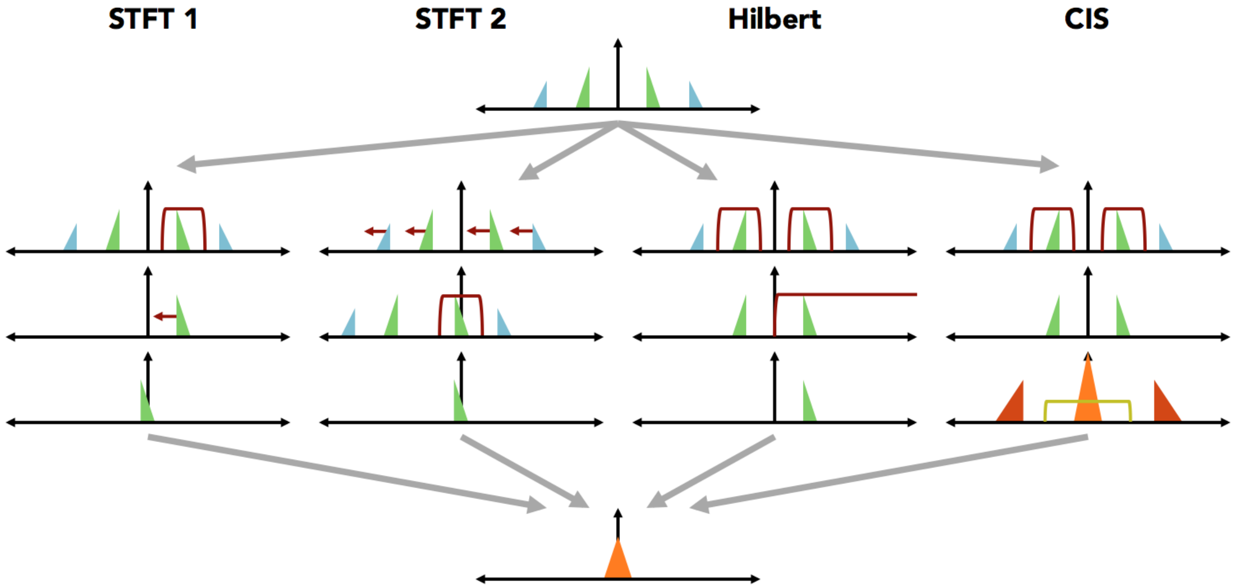

Figure 3.4: Method Comparison: magnitude spectrum at each step

The harmonic coherent method is missing from figure 3.4. This is because ignoring exact details of downshift frequency and filter coefficients, it is actually the same as the STFT method: downshift followed by lowpass filter.

Let's move on the the Hilbert envelope. In figure 3.4 we first bandpass filter, then acquire the analytic signal, which is equivalent to setting the negative frequencies to zero. The final operation is to take the magnitude squared, which is invariant to frequency shifts. From this abstract view, we should expect the same result as the STFT method.

In CIS, we see that taking the magnitude squared of the real bandpass signal causes double frequency terms, and the baseband term is scaled by a factor of 2. The final filter operation rescales the baseband term and eliminates the double frequency terms.

3.5 Summary

So what are the differences? To come to the conclusions made, some assumptions had to be made. We found that the Hilbert and CIS methods are approximately the same. STFT decomposition is a subset of the Hilbert method where the filterbank is comprised of uniform-bandwidth linearly spaced filters. Coherent harmonic is a expansion of STFT decomposition using the fundamental frequency of a signal to adaptively change downshift frequency and filter bandwidth.

In 3.86 we generalize the considered methods. $h_k\Big[n, F_0[n]\Big]$ is a function of k allowing for non-uniform bandwidths and a function of $F_0[n]$, allowing for coherent filter adaptation. Similarly, $\omega_k\big[F_0[n]\big]$ is a function of $F_0[n]$, allowing for coherent downshift frequencies.

$$m_k[n] = \Big| x[n]e^{-j\omega_k\big[F_0[n]\big]n} * h_k\big[n, F_0[n]\big]\Big| \tag{3.86}$$

In the next chapter we will investigate encoding harmonics in cochlear implants using our generalized envelope extraction equation.

Chapter 4

HARMONIC ENVELOPES

We want to come up with an envelope extraction system that best represents harmonic signals. Since harmonic signals have a specific structure, we model our harmonic signal as a restricted sum-of-products model. We can define our carriers from equation **??** as centered at multiples of F_0. In this representation $x_0[n]$ is the fundamental centered at F_0, $x_1[n]$ is the 1st harmonic centered at $2F_0$, etc. Without loss of generality, we will consider the analytic signal, $\widehat{x}[n]$.

$$\theta_k[n] =2\pi(k+1)\frac{F_0[n]}{F_s}n + \phi_k[n] \tag{4.1}$$

$$x[n] =\sum_{k=0}^{K} m_k[n]cos(\theta_k[n]) \tag{4.2}$$

$$\widehat{x}[n] =\sum_{k=0}^{K} m_k[n]e^{j\theta_k[n]} \tag{4.3}$$

We change our notation slightly from chapter 3. In this chapter $m_k[n]$ is the unknown desired envelope, and $\tilde{m}_k[n]$ is our extracted envelope estimate.

$$\tilde{m}_k[n] =\left|x[n]e^{-j\omega_k\left[F_0[n]\right]n} * h_k\left[n, F_0[n]\right]\right| \tag{4.4}$$

Provided our envelope extraction equation, 4.4, our goal is to best represent the desired $m_k[n]$.

The design can be summarized by two things:

- downshift frequency, $\omega_k\left[F_0[n]\right]$

- lowpass filter, $h_k\big[n, F_0[n]\big]$

If $w_k[\cdot]$ and $h_k[\cdot]$ are functions of $\widehat{x}[n]$ we have coherent envelope extraction. If they are time-invariant, we have incoherent extraction.

4.1 Steady-State Analysis

We start with the simplest scenario, where $\widehat{x}[n]$ is a steady-state signal. The conditions we require for this are:

- constant pitch: $F_0[n] = F_0$
- narrowband modulator: $m_k[n] \approx$ *constant* over short periods of time
- constant phase term: $\phi_k[n] = \phi_k$, we choose $\phi_k[n] = 0$ for cleaner equations however this is not necessary

4.1.1 3 Harmonic Example: Desired Envelope

We visualize the frequency domain for a signal with three harmonics ($K = 2$) in figure 4.1. For this example we consider the 1st harmonic ($k = 1$), centered at $2F_0$.

Figure 4.1(d) is the spectrum of the squared envelope, $|\mathcal{F}\{m_1^2[n]\}|$. We see this relationship in equation 4.8

$$(a) \quad \widehat{x}[n] \Longleftrightarrow \widehat{X}[n, f) \tag{4.5}$$

$$(b) \quad \widehat{x}_1[n] \Longleftrightarrow \widehat{X}_1[n, f) \tag{4.6}$$

$$(c) \quad \widehat{x}_1^*[n] \Longleftrightarrow \widehat{X}_1^*[n, -f) \tag{4.7}$$

$$(d) \quad m_1^2[n] = \widehat{x}_1[n]\widehat{x}_1^*[n] \Longleftrightarrow \widehat{X}_1[n, f) * \widehat{X}_1^*[n, -f) \tag{4.8}$$

The envelope can always be acquired from the squared envelope by a final square root operation. This operation introduces nonlinearities at multiples of F_0 that are difficult to

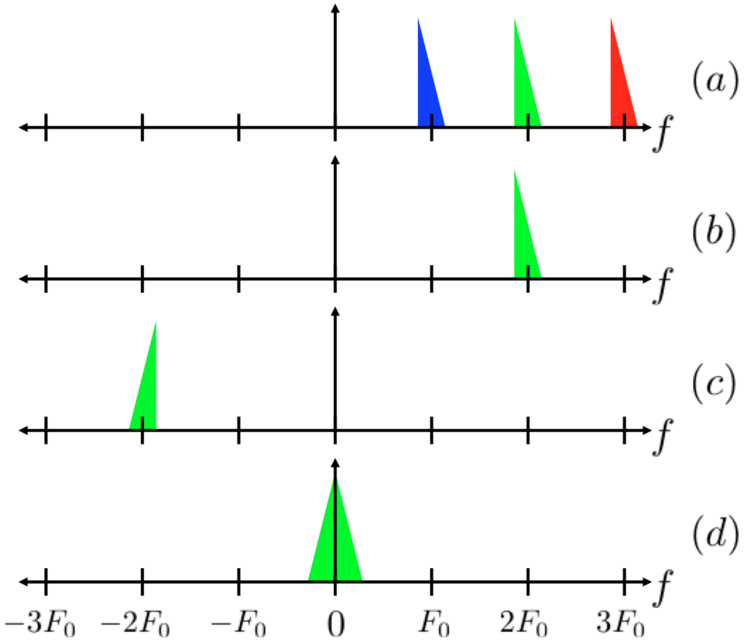

Figure 4.1: Magnitude of spectrum for equations 4.5 - 4.8

analyze. For mathematical convenience, during our analysis we can consider the squared envelope. This final square root operation will remain constant across all examples which allows us to not consider it.

$$m_1[n] = \left|\widehat{x}_1[n]\right| = \left[\widehat{x}_1[n]\widehat{x}_1^*[n]\right]^{\frac{1}{2}} \tag{4.9}$$

4.1.2 Estimated Envelope

Let's now evaluate our estimate, using equation 4.4. As stated above, we consider the squared envelope.

$$\begin{aligned}
\tilde{m}_k^2[n] =& \left|\widehat{x}[n]e^{-j\omega_k n} * h_k[n]]\right|^2 \\
=& \left|\sum_{l=0}^{K} m_l[n]e^{j(\theta_l[n]-\omega_k[n])} * h_k[n]\right|^2 \\
\approx& \left|\sum_{l=0}^{K} m_l[n]\Big(e^{j(\theta_l[n]-\omega_k[n])} * h_k[n]\Big)\right|^2 \\
=& \left|\sum_{l=0}^{K} m_l[n]e^{j\omega_{k,l}n} H_k\big(e^{j\omega_{k,l}}\big)\right|^2 \qquad (4.10) \\
\omega_{k,l} =& 2\pi\frac{(l+1)F_0 - F_{ds,k}}{F_s} \qquad (4.11) \\
h_k[n] \Longleftrightarrow& H_k\big(e^{j\omega}\big) \qquad (4.12)
\end{aligned}$$

$\omega_{k,l}$ is the downshifted center frequency of the l'th harmonic for the estimate of the k'th envelope. $H_k\big(e^{j\omega}\big)$ is the discrete Fourier transform (DFT) of $h_k[n]$.

Expanding equation 4.10 we get:

$$\begin{aligned}\tilde{m}_k^2[n] &= \sum_{l=0}^{K}\sum_{i=0}^{K} m_l[n]m_i^*[n]e^{j(l-i)F_0}H_k\left(e^{j\omega_{k,l}}\right)H_k^*\left(e^{j\omega_{k,i}}\right) \qquad (4.13)\\
&= \sum_{l=0}^{K}\left|m_l[n]\right|^2\left|H_k\left(e^{j\omega_{k,l}}\right)\right|^2\\
&+ e^{-j2\pi\frac{F_0}{F_s}n}\sum_{l=0}^{K-1} m_l[n]m_{l+1}^*[n]H_k\left(e^{j\omega_{k,l}}\right)H_k^*\left(e^{j\omega_{k,l+1}}\right)\\
&+ e^{j2\pi\frac{F_0}{F_s}n}\sum_{l=1}^{K} m_l[n]m_{l-1}^*[n]H_k\left(e^{j\omega_{k,l}}\right)H_k^*\left(e^{j\omega_{k,l-1}}\right)\\
&+ e^{-j2\pi\frac{2F_0}{F_s}n}\sum_{l=0}^{K-2} m_l[n]m_{l+2}^*[n]H_k\left(e^{j\omega_{k,l}}\right)H_k^*\left(e^{j\omega_{k,l+2}}\right)\\
&+ e^{j2\pi\frac{2F_0}{F_s}n}\sum_{l=2}^{K} m_l[n]m_{l-2}^*[n]H_k\left(e^{j\omega_{k,l}}\right)H_k^*\left(e^{j\omega_{k,l-2}}\right)\\
&+ \ldots\\
&+ e^{-j2\pi\frac{KF_0}{F_s}n}m_0[n]m_K^*[n]H_k\left(e^{j\omega_{k,0}}\right)H_k^*\left(e^{j\omega_{k,K}}\right)\\
&+ e^{j2\pi\frac{KF_0}{F_s}n}m_K[n]m_0^*[n]H_k\left(e^{j\omega_{k,K}}\right)H_k^*\left(e^{j\omega_{k,0}}\right) \qquad (4.14)\end{aligned}$$

We can now think of $\tilde{m}_k[n]$ as a combination of terms each centered at iF_0 where the magnitude of each term is:

$$\left|\tilde{m}_{k,iF_0}[n]\right| = \left[\sum_{l=0}^{K-|i|}\left|m_l[n]\right|\left|m_{l+i}[n]\right|\left|H_k\left(e^{j\omega_{k,i}}\right)\right|\left|H_k\left(e^{j\omega_{k,l+i}}\right)\right|\right]^{\frac{1}{2}}, \quad -K \leq i \leq K \qquad (4.15)$$

Evaluated at DC:

$$\left|\tilde{m}_{k,0F_0}[n]\right| = \left[\sum_{l=0}^{K}\left|m_l[n]\right|^2\left|H_k\left(e^{j\omega_{k,l}}\right)\right|^2\right]^{\frac{1}{2}} \qquad (4.16)$$

4.1.3 3 Harmonic Example: Estimated Envelope

Let's go back to our three harmonic example. We are again trying to acquire the 1st harmonic, $m_1[n]$ (green). We define $\omega_1 = 2F_0$.

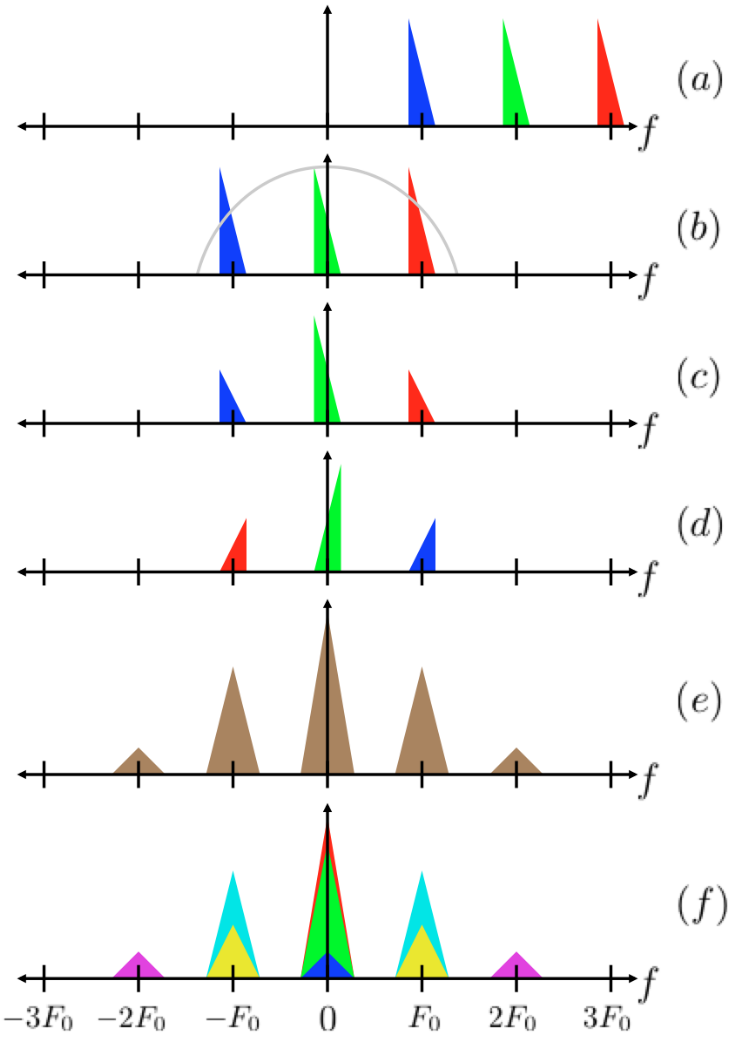

Figure 4.2: $(a)|\widehat{X}[n,f)|$ $(b)|\widehat{X}[n,f-2F_0)|$ $(c)|\widehat{X}[n,f-2F_0)||H_1(f)|$ $(d)|\widehat{X}^*[n,-f+2F_0)||H_1(-f)|$ $(e)|\mathcal{F}\{\tilde{m}_1^2[n]\}|$ (f) contributions of separate components of (e)

We can see the relationships

$$\widehat{x}[n] \Longleftrightarrow \widehat{X}[n, f) \tag{4.17}$$

$$\widehat{x}[n]e^{-j2\pi\frac{2F_0}{F_s}n} \Longleftrightarrow \widehat{X}[n, f - 2F_0) \tag{4.18}$$

$$\widehat{x}[n]e^{-j2\pi\frac{2F_0}{F_s}n} * h_2[n] \Longleftrightarrow \widehat{X}[n, f - 2F_0)H_1(f) \tag{4.19}$$

$$\tilde{m}_1^2[n] \Longleftrightarrow \widehat{X}[n, f - 2F_0)H_1(f) * \widehat{X}^*[n, -f + 2F_0)H_1^*(-f) \tag{4.20}$$

Equations 4.17 -4.20 are visualized in figure 4.2. The interesting part of figure 4.2 is (f). We see our green component that we were looking for, however there are a whole lot of other things that we didn't want.

Figure 4.1(d) is equivalent to the green component of figure 4.2(f) if our filter $|H_1(f)| = 1$ when $f \approx 0$.

The other components come from interactions with the unwanted harmonics that we failed to completely filter out. For clarity the convolution is visualized in figures 4.3, 4.4, 4.5. Positive and negative components are mirror images so the positive components are not explicitly visualized.

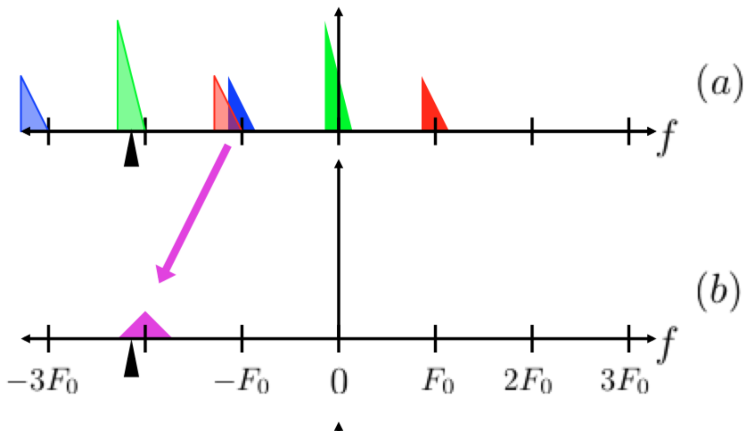

Figure 4.3: Envelope Estimate $-2F_0$ Component

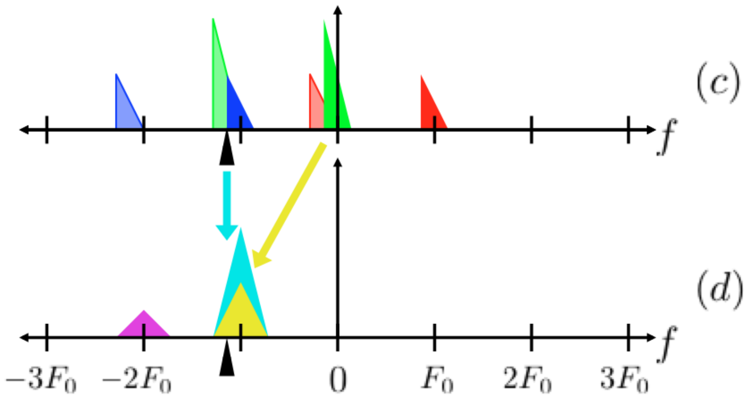

Figure 4.4: Envelope Estimate $-F_0$ Component

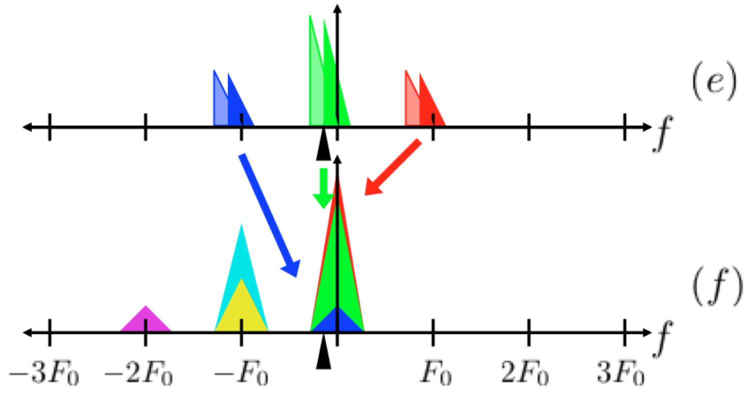

Figure 4.5: Envelope Estimate Baseband Component

4.1.4 2 Harmonic Example: Effects of Downshift Frequency

We saw clearly in our 3 harmonic example how a non-ideal filter causes distortions in the estimate. We now visualize the effects of downshift frequency by comparing $F_{DS} = F_0$ to $F_{DS} = F_0 + F_{err}$.

We can see from figure 4.6 that the non-ideal downshift affects the relative amplitudes of the desired harmonic to interference harmonics when filtering. As a result the baseband term will be lower amplitude and the terms at multiples of F_0 will be higher amplitude.

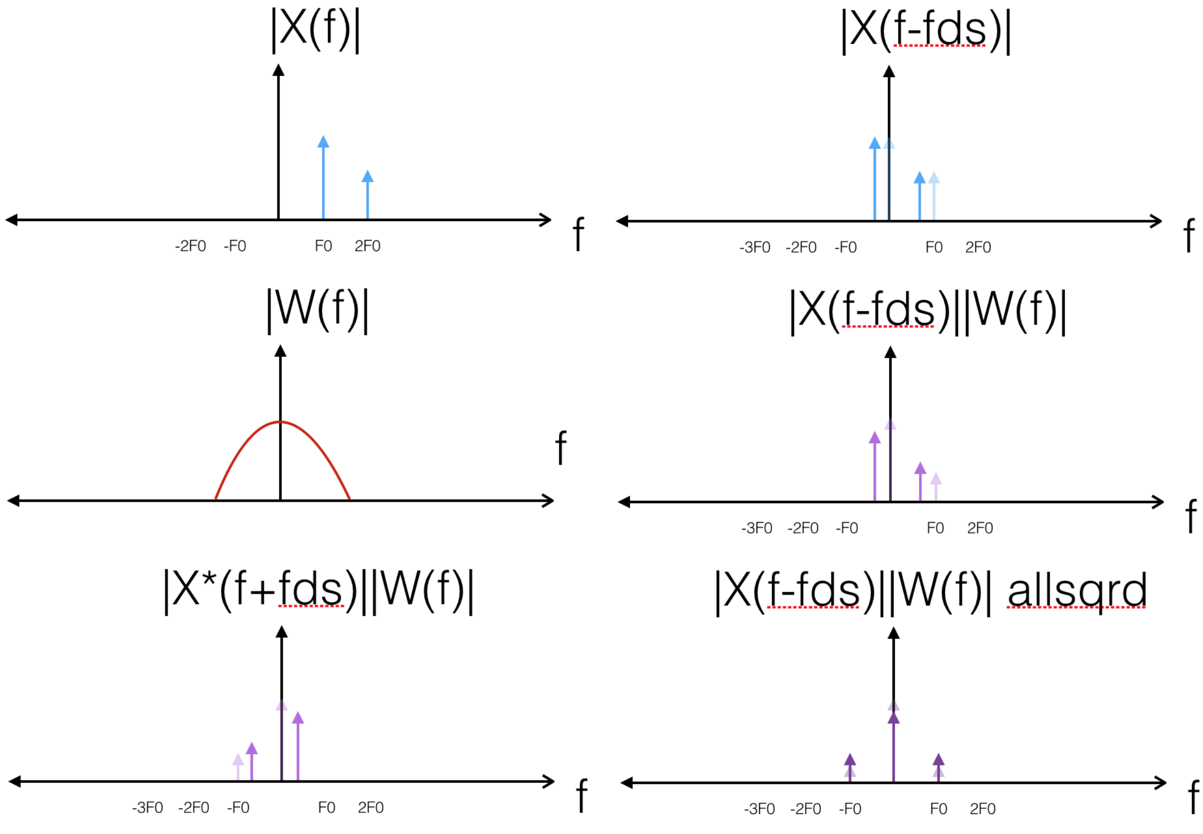

Figure 4.6: Comparison of ideal and non-ideal downshift frequencies, faded lines represent ideal $(a)|\widehat{X}[n,f)|$ $(b)|\widehat{X}[n,f-F_{DS})|$ $(c)|H(f)|$ $(d)|\widehat{X}[n,f-F_{DS})||H(f)|$ $(d)|\widehat{X}^{*}[n,-f+F_{DS})||H(-f)|$ $(e)|\mathcal{F}\{\tilde{m}_0^2[n]\}|$

4.2 Steady-State Metrics

In considering how well our envelope $\tilde{m}_k[n]$ estimates $m_k[n]$ there are three important metrics. We will now discuss each in detail.

4.2.1 Coherent Gain

Coherent gain is defined as the gain of the harmonic of interest, k.

$$G_k = \left| H_k\left(e^{j\omega_{k,k}}\right) \right| \tag{4.21}$$

Recalling equation 4.11, if $F_{ds,k} = (k+1)F_0$ then, $w_{k,k} = 0$ and the coherent gain is simply the DC gain of the filter.

$$G_k = \left| H_k(0) \right| = \sum_n h_k[n] \tag{4.22}$$

We may further simplify this by normalizing our filter such that $\left| H_k(0) \right| = 1$. Of course, our downshift frequency won't be ideal in real systems. Factors to consider include the quantization of $F_{ds,k}$ and the accuracy of F_0 estimation.

A similar metric, brought up in [windows for harmonic analysis] is termed scalloping loss, or picket-fence effect. This is the effect of the harmonic falling in between filter centers.

4.2.2 Harmonic SIR

Continuing our focus on the baseband, another question is: what is the contribution of the target harmonic versus the others? The baseband component is contributed to by spectral leakage due to non-ideal filters. This is visualized as the red and blue in figure 4.5(f). The harmonic signal-to-interference-ratio (SIR) quantifies the ratio of target harmonic to spectral leakage.

$$SIR_k = \frac{\left|H_k\left(e^{j\omega_{k,k}}\right)\right|}{\left[\sum_{l=0}^{K}\left|H_k\left(e^{j\omega_{k,l}}\right)\right|^2\right]^{\frac{1}{2}}} \tag{4.23}$$

The terms will roll off as the harmonic center frequencies get further away from $F_{ds,k}$, so typically SIR_k is sufficiently described by only one or two harmonics on either side of the k'th, i.e. $k-2 \leq l \leq k+2$.

Harmonic SIR does not describe the true signal-dependent SIR, as varying envelope magnitudes across harmonics will change this, however it does provide an objective measure of the quality of our system to arbitrary harmonic inputs.

4.2.3 Modulation Depth

Finally, we consider the magnitude of each bandpass component relative to baseband. These terms appear in our envelope estimate as modulations at rates that are multiples of F_0. Because of the forced symmetry of the real envelope we only need to consider positive frequencies, iF_0.

$$D_{k,i} = \frac{\left[\sum_{l=0}^{K-i}\left|H_k\left(e^{j\omega_{k,l}}\right)\right|\left|H_k\left(e^{j\omega_{k,l+i}}\right)\right|\right]^{\frac{1}{2}}}{\left[\sum_{l=0}^{K}\left|H_k\left(e^{j\omega_{k,l}}\right)\right|^2\right]^{\frac{1}{2}}}, \quad 1 \leq i \leq K \tag{4.24}$$

The largest value and, for that reason, most important value is $D_{k,1}$, the modulation depth at F_0.

4.3 Induced VS Explicit Temporal Modulation

So our three metrics are coherent gain, harmonic SIR and modulation depth. We aim for a coherent gain of $G_k = 1$ and maximized harmonic SIR.

We have mentioned in section 2.1.1 that we can have either induced or explicit temporal modulations. For explicit modulation systems our goal is minimum modulation depth. For induced that is not as clear.

In this document we argue that the latter, explicit modulation option is better. The reasoning is best shown by a motivational example.

Let's consider a single note played by two different instruments: clarinet and saxophone. In this example $F_0 = 261Hz$. The clarinet is interesting in that it only has energy at odd harmonics.

We attempt to estimate the 3rd harmonic, $m_3[n]$. We first downshift by $-3F_0$, then lowpass filter. The spectrum of each signal at this stage is visualized in figure 4.7. The top panel shows the output of a sufficiently narrow filter where the 3rd harmonic is isolated. The bottom panel shows a different filter design that intentionally allows the two adjacent harmonics to pass through. Here we start to see the problem, that despite the wide bandwidth filter, there is (almost) no energy around $\pm F_0$ for the clarinet because of the harmonic structure. (There is something present however it's down 30dB.)

Figure 4.8 shows the time-domain envelopes resulting from this processing. The input signals were normalized such that the top panel shows the same signal power for both instruments.

The problem is clearly represented in the bottom panel, were we have a very large F_0 modulation in the saxophone envelope but little to no change in the clarinet. The result is that we have a much stronger temporal pitch cue as well as louder overall volume to the saxophone.

Spectral leakage into other harmonic envelopes is not natural. It forces the envelope to modulate as a function of the adjacent harmonics which, as we just saw, is signal dependent. Furthermore, if we have uniform bandwidth filters, (as ACE does), the harmonic resolution will not behave as it does in the cochlea.

Beyond this example, explicit modulation decouples F_0 and modulation depth. This way we have much more control over modulation depth while still making optimal design

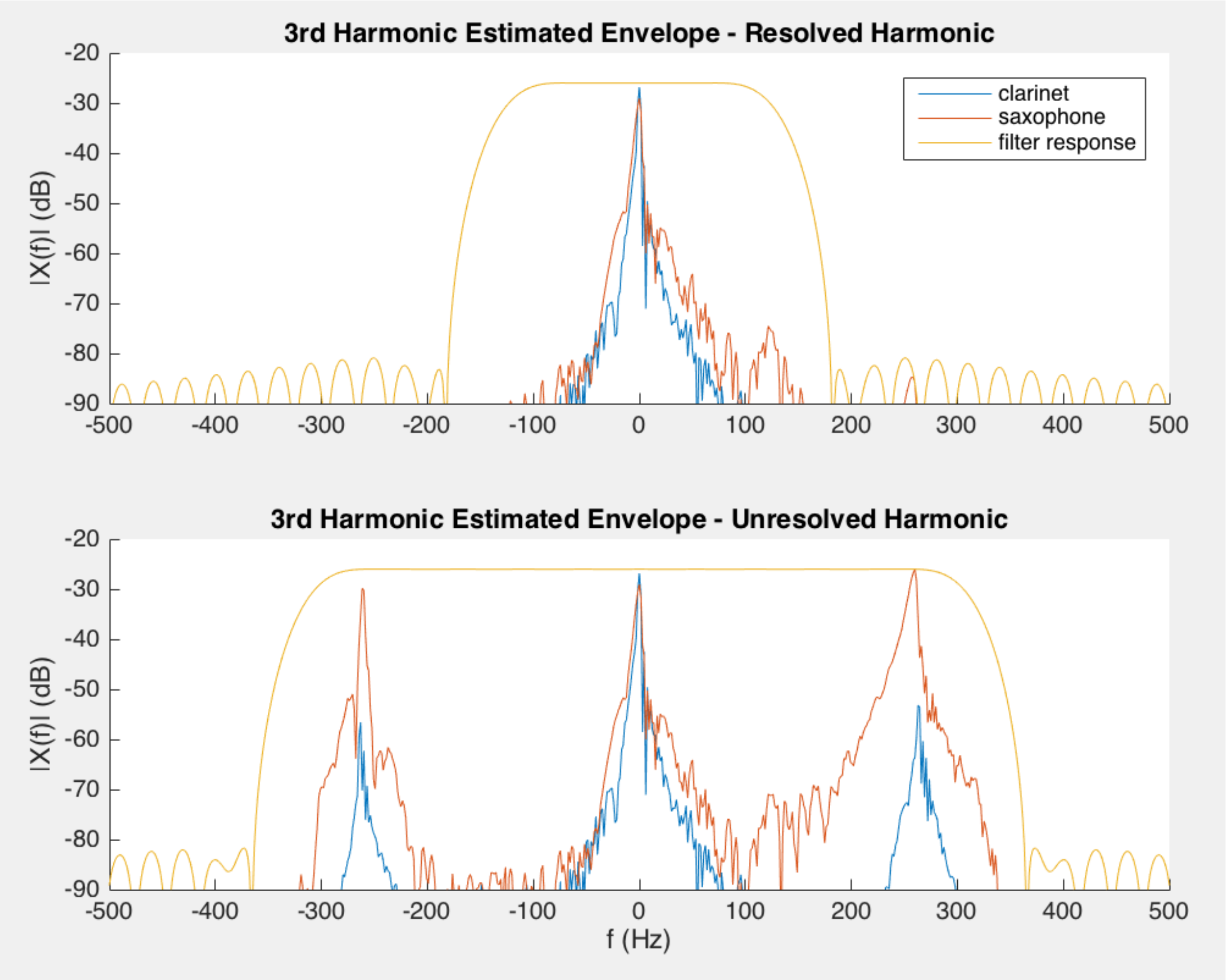

Figure 4.7: Clarinet vs Saxophone Harmonic Components

decisions for envelope extraction. We can decide modulation depth as a function of how harmonic the signal is. eTone [REF] uses a harmonic probability metric to do just that.

4.3.1 Followup Filter

Another thing to note is that regardless of downshift frequency, our harmonic envelope will always have it's energy centered at baseband and multiples of F_0. An alternative way of eliminating induced modulations is to add a lowpass filter to the end of the processing chain.

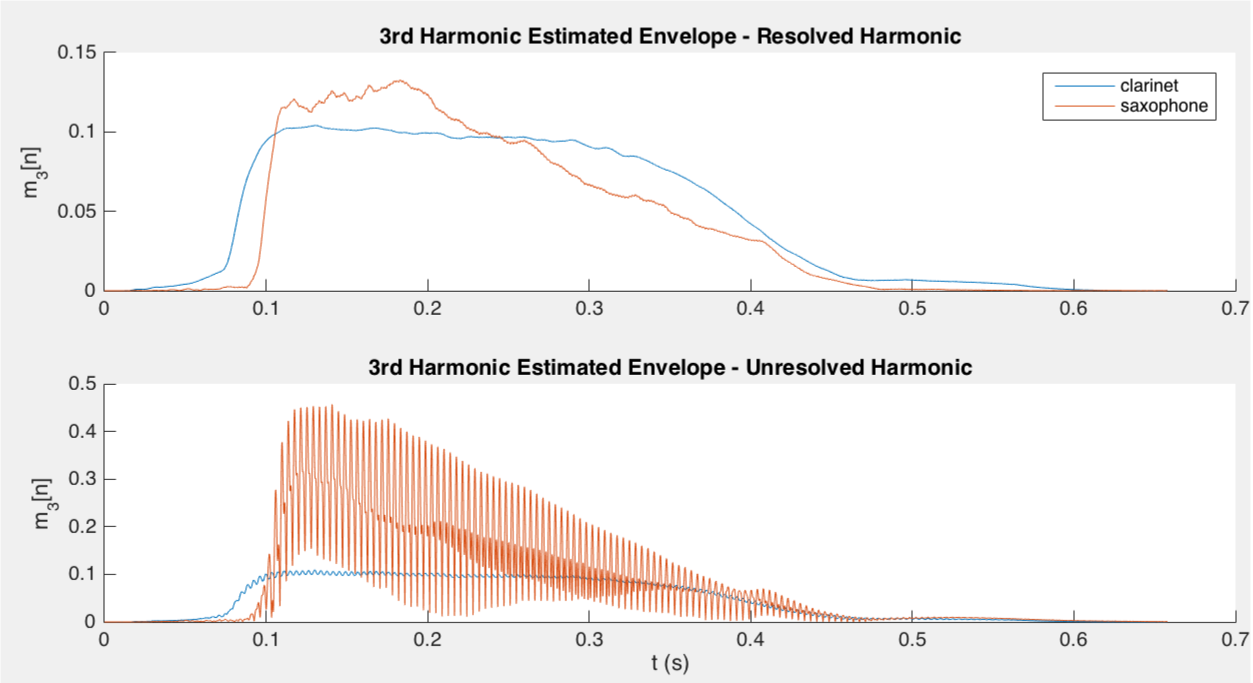

Figure 4.8: Clarinet vs Saxophone Envelope Estimates

There are a handful of research strategies [REF?] that have used this additional filter. eTone's envelope follower is an example of this.

The main improvement to a followup filter is that we can guarantee to eliminate temporal modulations. This could also be achieved by designing a sufficiently narrow filter, $h_k[n]$ however this brings about a tradeoff, where the narrower our filter is the more susceptible we are to error in downshift frequency.

In terms of our three metrics, the followup filter will provide us with a robust coherent gain and guaranteed low modulation depth at the cost of lower harmonic SIR.

Another point to consider is the cost of adding an extra processing stage. The additional stage means more memory, clock cycles and processing delay.

4.4 Time-Varying F0

We are only concerned with continuous changes in $F_0[n]$. Jumps would imply different harmonic envelopes.

$\tilde{m}_k[n]$ uses a window of samples of $\widehat{x}[n]$, equal to the length of $h_k\big[n, F_0[n]\big]$. If $F_0[n]$ changes significantly within this window we will have problems with our estimate. That being said, the longest windows considered in the document are 32ms long. In terms of music, 32ms is equivalent to a sixty-forth note at 120BPM (beats per minute), i.e. very fast. We will consider this sufficient for typical rates of change of $F_0[n]$.

The other problem that can arise is as $F_0[n]$, our steady-state metrics may change. This can be evaluated by simply looking at the continuous metrics as a function of F_0.

4.5 Transients

Nearly everything we have considered so far has suggested the narrower the filter the better. The problem with this is the time-domain response of filters with fast rolloffs. There is a tradeoff where the sharper a filter rolls off, the more transient smearing will incur.

Studies on timbre perception [REF?] have suggested that humans hear changes in rise time in the log domain, i.e. the shorter a transient is, the more sensitive our perception is to smearing distortion.

Of course if the pre-processing smears the transients, we can only do as well as that. Most cochlear implants nowadays use pre-processor dynamic range compression. We get some insight from a study performed on hearing aids, which would use a similar system. “Almost all of the hearing aids tested have attack times less than or equal to 10 ms. A little more than half of the hearing aids had release times of 50 ms or less. The range of the attack times varied from 1 to 23 ms” [attack and release times of AGC hearing aids] 1ms is faster than most typical sounds, so we should try to smear transients as little as possible in our processing.

All of this suggests filter bandwidth be as wide as possible without encompassing the other harmonics, which results in a cutoff of $\frac{F_0}{2}$.

4.6 Evaluation of Strategies

As stated above the design can be summarized by downshift frequency and lowpass filter.

The ideal downshift frequency is simply $(k+1)F_0[n]$. The question is what degree of quantization is sufficient to estimate our signal.

For filter design we need to consider bandwidth as a function of filter order and filter/window type. Ideally our cutoff is somewhere below F_0 but high enough to incorporate the bandwidth of $m_k[n]$.

The filters can be different as a function of k. This is a natural path to pursue if we consider the critical bands of the cochlea. This will be discussed in more detail later in this document however for now we will assume $h_k[n] = h[n]$. This is natural for harmonic envelopes as harmonics are linearly spaced.

The designs considered are:

downshift quantization - 1, 31, 63, 125Hz

filter order - 128, 256, 512

filter - rectangular, hanning, adaptive hamming

k - which harmonic, how do they relate to each other

Adaptive hamming is an adaptive bandwidth filter with a lowpass cutoff (-6dB point) of $\frac{F_0[n]}{2}$.

For practical considerations, we will set a maximum quantization as F_s / filter order.

$order = 256 \longleftrightarrow f_q \leq 63Hz$

$order = 512 \longleftrightarrow f_q \leq 31Hz$

4.6.1 Coherent Gain

We first look at different downshift quantizations, all else constant. This is visualized in figure 4.9. When F_0 is exactly at a quantized value, $G_k = 0dB$, however the gain decreases as F_0 drifts away until the worse case where it is exactly in between quantization values.

Decreasing the quantization increases the number of dips and in turn improves the worst case G_k.

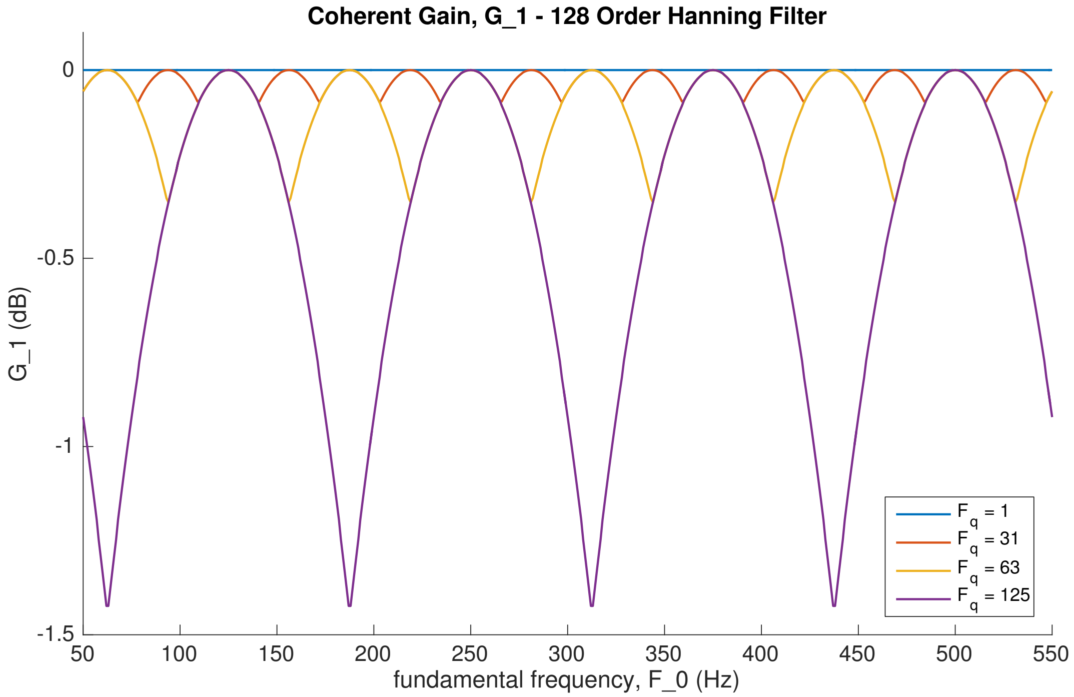

Figure 4.9: G_k Downshift Quatization

Figure 4.10 compares the three different filter orders. Using a hanning window, the lower order filters have slower rolloffs and better worst case G_k. This doesn't necessarily hold true for adaptive filters. Provided a high enough desired cutoff that the 128 order filter can achieve this reasonably well, the order become irrelevant.

Figure 4.11 compares the different filter designs. The wider bandwidth filters have smoother G_k across F_0 and as a result the adaptive bandwidth becomes optimal at high F_0's.

So lower quantization and wider bandwidth both improve G_k, but that's pretty intuitive. The interesting part here is the relationship between harmonics. If we consider the first three harmonics, figure 4.12 shows that the number of dips is proportional to k. As a result we

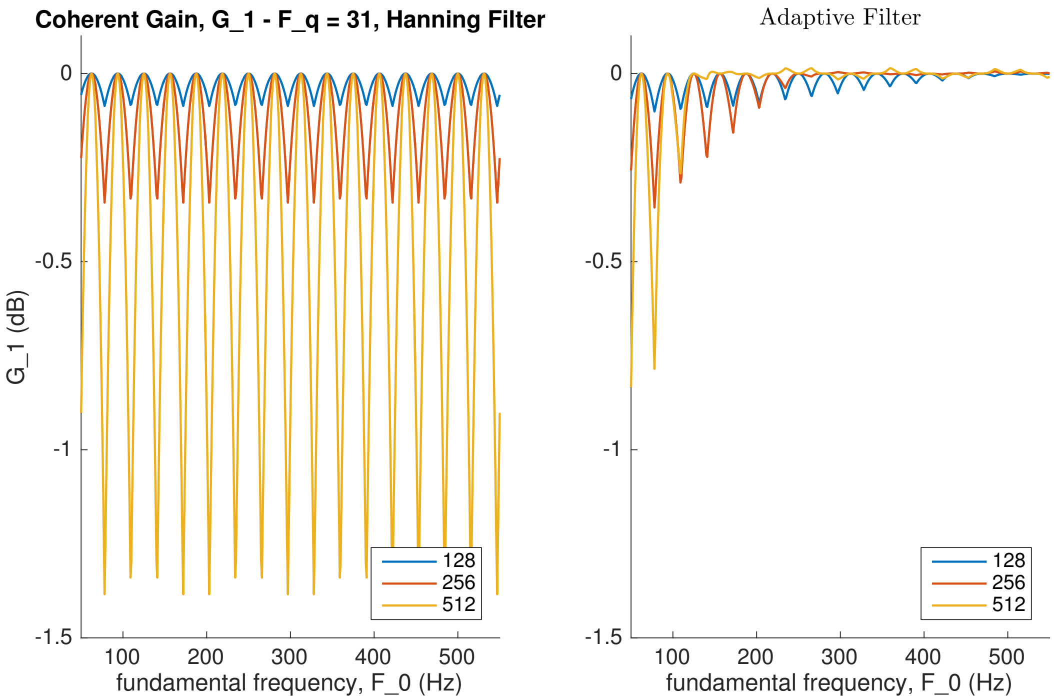

Figure 4.10: G_k filter order

get interactions at certain values of F_0. For example, if $F_0 = 1.5 f_q = 188\text{Hz}$, odd harmonics will be at a minimum and even harmonics will be at a maximum. This results in a distortion between harmonics where some are attenuated more than others.

It should be noted that pre-processing compression or automatic gain control (AGC) will cause harmonic distortions. This could arguably be used to either make the case that it is important to minimize further distortions, or alternatively that these further distortions are minimal in comparison and thus shouldn't be over engineered.

Considering maximum quatization is F_s / order and hanning filter as our baseline, worst case: $G_k \approx -1.5dB$. Increasing the filter order and decreasing quantization proportionally increases the number of dips while keeping depth the same. The relationship between harmonics and the case of $F_0[n]$ continuously changing over time put emphasis on minimizing the dynamic range of G_k.

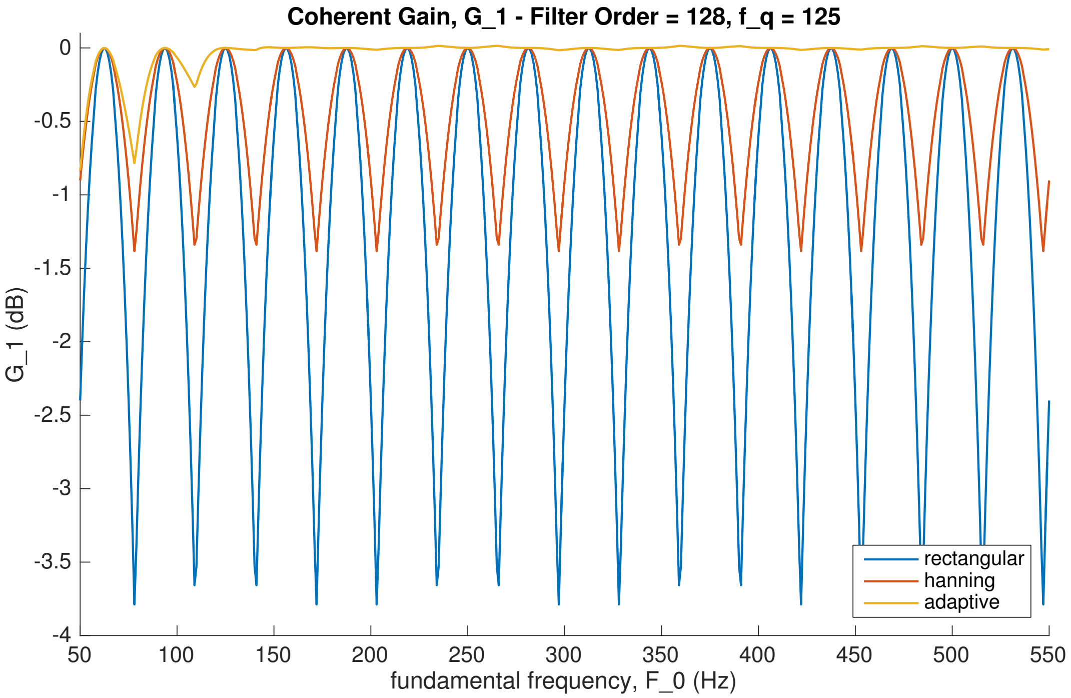

Figure 4.11: G_k filter design

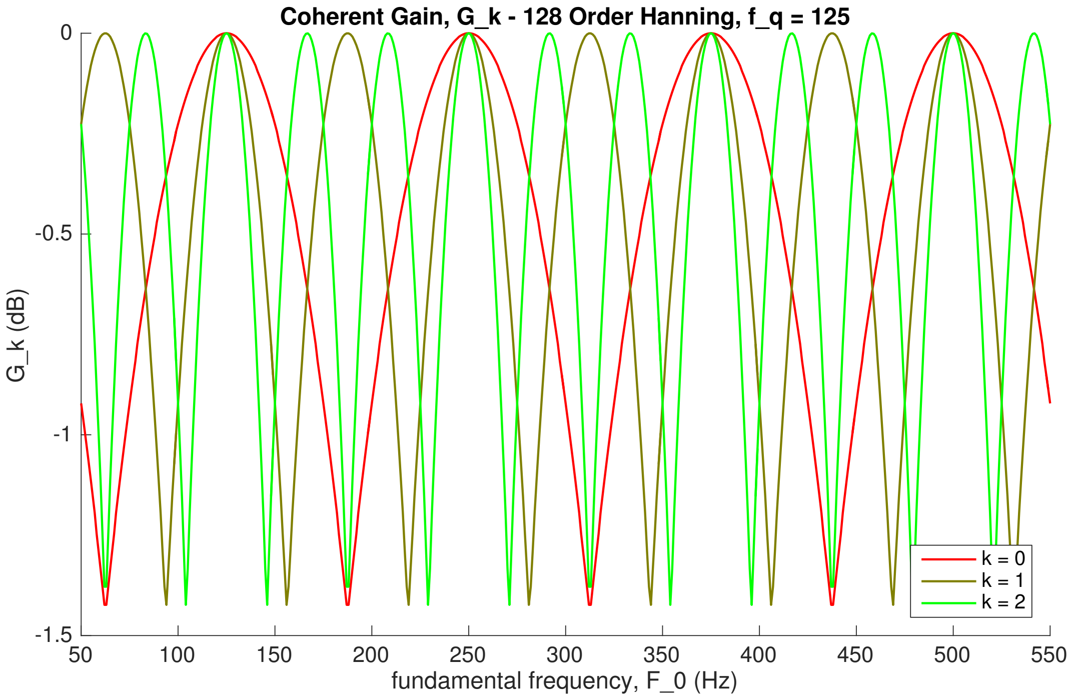

Figure 4.12: G_k variation across harmonics

4.6.2 Harmonic SIR

We first consider filter order and quantization. In figure 4.13 we consider all filter orders with and without quantization.

The downshift quandization doesn't actually affect performance significantly. This can be seen in figure 4.13 by looking at the two lines corresponding to order = 128. Above a $F_0 = 250$Hz the harmonics are spaced far enough apart that the quantization doesn't matter. Below $F_0 = 130$Hz the filter cutoff is not sharp enough to isolate the harmonic, in which case downshift quantization is irrelevant.

Also note that for order = 512 the cutoff is narrow enough that we get ideal harmonic SIR over all F_0.

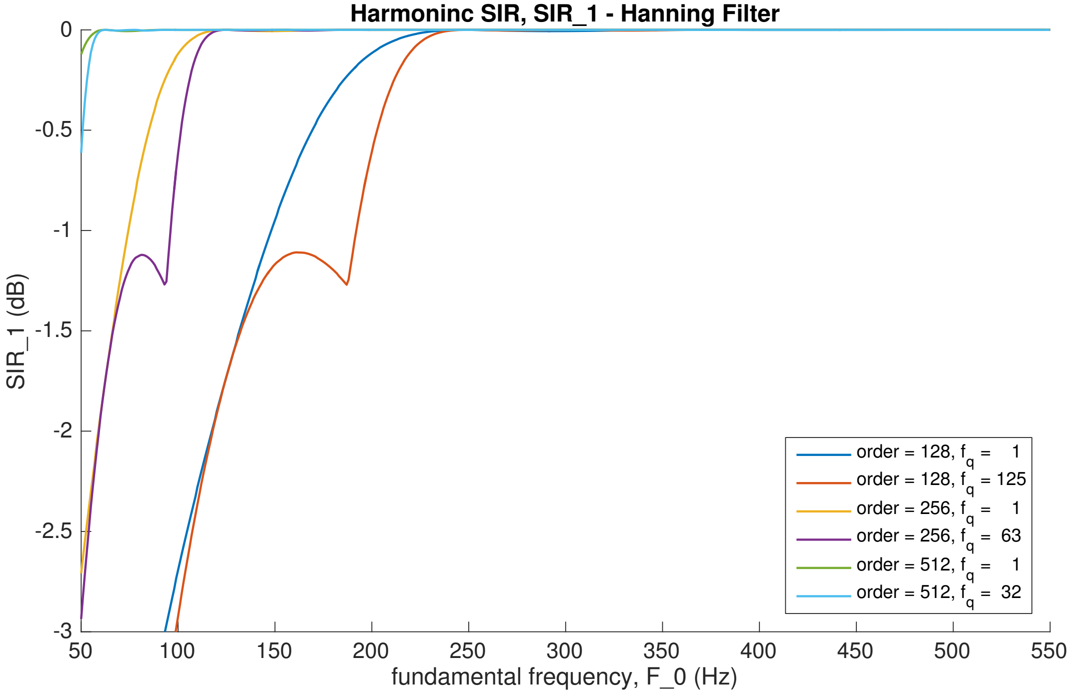

Figure 4.13: SIR_k filter order and quantization

Figure 4.14 compares filter design methods. Hanning and adaptive are essentially the same, showing that the limiting factor is still filter order. Rectangular provides a better lower limit for what F_0 the SIR breaks down at, and it does this at the cost of dips at higher frequencies. This agrees with the fact that rectangular windows have the sharpest rolloff at the expense of large sidelobes.

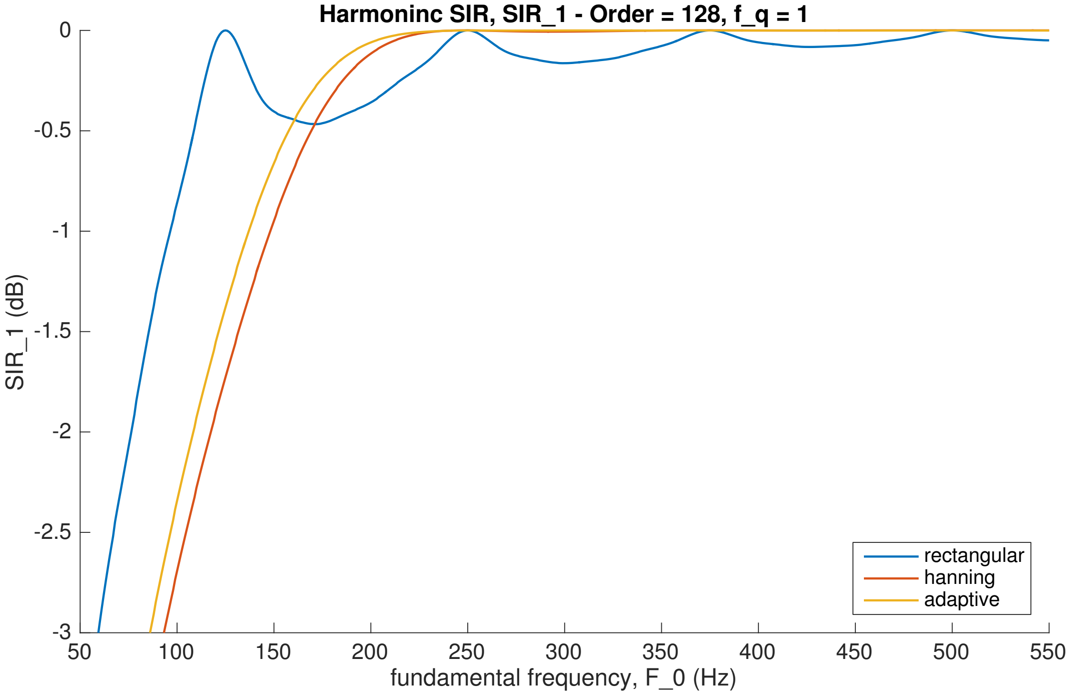

Figure 4.14: SIR_k filter design

The higher order harmonics are compared in figures 4.15 and 4.16. We see patterns similar to figure 4.12 where the number of dips is proportional to k. These figures reinforce that improvement from decreasing quantization, f_q, is bounded.

For hanning the incremental 1dB of improvement is arguably not important. For rectangular we actually see a significant improvement in the 80-130Hz region for $k > 3$.

Filter order is certainly the dominant factor for harmonic SIR. For order = 128, it starts to break down for $F_0 \approx 220$ Hz and degrades as F_0 decreases. For order = 256, it starts to break down for $F_0 \approx 110$ Hz. For order = 512 the harmonic SIR performs is essentially

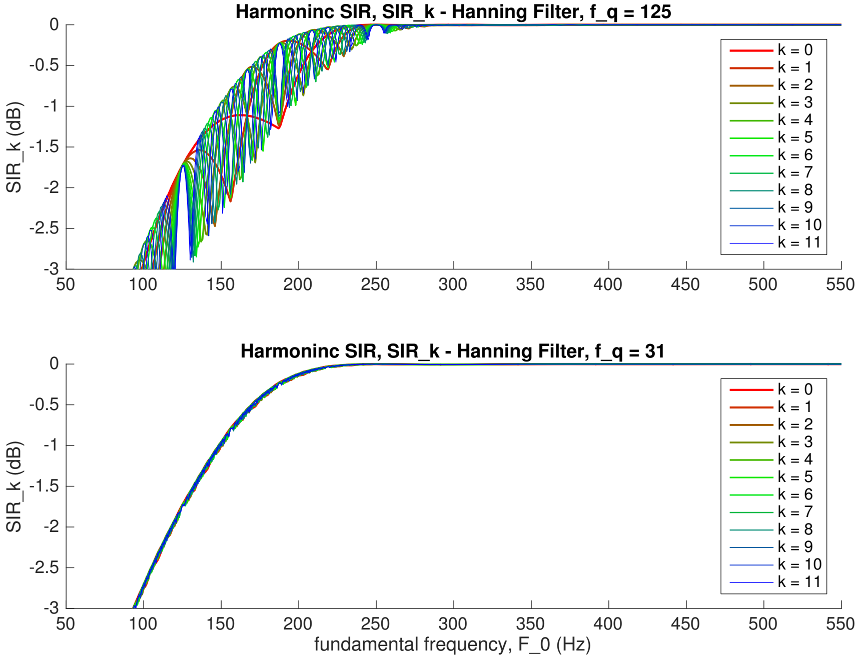

Figure 4.15: SIR_k variation across harmonics with hanning filter

optimal across all values of F_0.

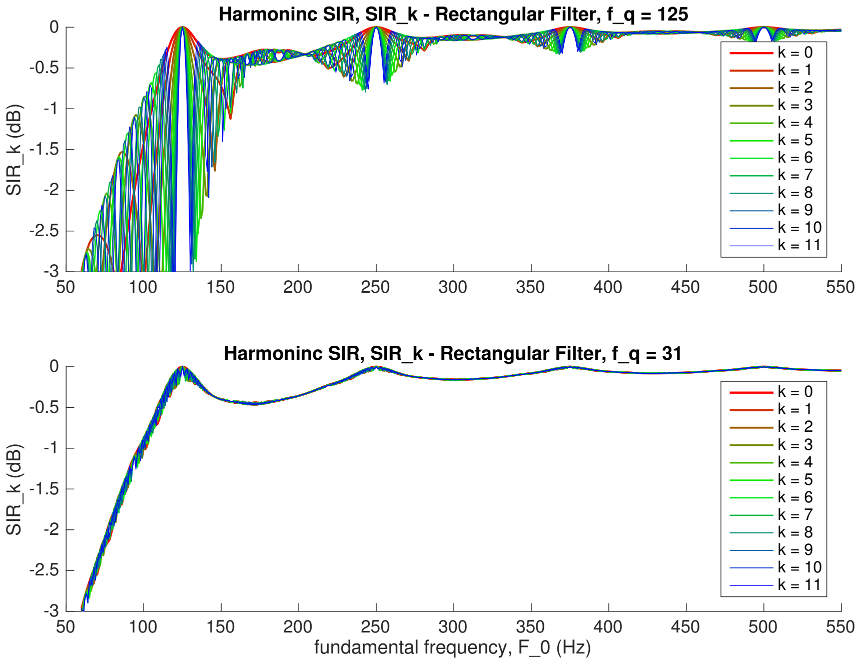

Figure 4.16: SIR_k variation across harmonics with rectangular filter

4.6.3 Modulation Depth

We discussed earlier that our goal is to provide explicit modulations, in which case we need minimal modulation in the extracted envelope.

We first compare each filter design method at the different filter orders, as seen in figure 4.17. For all orders rectangular windows do a poor job of suppressing modulations due to high sidelobe amplitude. Hanning and adaptive show similar responses with the dominant variation being the response to low F_0 as a function of filter order.

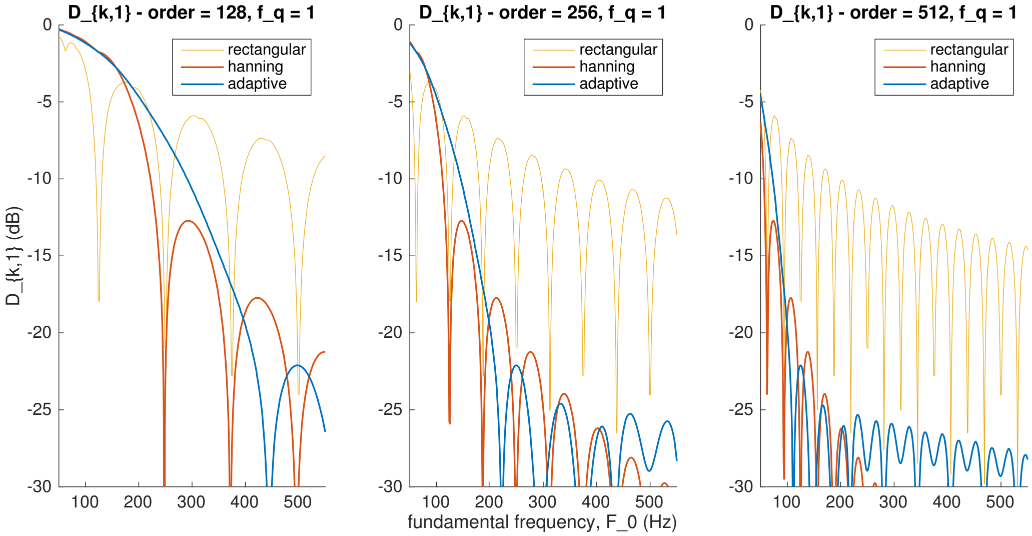

Figure 4.17: $D_{k,i}$ filter design and order

Downshift quantization shows little affect on modulation depth. This is shown for both hanning and adaptive filter designs in figure‘4.18.

Provided no downshift quantization, ,odulation depth won’t change as a function of k. Figure 4.19 shows this variation, however it has minimal impact.

Recall $D_{k,i}$ is the modulation depth of the estimate of the kth harmonic at a rate of iF_0. We should expect that as i increases we move further away from baseband and our filter does a better job of eliminating modulations. This is verified in figure 4.20.

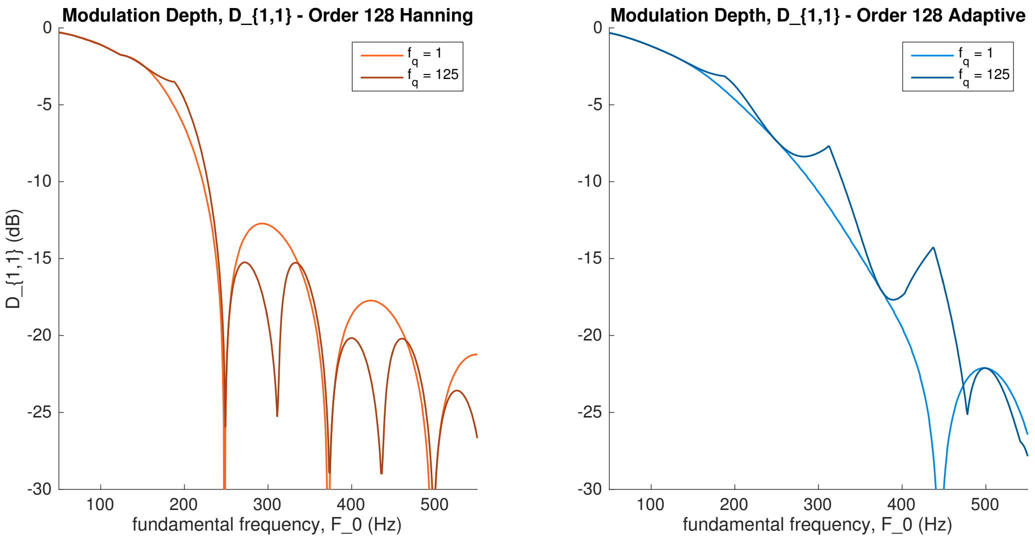

Figure 4.18: $D_{k,i}$ downshift quantization

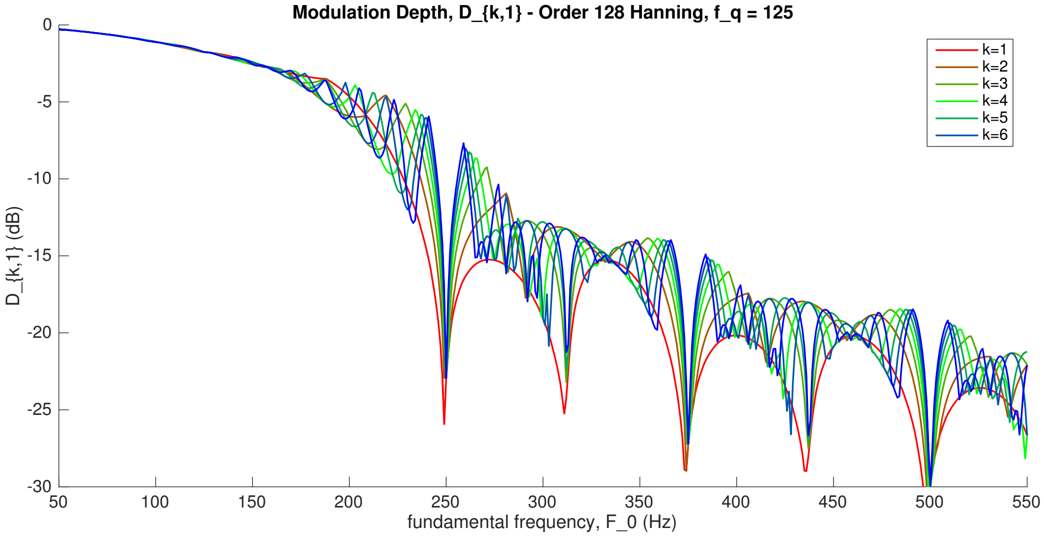

Figure 4.19: $D_{k,i}$ at rate of iF_0

This results suggest that $D_{k,1}$ is the most important measure, and that hanning and adaptive filter designs achieve approximately the same performance. At low F_0 filter order plays a large roll in modulation depth.

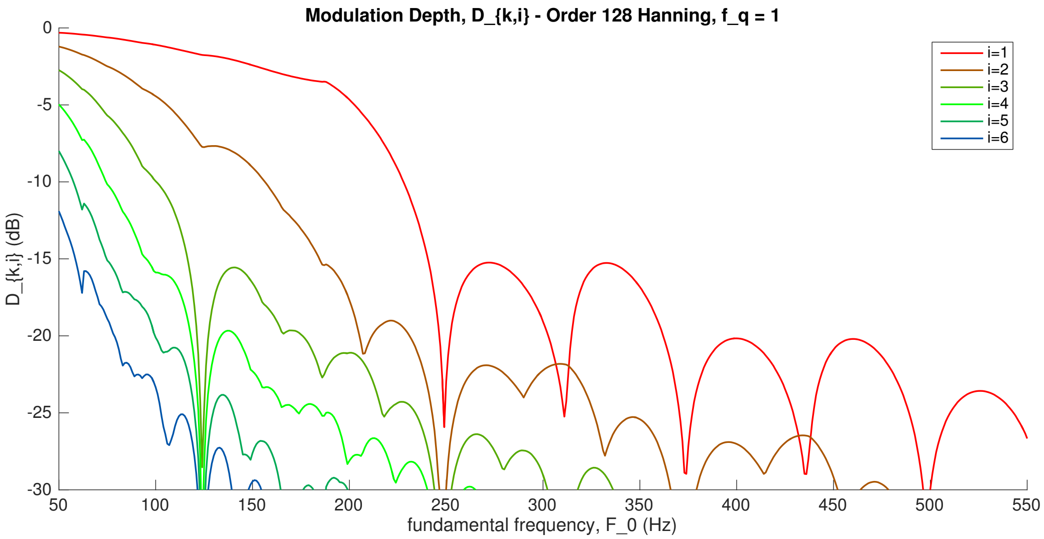

Figure 4.20: $D_{k,1}$ across harmonics

Psychophysical studies have found that for reliable pitch discrimination amplitude-modulations of approximately 10% to 40% are required on average. [REF]

$$10\% \rightarrow D_{k,1} = -20\text{dB}$$

$$40\% \rightarrow D_{k,1} = -8\text{dB}$$

This implies that depending on the user:

- order 128 breaks down at $F_0 \approx 240$ to 400Hz
- order 256 breaks down at $F_0 \approx 120$ to 200Hz
- order 512 breaks down at $F_0 \approx 60$ to 100Hz

In the best case, order 512 is sufficient for all F_0. In the worst case, order 128 will have artifacts across almost the entire F_0 range.

4.6.4 Transients

Time-responses are a bit more difficult to analyze, as we cannot use the standard dB measurements we are familiar with. We will consider transient responses of the different filter designs and filter orders.

We start with the unit step response, shown in figure 4.21. Latency on the order of 15ms isn't of much concern. The more important difference in the rise time. The 10-90% rise times are displayed in table 4.1.

The adaptive filters all have the same rise time at high enough F_0 however the lower order filters are fundamentally constrained on how slow the rise time can be. The rectangular window is the worst of them all.

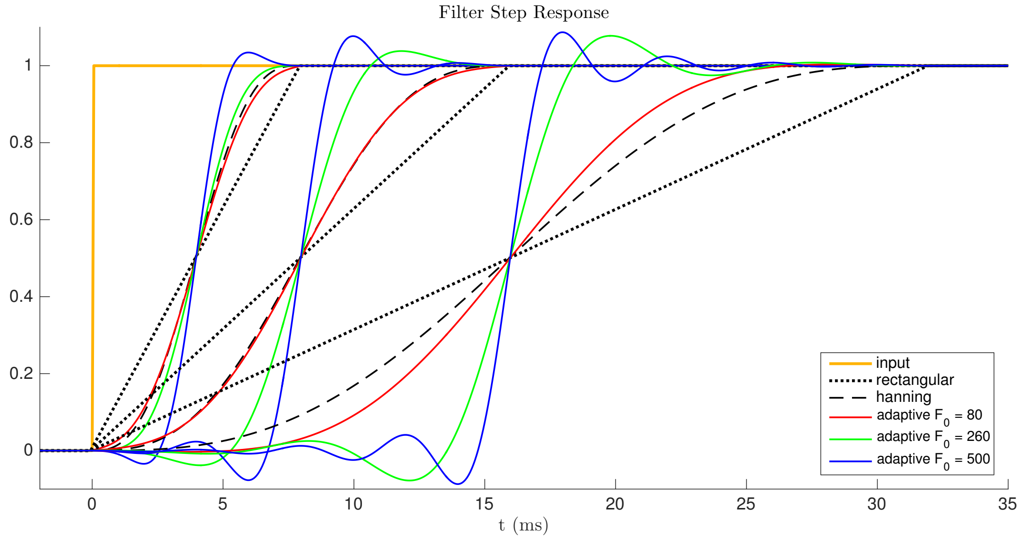

Figure 4.21: Transient Step Response, order = 128, 256, 512 (increasing order corresponds to longer reponse time)

An alternative view is shown in figure 4.22. For typical attack times in the range of 5-200ms and input to output change in attack time is plotted.

As mentioned in section 4.5 humans hear transient changes in the log domain, and thus the axes are log scaled.

	rectangular	hanning	adaptive 80	adaptive 260	adaptive 500
Order	**Rise Time (ms)**				
128	7	4	4	3	2
256	13	8	8	4	2
512	26	16	12	4	2

Table 4.1: filter rise times

For the worse case, rectangular order 512, more than half the dynamic range is lost due to smearing.

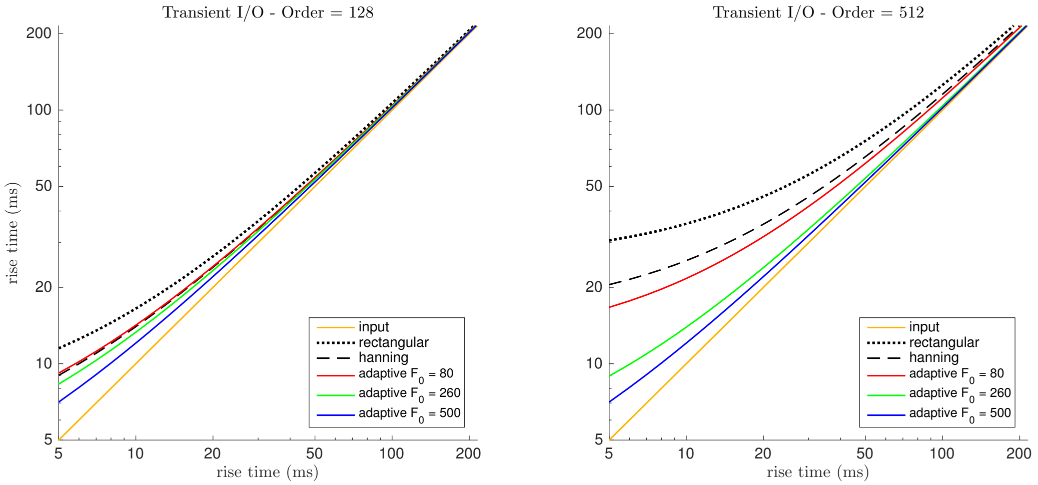

Figure 4.22: Transient Input/Output Change

As a final perspective on transients, we consider typical instrument attack times. Figure 4.23 shows the shifted attack times of twelve instruments typical attack times. The vertical scale has no meaning, it is simply for visual clarity.

What's interesting is that on a log scale, the instruments generally bunch into two groups. The slow attack-time group seems robust to the distortions of any of these filters. On the other hand the fast-attack time instruments change dramatically. For the narrow bandwidth 512 order filters, the smeared guitar output is closer in attack-time to an English horn than itself!

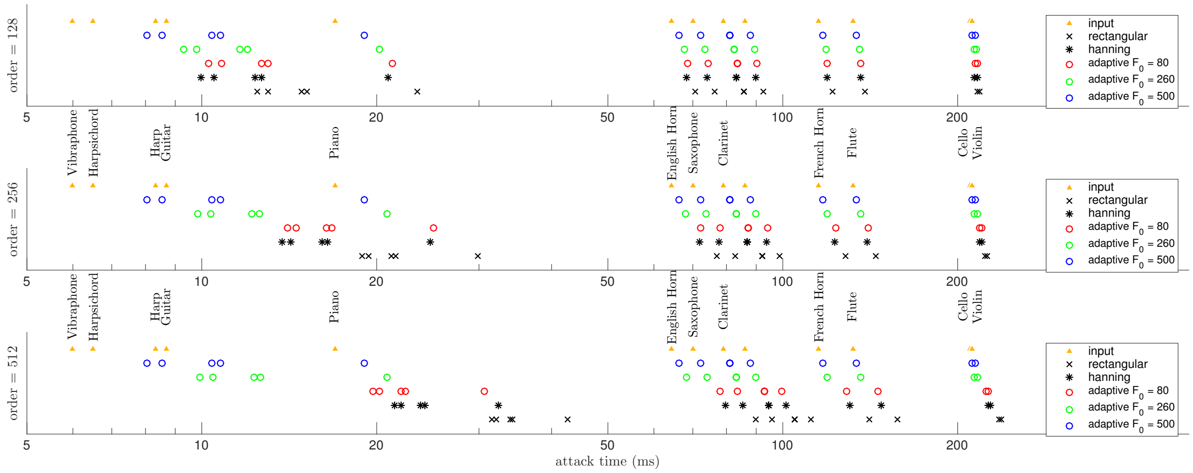

Figure 4.23: Transient Distortion for Common Instruments

4.6.5 Summary

For the most part the hanning and adaptive filters outperformed rectangular. The rectangular window's performance on modulation depth makes it essentially unusable.

For coherent gain we get a worst case of roughly -1.5dB. It doesn't appear from our results to be an overly critical design consideration.

For harmonic SIR and modulation depth the critical performance variable was filter order. To very loosely summarize, order 128 fails for $F_0 < 240$Hz, order 256 fails for $F_0 < 120$Hz and 512 does sufficiently well for the full range considered.

Downshift quantization also did not seem to play a prominent role. This is in part affected by the restriction that quantization can't be worse than F_s / filter order.

There is clearly an envelope bandwidth tradeoff, where the wider a filter is the less transients are smeared but the more the other harmonics interfere in the estimated envelope.

The sharp-cutoff order 512 filters smear the fast transients a significant amount, however the adaptive bandwidth filters seem to do well at smearing as little as possible while still achieving good performance on the other metrics. This could be the best solution to the posed bandwidth tradeoff.

4.7 Non-ideal Pitch Estimators

The critical assumption thus far has been accurate pitch estimates. One problem to consider is error in the pitch estimator. The other that we will consider is pitch estimator quantization.

We consider a specific pitch estimator that uses autocorrelation. To summarize this method, an autocorrelation is performed on the windowed input. A maxima is selected from this autocorrelation and the fundamental frequency is computed from the index of the maxima.

$$\begin{aligned} R_{xx}[n,\tau] =& x_{windowed}[r] * x_{windowed}[-r] && (4.25) \\ \tilde{F_0}[n] =& F_s \Bigg[\arg\max_{\tau} R_{xx}[n,\tau] \Bigg]^{-1} && (4.26) \end{aligned}$$

This can be implemented efficiently using the fast-autocorrelation method

$$R_{xx}[n,\tau] = \mathcal{F}^{-1}\Big\{X[n,k]X^*[n,k]\Big\} \tag{4.27}$$

Defining the FFT order as N, for this method the possible values of F_0 are

$$F_0 = \frac{F_s}{\tau}, \quad 1 \le \tau \le \frac{N}{2} \tag{4.28}$$

By then bounding the considered F_0 values to roughly 50-550Hz we can get better resolution by resampling the signal such that more values of F_0 fall within these bounds.

$$max\Big(\frac{2Fs}{N}, 50\Big) \le F_0 \le min\Big(\frac{Fs}{2}, 550\Big) \tag{4.29}$$

Choosing F_s is important, since the quantization of F_0 is not linearly spaced and becomes worse at higher values of F_0.

To be clear that this different sampling rate is only relevant to pitch estimation and not any of the other envelope extraction process, we define a new pitch estimator sampling rate, $F_{s,p}$. Having N as the filter orders we have previously considered we choose $F_{s,p}$ for maximal possibilities for F_0 within our region of interest. The results are shown in table 4.2.

With this design each N covers approximately the same range, however the high orders have 2 or 4 times as many samples as $N = 128$. This is especially important at high values of F_0 where the quantization is the worst.

We revisit harmonic SIR and modulation depth with non-deal pitch estimates. Downshift quantization is assumed: $f_q = \frac{F_s}{N}$.

Order (N)	$F_{s,p}$	**min** F_0	**max** F_0	**best quantization**	**worst quantization**
128	4kHz	63Hz	500Hz	1Hz	56Hz
256	8kHz	63Hz	533Hz	1Hz	33Hz
512	16kHz	63Hz	533Hz	1Hz	17Hz

Table 4.2: F_0 estimate quantization

4.7.1 Harmonic SIR

Harmonic SIR is visualized for two different filter design methods in figues 4.24 and 4.25. The pitch quantization, which is worse for lower orders, causes harmonic SIR to degrade for higher harmonics. This makes sense as the quantization error will be scaled by harmonic index k.

The hanning filter performs slightly at high F_0s bettter due to narrower filter bandwidth. Depending on the desired performance, harmonic indices above a certain threshold will no longer provide accurate harmonic envelopes. This threshold is slightly lower for adaptive filters than hanning filters and it is significantly lower for lower order filters.

We now consider the same designs but with ±5Hz pitch estimation error. The worse case SIR_k is shown for hanning filter in figure 4.26 and for adaptive filter in figure 4.27.

The error degrades performance in two dimensions. Similar to quantization error, the perfomance degrades proportional to k. The other problem is at low values of F_0, where harmonics are more closely spaced.

We can take the right plot in figure 4.26 as an example. For the first 3 harmonics we get good harmonic SIRs for $F_0 > 80$Hz, however for $k = 3, 4$ this increases to roughly $F_0 > 180$Hz and for even higher harmonics we never achieve satisfactory SIR.

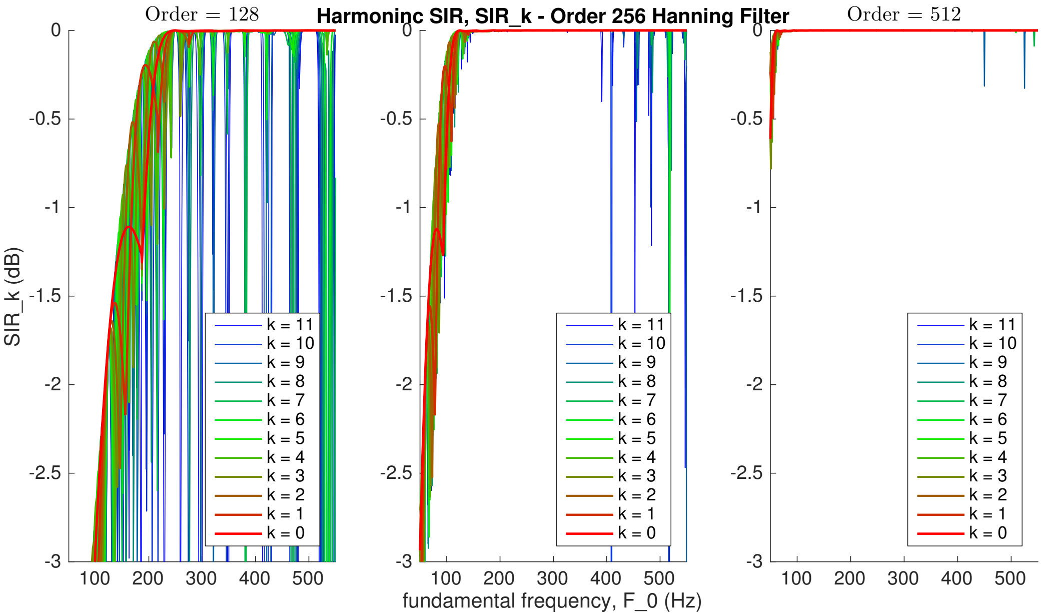

Figure 4.24: SIR_k, hanning filter and pitch estimate quantization

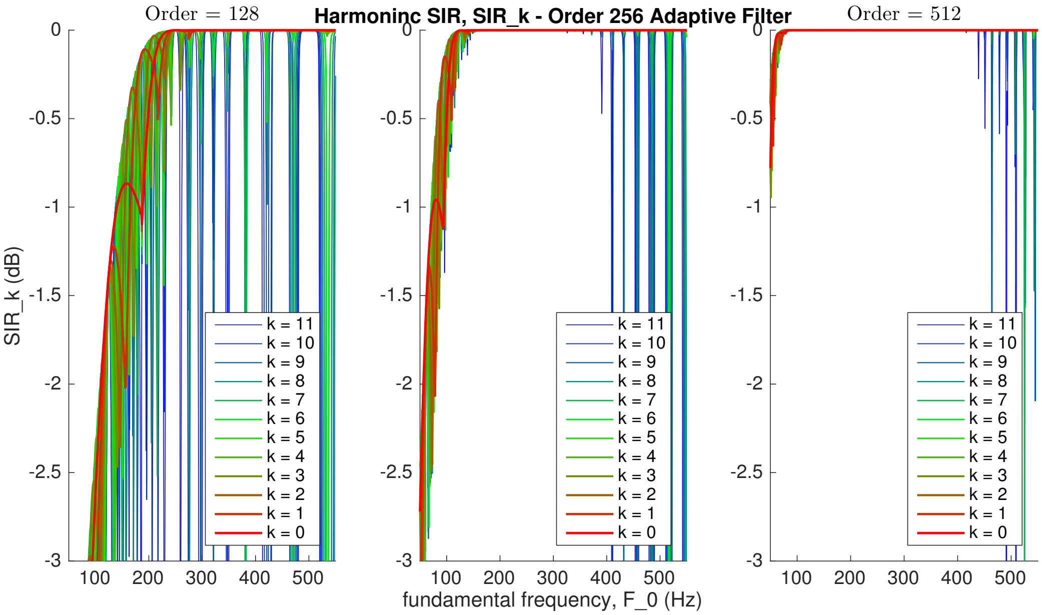

Figure 4.25: SIR_k, adaptive filter and pitch estimate quantization

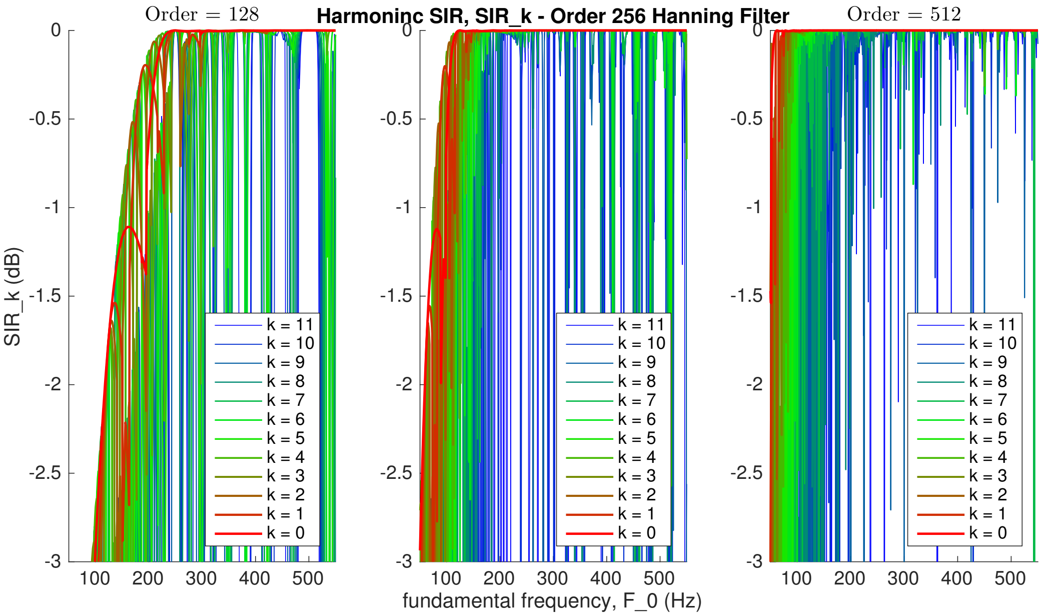

Figure 4.26: SIR_k, hanning filter, pitch estimate quantization and ±5Hz estimation error

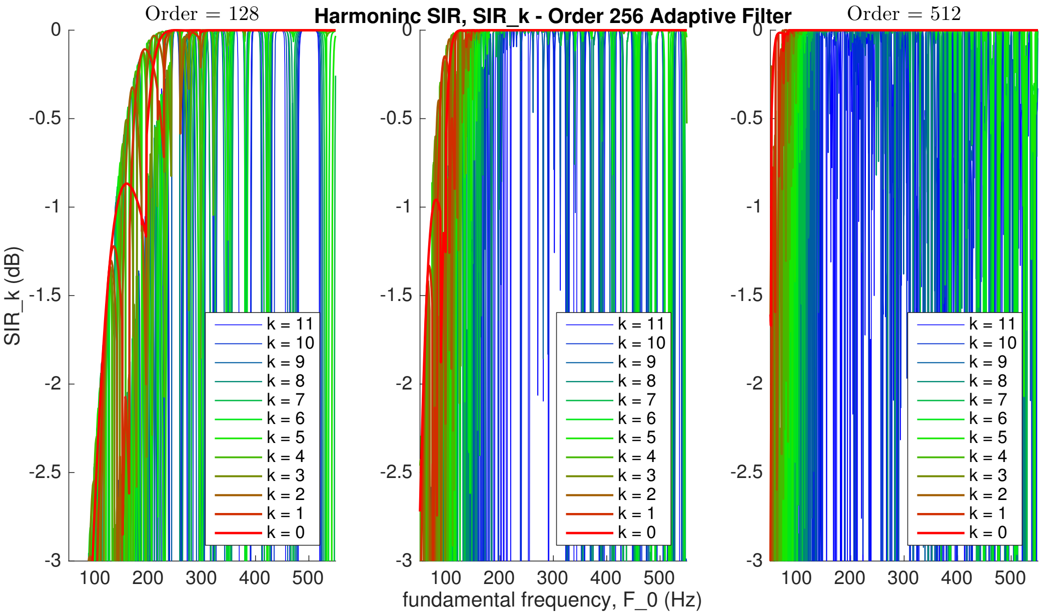

Figure 4.27: SIR_k, adaptive filter, pitch estimate quantization and ±5Hz estimation error

4.7.2 Modulation Depth

We repeat these same comparisons for modulation depth. Looking at figure 4.28, with hanning filter and pitch estimate quantization, high harmonics have very high modulations. Around the 6th harmonic ($k = 5$) we start to see big spikes in modulation depth at high F_0. Interestingly the same harmonics have poor performance regardless of N, however there is a far broader region of failure for lower N.

In figure 4.29 we see much better performance for $N = 512$ in comparison to the hanning filer. This is because despite having wider bandwidth at high F_0, the sidelobes are much lower than the hanning filter. The first hanning sidelobe has a gain of -31dB, whereas for $F_0 = 500$Hz the adaptive filter has a first sidelobe gain of -56dB.

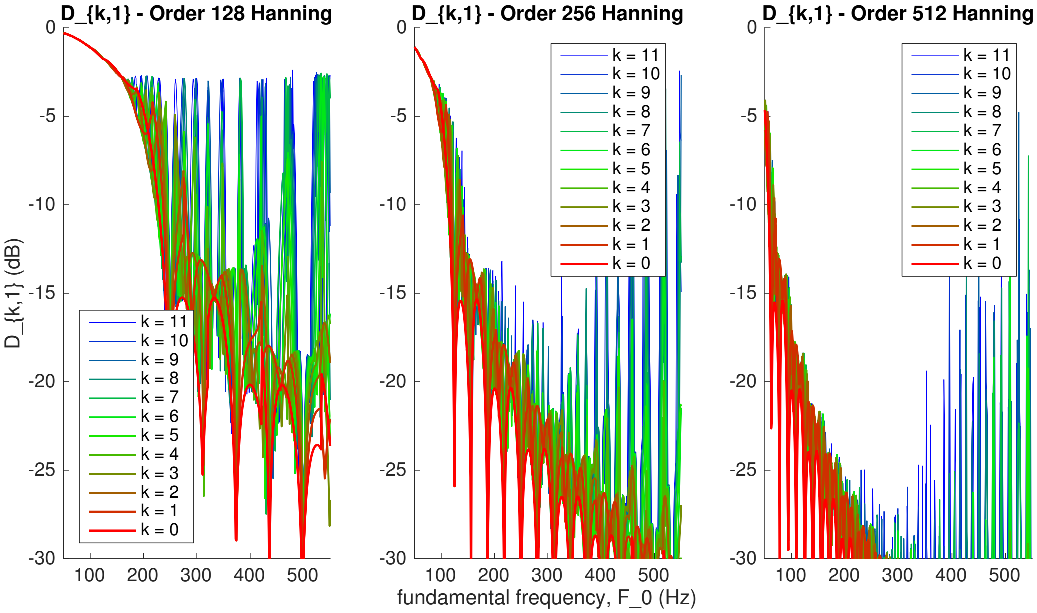

Figure 4.28: $D_{k,1}$, hanning filter and pitch estimate quantization

Now considering ±5Hz estimation error, we see from figures 4.30 and 4.31 the same shift right where higher harmonics at low F_0 perform worse. The adaptive order 512 filter performs the best, being very robust error.

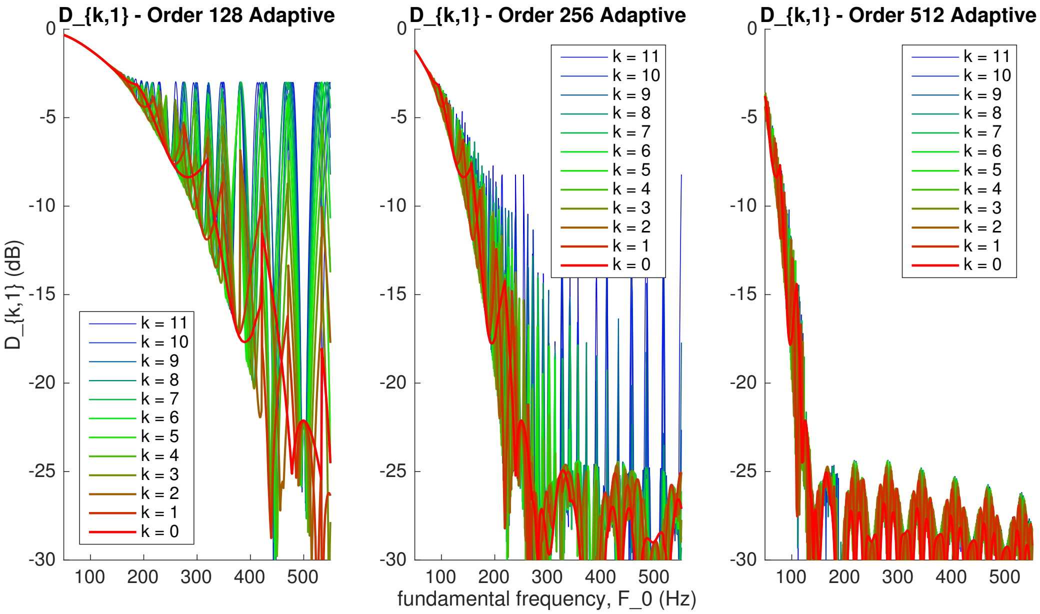

Figure 4.29: $D_{k,1}$, hanning filter and pitch estimate quantizationr

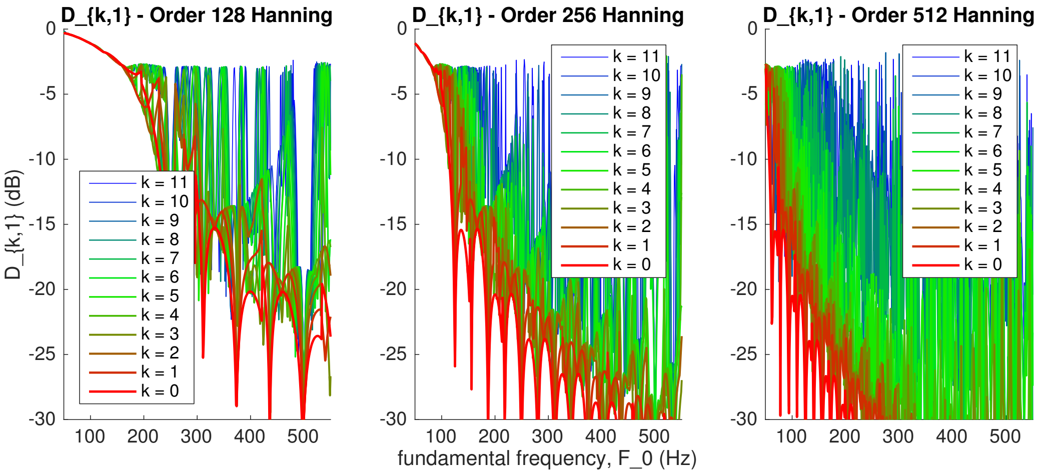

Figure 4.30: $D_{k,1}$, hanning filter, pitch estimate quantization and ±5Hz estimation error

Regardless of extraction method there will be a fundamental limit on performance as the pitch estimate becomes worse. We see from all of the above example that performance degrades proportional to $k \times \text{error}/F_0$.

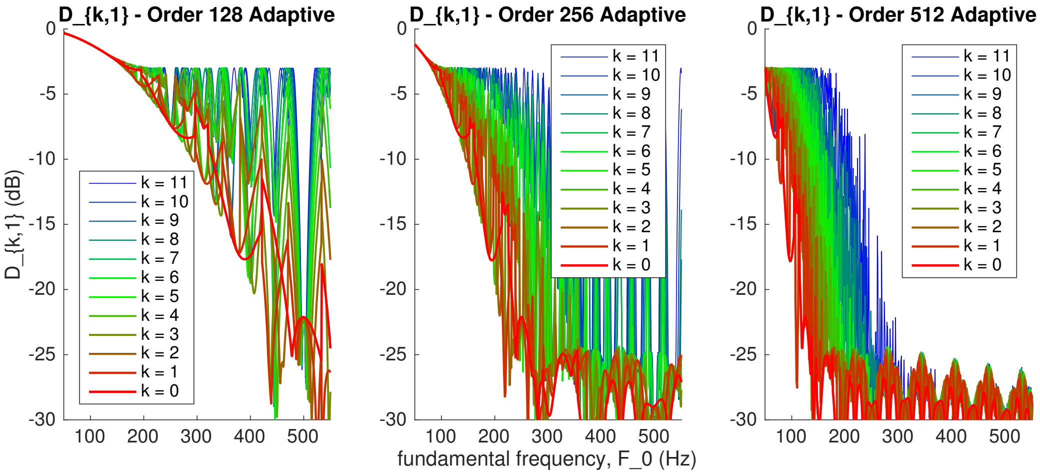

Figure 4.31: $D_{k,1}$, hanning filter, pitch estimate quantization and ±5Hz estimation error

Chapter 5
CONCLUSION

5.1 Mapping and Selection

Fixed Greenwood bands are determined offline, corresponding each electrode with a bandwidth. The N envelopes are then mapped to electrodes by finding the greenwood bands each harmonic falls within.

Two general solutions

1) Adaptive (select loudest)

similar to ACE, we can choose the loudest channels. This suffers from stability issues. We can apply another heuristic to stabilize the decision based on consistency of signal energy and fundamental frequency

2) Fixed

stable, each option suffers from missing key harmonics to the signal

lowest channels will imply no high frequency energy, which could be bad for unvoiced signals

other relationships such as odd harmonics or prime numbered harmonics could miss harmonics critical to timbre perception.

fixed is more computationally efficient!

how to allocate harmonics to channels: combine, select 1st, select largest

hint at hybrid via failure to represent of selection options to represent the high channels well.

how to choose channels, largest vs fixed

heuristics on both!

Because we have isolated individual harmonic envelopes there is no issue of signal energy falling in between channels.

Another bonus to HSSE is that we may add a simple heuristic to maintain channel mapping stability. For example, if F0 has not varied significantly since the previous frame, we can allocate to the same channels to avoid unnecessary switching between channels induced by vibrato or inaccuracies in pitch estimation.

5.2 Implementation Considerations

5.2.1 Carrier Synthesis

what waveform?

raised vs rectified, (raised seems better)

cite that 4-waveform paper

Swanson thesis: “A high-rate pulse train, modulated on and off at frequency F0, had a higher pitch than a train of pulses at the rate of F0. If amplitude modulation of high-rate pulse trains is to be used to convey pitch, then the shape of the modulating waveform is important: a half-wave shape is better than a square-wave (on-off) shape.”

5.2.2 Pitch Estimator

importance of a good pitch estimator

improving pitch estimator

haven’t yet mentioned octave errors

5.2.3 Adaptive Filters using FFT

...

improved filter quantization using interpolation

5.2.4 Efficient FFT Interpolation Algorithm

FFT with changeable window, and interpolate

Can this be done with different filter as function of F0? We probably need to design the filters such that they pass reconstruction requirements

Is the actual equation just a sinc function times a phase shift?!

READ THIS: [An Intelligent FFT-Analyzer with Harmonic Interference Effect Correction and Uncertainty Evaluation]

5.2.5 Stimulation Rate

“ it was important not to change other 12 HWR strategy take-home study 224 aspects of the strategy, in particular, stimulation rate. It would not be a fair comparison to trial HWR at 1800 pps against ACE at 900 pps, as the increased stimulation rate in itself could affect performance. A higher rate could potentially represent amplitude modulation cues more faithfully (McKay et al. 1994). Conversely, there is evidence that sensitivity to temporal modulation is worse at higher rates (Galvin and Fu 2005).” [swanson thesis]

5.2.6 encoding transients

explicit transient encoding using transient detectors [find a vocoder ref]

5.2.7 Hybrid Methods

Most everything so far has assumed the signal has an F_0, what if it doesn't? What if it is well outside the boundaries of F_0? What about polyphonic music? What about SNRs below what is needed for accurate F_0 estimation.

hybrid considerations

1) to achieve harmonic and inharmonic at same time

2) to better model the critical bands in the cochlea

maybe narrower filters could improve SiN, this was not investigated

effective information bandwidth: should I really get into this? From the theoretical standpoint, envelope extraction is exactly the same in ACE and F0mod. In implementation

ACE typically uses a lower order FFT. In [laneau] the authors consider 128-point for ACE and 512-point for F0mod and both will be considered here. with respect to bandwidth we actually have to different things, filter bandwidth and effective information bandwidth. The former is obvious, the later refers to what frequencies are encoded on a electrode channel. If multiple narrowband filters are somehow combined on the same channel, they may have the same information bandwidth as one wideband filter.

soft decisions (e-Tone)

5.3 Summary

5.4 Future Work

Appendix A
DERIVATIONS

$$\begin{aligned}
\phi_0[n+r] =& \phi_0[n] + 2\pi\frac{F_0[n]}{F_s}r, \quad 0 \le r < N && \text{(A.1)}\\
m_{k,harmonic}[n] =& \left| x[n]e^{-jk\phi_0[n]} * \frac{1}{Nw[0]}w[-n] \right| \\
=& \frac{1}{Nw[0]}\left| \sum_{r=-\infty}^{\infty} x[n-r]e^{-jk\phi_0[n-r]}w[-r] \right| \\
\text{Let} \quad & r' = -r \\
=& \frac{1}{Nw[0]}\left| \sum_{r'=0}^{N-1} x[n+r']e^{-jk\phi_0[n+r']}w[r'] \right| \\
=& \frac{1}{Nw[0]}\left| e^{-jk\phi_0[n]} \sum_{r'=0}^{N-1} x[n+r']e^{-j\frac{2\pi F_0[n]}{F_s}kr'}w[r'] \right| \\
=& \frac{1}{Nw[0]}\left| e^{-jk\left(\phi_0[n]-\frac{2\pi F_0[n]}{F_s}n\right)} \left[e^{-j\frac{2\pi F_0[n]}{F_s}kn} \sum_{r'=0}^{N-1} x[n+r']w[r']e^{-j\frac{2\pi F_0[n]}{F_s}kr'} \right] \right| \\
=& \frac{1}{Nw[0]}\left| X\Big[n, \frac{N}{1}\frac{F_0[n]}{F_s}k\Big) \right| \\
=& \frac{1}{Nw[0]}\left| X\Big[n, \lambda[n]k\Big) \right| && \text{(A.2)}
\end{aligned}$$

VITA

Tyler Ganter grew up in upstate New York, where the long cold winters inspired him to pick up guitar. While pursuing a BSEE at the University at Buffalo he took an interest computer programming, which he decided to minor in. Through a continued interest in mathematics, software and music he found himself at home in the field of audio digital signal processing. The desire to delve deeper into this field has led him to study at the University of Washington.