# A PERCEPTUAL REPRESENTATION OF AUDIO

by

Daniel Patrick Whittlesey Ellis

B.A.(hons) Cambridge University, 1987

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements
for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

5th February 1992



Signature of author _____
Department of Electrical Engineering and Computer Science
5th Feb 92

Certified by _____
Barry L Vercoe
Professor of Media Arts & Sciences, MIT Media Laboratory
Thesis supervisor

Certified by _____
Thomas F Quatieri
Research Scientist, MIT Lincoln Laboratory
Thesis supervisor

Accepted by _____
Campbell L Searle
Chair, Department Committee on Graduate Students

# Abstract

## A PERCEPTUAL REPRESENTATION OF AUDIO

by Daniel Patrick Whittlesey Ellis

Submitted to the Department of Electrical Engineering and Computer Science on Feb 5th 1992, in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science

The human auditory system performs many remarkable feats; we only fully appreciate how sophisticated these are when we try to simulate them on a computer. Through building such computer models, we gain insight into perceptual processing in general, and develop useful new ways to analyze signals.

This thesis describes a transformation of sound into a representation with various properties specifically oriented towards simulations of source separation. Source separation denotes the ability of listeners to perceive sound originating from a particular origin as separate from simultaneous interfering sounds. An example would be following the notes of a single instrument while listening to an orchestra. Using a cochlea-inspired filterbank and strategies of peak-picking and track-formation, the representation organizes time-frequency energy into distinct elements; these are argued to correspond to indivisible components of the perception. The elements contain information such as fine time structure which is important to perceptual quality and source separability . A high quality resynthesis method is described which gives good results even for modified representations.

The performance and results of the analysis and synthesis methods are discussed, and the intended applications of the new domain are described in detail. This description also explains how the principles of source separation, as established by previous research in psychoacoustics, will be applied as the next step towards a fully functional source separator.

Thesis supervisors:

| | | |
|---|---|---|
| **Barry L Vercoe** | title: | Professor of Media Arts & Sciences, MIT Media Lab |
| **Thomas F Quatieri** | title: | Research Scientist, MIT Lincoln Laboratory |

# Acknowledgements

First and foremost my thanks go to Barry Vercoe for persuading me that there was anything interesting about perception in the first place, but also for his faith, support and advice over the two and a half years I have worked in his Music & Cognition group at the MIT Media Lab. I could not imagine a more agreeable supervisor. Both he and the many others instrumental in creating the Lab deserve enormous credit for having built such a marvellous place to work and learn.

Secondly I am eternally indebted to Tom Quatieri for his limitless patience and conscientiousness in discussing this work with me regardless of his many other commitments, and his unfailingly solid advice even though my own weaknesses sometimes prevented me from taking full advantage of it.

I want to thank all my colleagues from Music & Cognition, past and present, for having provided a very rare kind of stimulating and supportive environment. Particular thanks to Alan, Andy, Betty-Lou, Bill, Dave, Janet, Jeff, Joe, Judy, Kevin, Marc, Mike H, Molly and Shahrokh. However, every member of the group must take credit for the atmosphere since each one has a part in creating it.

For moral and emotional support, I am very grateful to Bill Gardner, Amy Bruckman, Dave & Kim Rosenthal, Alan Ruttenberg, Paul Resnick, Karen Flint, Paul Smith and Graham Johnson. Special thanks go to Bill for being an unbelievably tolerant officemate, and as an endless source of excellent technical advice and insight, including very valuable comments on this thesis. Special thanks also to Amy for tolerance of my neuroses beyond the call of duty.

For material support during this work I am indebted to the Commonwealth Fund of New York, in particular Roy Atherton and Robert Kostrzewa, for the Harkness Fellowship that originally brought me to MIT. I am also grateful for support from Nichidai corporation, the Television of Tomorrow consortium, and ESL.

Finally my thanks and love go to Sarah, so much a part of me I cannot see which parts to thank you for. Suffice to say that they are the essence and core that allow me to go on working.

# Contents

# 1 Introduction

When we try to build computer systems that mimic human functions of perception and cognition, we rapidly discover a paradox: "easy things are hard" [Mins86]. In trying to build a model of a process one is aware of performing, it becomes apparent that there are several basic components required that were not even recognized as part of the problem. Minsky introduces it in reference to accounting for "obvious" constraints such as not reusing blocks from the bottom of a tower to extend it at the top. But it applies in the fields of visual and auditory perception with equal validity and for the same reason : particularly difficult processes cannot be left to the conscious mind, but instead require the evolution of special-purpose machinery. This separation means that we are not consciously aware of that machinery's function. Since the role of perception is to inform us about the outside world, and since it is successful most of the time, we often overlook the distinction between reality and our perception of it, so the processes of perception are completely ignored.

The motivation for this thesis is the desire to have a computer that can 'understand' sound in the same way as a person. By this we do not mean language recognition, but refer instead to something that occurs at an earlier stage : before any sound can be recognized, it must have been isolated and identified as a complete and distinct entity, and not several simultaneous sounds, nor some subcomponent of a more complex sound. This is the kind of unexpected and very difficult problem that, at first, we might not have expected to need to solve.

This thesis considers how to program a computer to perform *source separation*. It describes our first stage of such a program, the initial analysis and representation of sound. Source separation is the process of hearing a mixture of overlapping sounds from different physical sources and being able to detect this multiplicity of origins, and indeed focus one's attention on a meaningful subset. Examples of this range from continuing a conversation in a noisy room (the 'cocktail party effect'), to following the part of a single instrument in an ensemble, to hearing an unseen bus approaching over the noise of other traffic. In this thesis, we describe a transformation of sound into a new representation which is suitable for the application of known principles of source separation. These have been discovered

over the past few decades through carefully designed psychoacoustical experiments. Although the transformation is all we have built, we will describe in detail why and how it is the necessary first stage of a full simulation of human processing of sound.

One reason for giving computers human-like perceptual capabilities is to allow them to interact with people in a natural manner. Current speech recognition systems are easily confused by background noise, and this can only be overcome by good emulations of sound segregation. Building a successful separator will require us to gain a deep understanding of how sound information is used and represented within the brain. This understanding will have many other benefits ranging from sophisticated compression schemes that transmit only the truly important aspects of a signal, through to new sound production methods allowing composers to evoke novel and precisely calculated responses from their audiences.

## 1.1 THE GOAL OF THIS WORK: A REPRESENTATION FOR SOURCE SEPARATION

Source separation is a process of extracting particular information from a signal, and would *appear* to fall firmly into the field of signal processing. However, the kinds of operations involved in this and other examples of cognitive information processing turn out to be rather different from the rigorous and uniform techniques we normally encounter in engineering.

There are many components required for a full simulation of auditory signal processing, and we list some of them at the end of chapter 2. It was beyond the scope of the current project to build a complete source separator, but that continues to be our ultimate goal. The first step towards such a goal is to consider the nature of the internal representations used for processing sound in the brain. The innovation of this thesis is the design of a transformation of sound that is a useful model of these representations. It has the following key properties:-

- The sound is analyzed into *perceptually atomic elements*, i.e. subsets of the sound that we assume will always be perceived as coming from just one source, and broken up no further during perceptual organization. Each element is representing a particular 'packet' of acoustic energy, and it is regarded as indivisible because this energy has such uniform frequency and amplitude over its duration that the only reasonable hypothesis is that it was generated by a single physical process. In order to satisfy this minimum

assumption, many of the elements will be necessarily small compared to complete acoustic events, but they are still very large compared to the samples of the continuous representation of the sound that they describe. In this way, forming these elements may be considered a structuring or organizing of the acoustic information. This move towards a more symbolic representation of a continuous physical property is a common feature of models of perception and cognition [Dove86].

- The transformation must exhibit and encode all perceptually significant cues. This requires it to have *fine time resolution*, since short time effects such as 'jitter' have been shown to be important in distinguishing sources [McAd84]. Such qualities should follow automatically from our perceptual modeling; we mention it specifically because it is a weakness of other representations such as the phase vocoder [Dols86] or the McAulay-Quatieri Sinusoid Transform System [McAu86].
- The representation is intended to be *invertible* to a perceptually equivalent sound. This feature is particularly motivated by our goal of an ideal source separator as described below. It is a motive for homogeneity and against nonlinearity in the processing stages. It was necessary to balance this motive against our desire for an accurate model of auditory processing, which is far from linear.

## 1.2 A MORE RIGOROUS DEFINITION OF SOURCE SEPARATION

We can make an initial list of the kinds of information that may be employed when human listeners separate sources:-

a) spatial cues (both binaural such as interaural delays, and monaural such as pinna-derived coloration) i.e. perceived physical origin of the sound;

b) recognition of a learned sound or pattern - employing specific experience;

c) visual cues (e.g. lip reading);

d) non-spatial cues to common origin somehow intrinsic to the sound.

Of these (d) seems to be less obvious, but as is so often the case with introspection regarding perception it is possibly the most basic effect. We can imagine a musical example to demonstrate that such intrinsic cues do exist: It is easy to generate a fragment of sound using computer synthesis which will leave most listeners in confident agreement over the number of different 'sources' or instruments involved - even though the sound was generated solely by the computer with unfamiliar tone qualities and perhaps played through a single speaker. Thus all of the above strategies except (d) can be defeated, since there was no spatial distinction between the sources, the different voices did not have a previously-learned quality, and there was no visual component. Yet by employing sufficiently distinct timbres and organizing them into well-defined (if overlapped) events, the impression of a specific number of instruments is successfully conveyed. It is this problem, the recognition and tracking of distinct events, that must be solvable through the representation developed in this thesis; as a convenient restriction of the domain, we are only going to deal with this final class of intrinsic cues.

To further clarify the problem, we can sketch a block diagram of auditory source separation:-

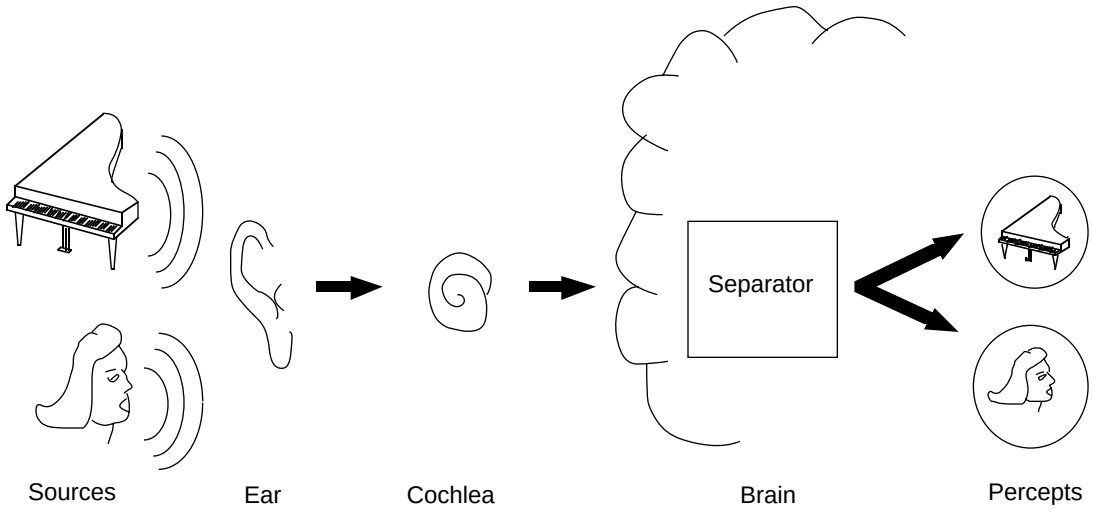



*figure 1.1 - Schematic block diagram of monaural source separation*

The points to note about this figure are :-

- There are (just) two separate sources of sound on the left, but they are superimposed into a single sound channel before reaching the ear.
- There is only one ear.
- We have drawn the images in the brain (the 'percepts' of the different sources) as identical to those sources, but they are only impressions or reduced representations.

This last point deserves clarification. The percepts in the brain do not contain all the details of the source to which they relate. There are many minor modifications we could make to a source and have it generate an identical percept. (We can establish that two percepts are identical by asking listeners if they can tell the difference between them). Thus we are not suggesting that the brain or any source separator can analyze a mixed sound back into the precise waveforms that were added together; this is mathematically impossible in general. However, the brain does form some idea of what the individual sources would sound like alone, and it is meaningful to try and reconstruct a signal out of the mixture which exhibits all the qualities attached to the brain's partial impression.

So this is the paradigm towards which the representation developed in this thesis is directed: a black box with a single input accepting complex sound mixtures, and a

series of outputs each presenting the signal for just one of the input components. Such a realization is many years away, but we have argued that it will eventually be possible and that we are moving the right way towards it.

## 1.3 AN OUTLINE OF THIS THESIS

This is chapter one, the **Introduction**, which presents our basic orientation and introduces the problems we addressed and why they were of interest.

Chapter 2, **Background**, surveys the theoretical underpinning of this work and other research in the field of source separation. It then considers what will be the necessary parts of any computerized source separator.

Chapter 3, **Design overview**, introduces the main stages of our transform . Each is considered in more detail in the subsequent chapters (4 - **Filterbank design**, 5 -**Track formation**, 6 -**Track resynthesis**).

Chapter 7, **Results and assessment**, considers the nature of some analysis and synthesis by the system, and chapter 8, **Comparison with the Sinusoid Transform**, contrasts it specifically with that system, upon which it is heavily based.

Chapters 9, **Track domain processing**, and 10, **Track-based source separation** present our as yet unimplemented plans for employing the representation.

Chapter 11, **Conclusions and future work**, ties up by summarizing the success of the project and highlighting some important outstanding issues for attention.

# 2 Background

## 2.0 INTRODUCTION

In this chapter we survey the previous results upon which the current work is based. We start with results of investigations into the human auditory system, from low level sensory processing through to recent results in perceptual sound organization. We then consider previous work in separating simultaneous sounds. Finally, we assemble some basic prerequisites for any computational system capable of imitating human source separation.

## 2.1 PRIMARY AUDITORY SENSATION

It is generally accepted that hearing is a process involving many layers of information analysis and processing. One is repeatedly surprised by the differences between perceived qualities of sound, filtered through these layers, and objective measurements of related properties of the sound, revealing how seamlessly our perceptual machinery presents its own construction of reality as the thing in itself. Despite this large gap between 'real' and perceived sounds, we can at least chip away at the very lowest levels in the chain with the tools of dissection, direct measurements of nerve firings, and perceptual experiments designed to defeat the interference of the impossibly-complicated higher layers. Several decades of experiment are beginning to give us a reliable understanding of the basic information extracted from a sound and passed to the subsequent stages.

This is not the place for a detailed tutorial on the physiology of the ear (see instead the section on preprocessing of sound in [Zwic90]) but rather to summarize what are considered the significant results. Sound incident upon the two ears is converted into nervous impulses by the hair cells of the cochlea. The hair cells detect motion of the basilar membrane, which is mechanically organized to distribute the spectral energy of its excitation along its length; thus the firings of a nerve connected to a particular hair cell show a band-pass response to the input signal. The density of firings for a particular fibre vary with the intensity of the stimulus over a certain range.

Inconsistencies in measurements of absolute threshold and frequency selectivity make a passive, linear model of the cochlea impossible, and it implies the existence of active nonlinear feedback mechanisms to improve sensitivity for quiet sounds -- not

least because a small proportion of fibres in the auditory nerve are carrying information *from* the brain *to* the cochlea. Just how much influence higher-level perceptual processes can have on the properties of the cochlea is open to debate; [Zwic90] presents a feedback model that is local to the lower levels but very successful in explaining several auditory phenomena.

Asking subjects when a particular tone becomes audible can give consistent results which we may assume to reflect the ultimate performance of this basic machinery (for instance, the absolute threshold in quiet across frequency). Similar experiments with combinations of sounds reveal the important phenomenon of *masking*, where stimulation at a certain point of the basilar membrane by a 'masker' sound reduces the sensitivity in that region so that other sounds, audible in the absence of the masker become unnoticeable or 'masked' when it is present. This effect extends somewhat before and after the strict temporal extent of the masker, known as temporal masking as distinct from simultaneous masking. These effects have been exploited very successfully to hide quantization artefacts in recent audio compression schemes [Veld89].

A curiosity in perception is that the relative phase of sine tones is largely imperceptible, except where the tones fall into the same critical band, i.e. are perceived as an interfering unit rather than separate partials [Patt87]. None the less, perception of spatial direction is strongly influenced by small phase differences between tones at the same frequency presented to different ears [Blau83], so this information is being represented at some level but not at others.

## 2.2 SUBSEQUENT AUDITORY PROCESSING

Once we start trying to understand auditory function above the basic level, it becomes much harder to make sense of the results. A recent contribution to this field which has been a major influence on the current work is Bregman's *Auditory Scene Analysis* [Breg90]. In this he presents the results of many years of experiments into how people organize simple auditory sensations into higher level perceptions.

Bregman investigates the phenomena of *fusion* and *streaming*, and the competition between them. 'Fusion' refers to the process of distinct elements of acoustical energy (such as different harmonics) being assembled into a single, often indivisible, auditory sensation (such as a note played on a musical instrument). 'Streaming' is related to selective attention; if a series of acoustic events can be concentrated upon

amidst other events which are ignored, the events are said to belong to a single stream. Since this ability varies considerably with intention, it is typically investigated by generating signals that the subject cannot help but hear as several streams i.e. they cannot 'listen' to the sounds as if they are coming from the same source. That different streams are irresistibly formed can be confirmed by the surprising result that it is very hard to make judgements about the relative timing of events in different streams.

The intention is to build a comprehensive set of the rules and principles that govern the conversion of raw sound stimulus into these percepts of 'event' and 'stream'. The model is that the sound at the ears is broken down into its 'basic elements', and then subsets of these elements are 'recognized' as some particular event, thereby giving a fused percept. In a typical experiment using gated sine tones, each simple tone is a basic element; it turns out that in most cases such an element can only contribute to one percept : once it is fused, it is unavailable for other recognizers.

The properties that lead to elements being fused or assembled into a single stream, as described in [Breg90], are:-

- Spatial location i.e. sounds perceived as originating in the same place;
- Common fate i.e. features shared between elements reflecting their common origin. This has two main instances: first, common onset and offset, where components appear and/or disappear at the same time, and second, common modulation where different components have synchronized and parallel changes in frequency or intensity.
- Harmonicity - partials that fall into a pattern as harmonics of some common fundamental pitch;
- Proximity in pitch and loudness (similarity) between successive elements or events;
- Conformity to some previously learned sound pattern.

These characteristics are related to the ideas of perceptual grouping from Gestalt psychology, but Bregman nicely justifies them on the basis of what he calls *ecology*; hearing is not some bizarre and arbitrary sense, but exists simply for the evolutionary advantages it confers. This ensures that within the constraints of available data and processing potential, it will do a good job of building perceptual objects that closely reflect actual external events, these being the inescapable facts

that we need to deal with to survive. Any spurious or error-prone rules of grouping would make the sense much less useful. But the rules that work depend totally upon the kinds of stimuli which we have to deal with i.e. the environment that we have evolved to survive in, so these rules must reflect profound or reliable regularities in the sounds of the 'real world'[1].

## 2.3 PREVIOUS WORK IN SOURCE SEPARATION

The problems of crosstalk in telephony and other communication systems, as well as the apparent tractability of the problem have led to a number of efforts at separating simultaneous sounds. Initially, many of these focussed on separating two human voices sharing a communication channel, and within that problem set about segregating the harmonic spectra of two superimposed vowels based on their different fundamental pitches [Pars76]. This is a nicely limited problem, but despite a fair deal of attention has had limited success because it ignores many of the other cues employed by human listeners.

This was improved in [Quat90] by using interpolation between unambiguous time frames to resolve poorly conditioned frames that occurred between them. This relied on having some sense of the continuity of peaks in time, furnished by their use of the McAulay-Quatieri Sinusoid Transform System (STS). Clearly the constraint that separation of voices must be consistent across time is a powerful aid to success. The STS encodes sound as the maxima in its short-time Fourier transform, and forms links between peaks in adjacent time frames considered to be continuations. As such, it is the basis of the representation developed in this thesis. The STS was also used in [Mahe89] to separate musical signals in a duet i.e two superimposed signals with reasonable time stability.

The orientation of this project is rather more grandiose than the highly constrained double-vowel problem. The problem of general-case auditory organization has received less attention, but in the past few years it has been addressed by several researchers, mainly inspired by the ideas of Bregman described above. Cooke

---

[1]There may be a clue to one 'reason' for music in this realization: our machinery of auditory organization is 'tickled' by the fact of several instruments making sounds with aligned onsets (i.e. in time) and common harmonics (in tune), violating many of the main assumptions for distinguishing sources.

[Cook91] is addressing this problem, and cites Marr's work [Marr82] in visual perception and organization as a good example to follow. Cooke has built a system to model general human auditory organization and implemented several grouping heuristics, but without such a strong emphasis on resynthesis. Although he does convert his representations back to sound, it is more as a diagnostic than a serious attempt to create an equivalent perception. Cooke's 'synchrony strands' bear a stark resemblance to the time-frequency energy tracks of the current work, although they are derived in a very different manner based on firing-rate synchrony of a complex auditory model [Patt91].

Mellinger [Mell91] has also sought to build a computer model of auditory organization, focussing on feature detectors and the architectural problems of combining their outputs. He has made a careful consideration of the problems of combining the information from different cues, as well as in depth consideration of what certain of those cues would be and how they should be used, again inspired by Bregman. Mellinger employs the very detailed and accurate auditory model developed by Lyon [Slan88], which has also been used as a basis for modelling perceptual grouping by comodulation in [Duda90].

## 2.4 REQUIREMENTS FOR A FUNCTIONAL SOURCE SEPARATOR

In this section we will describe several elements that we consider prerequisites of any simulation of human auditory source separation on a computer. It is intended that this discussion be free of considerations tied to a particular implementation, although it would be difficult to claim to be without interest or bias. However, since the problem is defined in terms of the performance of existing human listeners, we would expect limited flexibility in possible approaches; it will be difficult to simulate this skill in much detail (especially in the cases where it is simplifying or mistaking the actual sound origins) without taking essentially the same approach as our brains.

The principles described by Bregman are very appealing in that they seem consistent with our experiences and they are well supported by experiment. But if we are to employ them, we must have some method of breaking down input sounds into

indivisible elements; the principles are essentially synthetic i.e. they relate to the *reassembly* of such elements into plausible separate sources[2].

One important ecological principle is known as 'old plus new' : when the sound signal 'changes', it is most likely that *only one* event has occurred i.e. either one source has stopped, or one has been added. In the latter case, we can ascertain the qualities of the added source by forming the difference between what we hear and what we 'expected' to hear based on the 'old' sound. This important principle presents difficulties in realization since it requires a sophisticated sense of what to expect, well beyond a simple first-order difference along each input channel.

Research in both vision and hearing supports the significance of features -- qualities of a signal defined over a local region. The general principle seems to be that a wide variety of heterogeneous features is employed in concert to reach conclusions about the signal they reflect. More important than the detection of any particular class of feature is the problem of integrating this wide variety of information to give meaningful composite results. Much of the robustness of human perception, clearly a high priority in evolution, comes from its independence from any one kind of information and flexibility to integrate different cues. Good computational solutions to this have yet to be developed; at the moment we are typically reduced to projecting everything onto a single dimension (most often a probability of that set of observations under a particular hypothesis) which seems too narrow.

Probably the hardest part of any complete source separator will be the simulation of the functions served by memory and experience in human listeners. It is not clear how well we would be able to organize and segregate composite sounds if we did not already have a good idea of the character of the individual sources based on previous examples. We can imagine many roles of previous knowledge in hearing. One is the construction of special recognizers that are attuned to a particular sound i.e. if the pitch and decay sound like a plucked guitar string, then a recognizer might look for and capture the noisy 'pluck' sound at the beginning which it has learned to expect,

---

[2]It is possible to imagine the auditory system working in a more *analytic* way; for instance, by initially assuming that the whole sound was one source, then gradually introducing divisions to remove contradictions. These two approaches, synthetic and analytic, might be difficult to distinguish by their successful results alone.

even though there may be competing streams (based, for instance, on similarity) to which the pluck could belong.

Another function of this kind of knowledge, known as 'auditory restoration' [Hand89], is even harder to include in a machine that seeks to regenerate the sound components as perceived by the human listener. If the guitar pluck noise was essentially masked by some simultaneous loud event, the auditory system will typically make its best estimate that the pluck was there even if it was not specifically identified, and very often the segregated percept will include this 'assumed' sound just as if it had been clearly evident in the sound -- the listener will not be aware of the illusion. Bregman describes an even more alarming version of this effect: Take a simple spoken phrase, replace an easily-inferred phoneme by silence, play the modified sound to a listener with some kind of noise burst masking the deletion; the deleted portion will be perceptually restored, implying that feedback from the highest levels of language processing influences perceived sound.

So we must conclude that a fully functional source separator needs to solve the rather tough problems of recognition (to make a recognizer for specific sounds) and, indeed, language understanding! This kind of co-dependency is a general problem in modelling human cognitive processes; as engineers, we like to break up problems into independent simplified modules to be solved one at a time. The human information processing function, freed from constraints of conscious design or debugging, has evolved to be hugely interrelated, obeying no rules of modularity or stratification, making its analysis vastly more demanding.

# 3 Design Overview

## 3.0 INTRODUCTION

This chapter introduces the analysis-synthesis system, explaining the basic design motivations and the broad relations of the separate parts. Chapters 4 through 6 then describe these parts of the system in more detail.

## 3.1 DESIGN MOTIVATION

In order better to explain the design, it is worth describing some of the major influences. Primarily, this system is modelled after the McAulay-Quatieri Sinusoid Transform System (STS), which it closely resembles [McAu86]. But there were several aspects of existing STS implementations inappropriate for the intended function of auditory modelling. Most significantly, the STS employed a fast Fourier transform (FFT) based frequency transform, resulting in fixed-bandwidth frequency resolution -- a very significant divergence from the observed behavior of the cochlea. For this reason, we used a bank of bandpass filters with constant Q instead (i.e. a fixed ratio of center frequency to bandwidth), with all the attendant implications of multirate output signals.[1].

One specific objection to the consistently narrow bands of the FFT in the STS is that they sacrifice fine time resolution, coding it instead in the phases and amplitudes of more slowly moving harmonics. Yet evidence of source segregation based on short-term variations in period (jitter) suggests that time-domain features important to source separation exist on a scale smaller than that observed by the STS, but easily captured in the high frequencies of a constant-Q filterbank. This comparison is expanded in chapter 8.

---

[1]The uniform bin bandwidth B of an ST model means that each band must be sampled every 1/B seconds, so that the data can be stored as parameter points for all active tracks at each multiple of 1/B. For a variable bandwidth system, the required sampling rate for a track will depend on its exact frequency at that moment, and will therefore be different for each simultaneously active track : the data can no longer be organized along a common time base.

Our system is particularly interested in breaking sound into spectral energy concentrations with a direct correspondence to perceptual objects, since we are planning on modifying the sound at the level of these objects; this is not a concern for the STS. In our system, we would like to be able to resynthesize any single element of the representation, and have the resynthesis clearly sound like part of only one sound, not the combination of several sounds in the original. At the other extreme, we would also like this single-element-resynthesis to avoid being stripped down so far as to be completely meaningless and denatured, but to be recognizable as a particular component of the input sound. In a system that lacked this property, the combination of several elements might be required before the nature of the original sound emerged, as is the case for for high harmonics in a fixed-bandwidth system who exhibit properties of the original sound only through mutual interference. The motivation for breaking up a sound into such perceptually indivisible or atomic elements is of course to create a domain suitable for the application of Bregman-style rules of sound organization.

## 3.2 BLOCK DIAGRAM

The complete system is represented by the diagram below:-

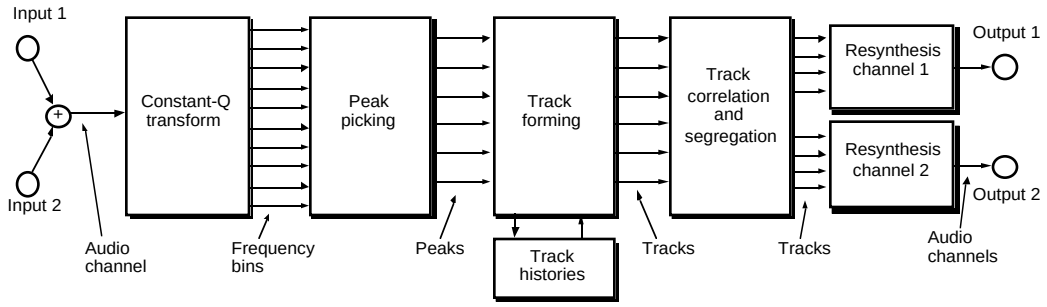



*figure 3.1 - Block diagram of the overall system*

The time waveform of the input sound, here shown as the sum of two sources to emphasize the application to separation, is converted into excitations in different frequency bands by the constant-Q transform (described in chapter 4, filterbank design). This generates spectral magnitudes at different sampling rates as a result of the different bin bandwidths. These are combined into instantaneous spectra, which are then reduced to their peak values alone, discarding information on the shape of the peaks or the valleys in between. Peaks from successive spectra are organized into tracks, which collect all the peaks from a contiguous energy concentration in

time-frequency. (These functions are described in chapter 5, track formation). Thus the one dimensional sampled sound is converted via a two-dimensional sampled time-frequency distribution to a set of discrete tracks standing for what are identified as the most important parts of the sound.

The next block, track correlation and segregation, is an example of the kind of processing that could be applied to the sound in the track domain. Chapter 9, track domain processing, describes some general ideas for this stage, and chapter 10 deals specifically with processing for signal separation. None of these has been implemented, but as they are the motivation behind the representation, it is important to include their description.

The remaining modules show resynthesis of plain sound from the processed tracks. This is accomplished with simple sinusoid oscillators following the magnitude and frequency of each track. This simple scheme gave very satisfactory results, so more sophisticated inversion was not required. Chapter 6, track resynthesis, describes this function.

The nature of the tracks formed and of the resyntheses are presented in chapter 7, results and assessment.

# 4 Filterbank Design

## 4.0 INTRODUCTION

The conversion of a single-dimensional variable, instantaneous sound pressure, into a two dimensional distribution of time-frequency energy is a very beneficial strategy on the part of our perceptual equipment. For while a single neural pathway is limited to detecting variations up to at most a few tens of events per second, the resonant hair cells of the ear allow us accurately to characterize sound pressure variations almost one thousand times faster than this. Clearly, the cochlea, by performing this radical transformation, must influence very strongly the entire auditory modality; the same is true for the corresponding module in any computer simulation. The poor match between the ubiquitous Fast Fourier Transform and the mammalian cochlea was one of the main motivations behind this project.

Research has created a wealth of knowledge describing the basic function of the cochlea. These results are obtained by direct experimentation upon and dissection of cochleas from cadavers, neurological experiments on the auditory nerve of live nonhuman mammals (mainly cats) and psycho-acoustical experiments on human subjects. This gives us the basic limits of intensity sensitivity and frequency discrimination etc. It has proved slightly more difficult to assemble these results into a single model, since the ear has many nonlinear properties, and indeed may vary its characteristics under the control of higher brain functions (via the efferent fibres of the auditory nerve). Nonetheless, many computational models have been built exhibiting good agreement with particular experimental results.

The filterbank used -- a uniformly-overlapped, constant-Q array -- was chosen as the best match to these results that still retains, in some loosely defined sense, linear homogeneous features. A more accurate model would vary its behaviour across the frequency axis and incorporate a time-varying nonlinear magnitude response, making inversion much harder. The exponential frequency sampling of a constant-Q filterbank tallies with our intuitive perception of pitch distance - an octave sounds like the 'same' pitch distance over a wide range of frequencies. This is an equivalent definition of a constant-Q filterbank.

## 4.1 CHOICES IN FILTERBANK DESIGN - FREQUENCY AXIS

We had three choices for the frequency axis : uniform spacing (as returned by the Fast Fourier Transform), exponential spacing (a constant-Q or wavelet transform) or perceptually-derived spacing (for instance the Mel or Bark scales, which are somewhere between the other choices). The principle advantage of uniform spacing is the efficiency of calculation afforded by the FFT; a secondary advantage is that each filtered subband has the same bandwidth and can thus be sampled at the same rate, leading to a simpler architecture. The disadvantage is that this is significantly different from the transform effected by the ear. The choice of a psychoacoustically-derived scale is, implicitly, the closest we can come to modelling the ear. Since it has an empirical basis, it has no simply expressible properties and must be calculated and downsampled explicitly for each subband. The compromise of the constant-Q filterbank is a good approximation to standard cochlea models (at least between 1kHz and 6kHz) but has a simple mathematical characterization. This simplicity gave us no processing advantage; since we evaluated the filter outputs by direct convolution, a totally arbitrary filterbank (or one based more closely on cochlear responses) would have been almost as easy. Instead, by using a transformation that was mathematically homogeneous, it was possible to deal with the data in a more uniform fashion. In this question of the frequency axis, it is unlikely that the difference between exponential frequency and a 'true' psychoacoustic scale would have any substantial impact on our results.

## 4.2 WINDOW SHAPE

Having chosen the basis for spacing our filterbank elements we must choose the shape of each filter in either the time or frequency domain. We then have to choose the number of filters per octave, but this is largely defined by the filter contour since we will be looking for uniform frequency-response crossover between adjacent bands, with the ultimate objective of being able safely to interpolate between our frequency samples.

Filter shape is once more a compromise between psychoacoustic accuracy and simplicity of implementation and interpretation. One common area of concern related to filter shape is sidelobes. In general, a filter with finite time support will have a main lobe in its frequency response flanked by sidelobes of a certain prominence. Finite time support is unavoidable if the filters are to be implemented by the convenient and flexible method of direct convolution (FIR filters), but even

low-magnitude sidelobes are particularly undesirable for an algorithm that is sensitive to local maxima in the spectrum, such as the peak picking of the next chapter. These sidelobes, although presumably reduced in amplitude compared to the mainlobe, will tend to form parasitic tracks of their own and clutter the track based processing. Thus smoothness was more of a priority than the more commonplace consideration of compact frequency support (narrow mainlobe). In the figure below, a typical 'narrow' window shows its many flanking magnitude peaks, whereas the Gauss window shows side peaks only at the noise floor.

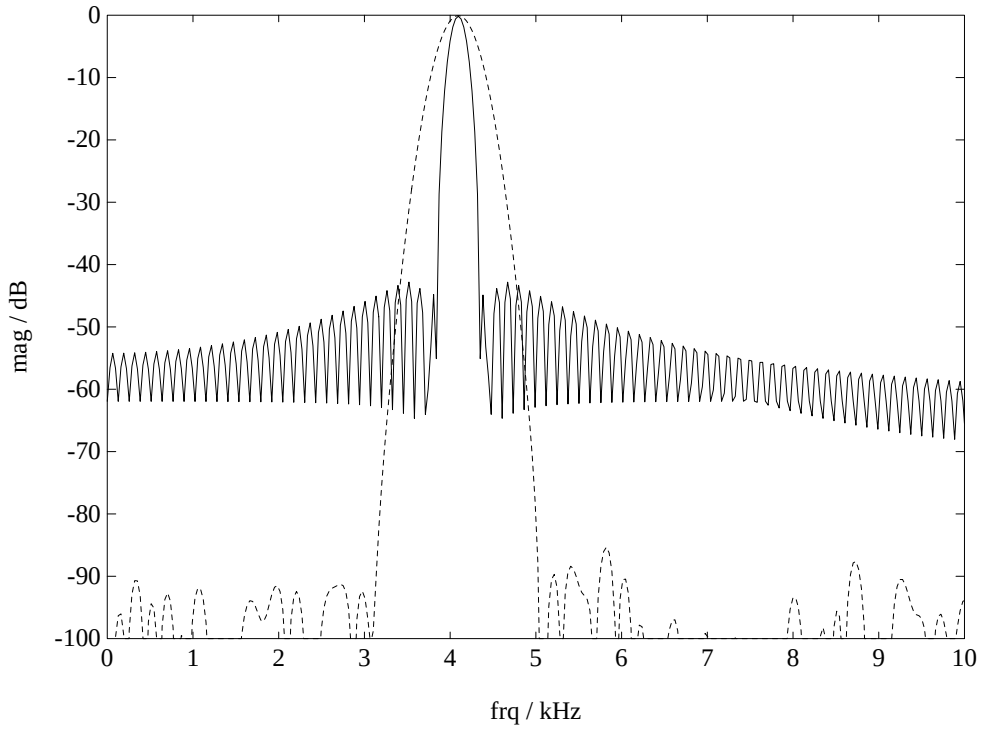



*figure 4.1- Comparison of spectra of 255 point Hamming (solid) and 511 pt truncated Gauss (dotted) windows showing comparison of mainlobe width and sidelobe height. Gauss sidelobes result from quantizing to 16 bits.*

Having chosen our time support (i.e. the number of points in our convolution kernel), we could have chosen from the Kaiser window family to push our sidelobes down to some arbitrary noise floor and taken whatever mainlobe width that gave us. Instead, we used a rectangularly-windowed Gaussian envelope - a sub-optimal compromise between sidelobes (from the rectangular windowing) and mainlobe, but an easy window to design and predict. A true, unbounded Gaussian has the attractive property of smoothness in both time and frequency (no sidelobes). The corresponding

disadvantage is that it is finitely supported in neither domain. The combination of Gaussian width and FIR length we used gave us the smoothness we desired.

The characteristic Q of the filterbank was chosen to approximate observed cochlear responses. This is a slightly ill-defined situation, since hair cells show a considerable variation in tuning as a function of level. Also, the ear's filter responses differ significantly from our Gauss envelopes, and neither is particularly well characterized by a conventional measurement of Q using -3 dB bandwidth. Glasberg and Moore's equivalent rectangular bandwidths for auditory filters suggest a Q of about 10 [Glas90]. As an ad-hoc match between the different filter shapes, we specify our Gauss filters by requiring the bandwidth at the $1/e^2$ or -17.4 dB points to be one-quarter of the center frequency. This gives a -3 dB bandwidth-based Q of 9.6.

The magnitude responses of one octave's worth of filters and the composite impulse response for a 6 octave system is shown below. See appendix A for a mathematical description of the filterbank system.

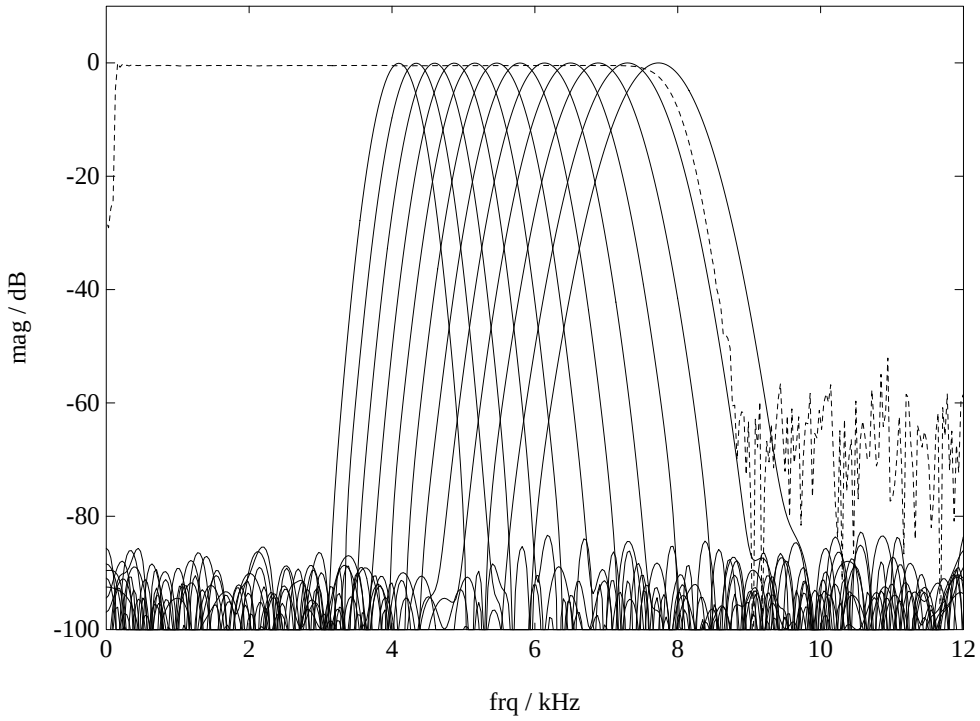



*figure 4.2- Magnitude responses of one octave of individual filters (solid) and composite impulse response of six octaves (dotted).*

## 4.3 MULTIRATE IMPLICATIONS OF VARYING BANDWIDTH SUBBANDS

Since we are dealing with discrete-time systems, sampling rates are always a consideration. As we mentioned above, a fixed bandwidth filterbank requires the same sampling rate for each of its output channels for a consistent Nyquist margin : this is typically effected by calculating an entire short-time spectrum, using an FFT, at the period required for the sampled subbands. For the kind of system we are describing where each subband has a different bandwidth, each will require a different sampling rate for a consistent aliasing margin. Note that the smooth Gaussian window approaches zero rather slowly, so we cannot talk about a Nyquist rate that completely avoids aliasing (other than that of the original sampled signal). Instead, we choose a low but acceptable aliasing distortion, and seek to apply it consistently to each of our subbands.

We stated above that the filters we used were designed to have a Q of 4, subject to our particular definition of bandwidth. This would imply a critical sampling rate of the center frequency divided by the Q (i.e. the same as the bandwidth, since it is a 'two-sided' bandwidth). In practice, the Gauss filters pass considerable energy outside of this band; to avoid problems of aliasing at this early stage, we doubled this sampling rate for safety to make our sampling rate for a given band half the center frequency. As can be seen in figure 4.2, a 4kHz band around the 8kHz-centered bandpass response (i.e. 6kHz to 10kHz) contains all the mainlobe energy down to the -90 dB noise floor This was very conservative, but furnished useful leeway in interpolation along time.

This requirement for unique sampling rates for each subband complicates the implementation. Instead, we reach a compromise between consistent sampling and system complexity by breaking the subbands into blocks one octave wide (from $f_0$ to $2f_0$), and sample all the subbands in that range at the same rate. (This constrains our filterbank spacing to provide an exact number of bands in each octave, which is acceptable.) We then have a set of sampling rates where each is exactly double and half of its lower and upper neighbors respectively, and each octave block of subbands is sampled consistently with the others.

## 4.4 FILTERING

As mentioned above, the filters were implemented as sine and cosine modulated truncated Gaussian windows. The filtering was implemented by direct convolution of the filter kernels with a section of signal.

## 4.5 RECURSIVE DOWNSAMPLING

As we have described, the subbands were arranged into octave-wide blocks with a separate sampling rate for each. The filter kernels for the prototype octave (the highest) were precalculated and applied as an ensemble against a segment of the input signal. This segment was then hopped forward to the next appropriate sample time for that octave's sampling rate, and the next set of coefficients for that octave's subbands was calculated.

Subsequent (lower) octaves would have required both filter kernels and (since the subband sampling rate was halved) input signal hops twice as long. But a simpler alternative was to low-pass filter and downsample the original signal by a factor of two, then apply the exact same top-octave processing to this new signal. By changing the sampling rate of the signal, the effective location of the filters moves down exactly one octave, and the fixed hop-size translates in to the correct, doubled sampling period. This process can of course be repeated for as many octaves as desired.

## 4.6 STORAGE OF SAMPLES

The fact that data is produced at different rates in different frequency bands presents difficulties for storing these results, for instance on disk. Given that we are going to want to reconstruct complete spectra from particular instants, we will need to access all the subband samples from the different octaves relevant to that instant. But since each octave is generating samples at a different rate, we need to index into each octave's sample stream at a different point. We could store each octave in a separate file, and then keep separate pointers in to each, but if we want to store the data in a single (linearly-organized) data file, we need a scheme to interleave the blocks from the different octaves so that information local to a particular time is stored contiguously in the data file. Apart from anything else, this means we can perform analysis of arbitrarily long input files in a single pass, with finite buffers for input and output. The block sequence was derived from the order in which the different

sized frames at each octave were completed, given that the initial frames have the centers of their time supports aligned, as shown in the following diagram:-

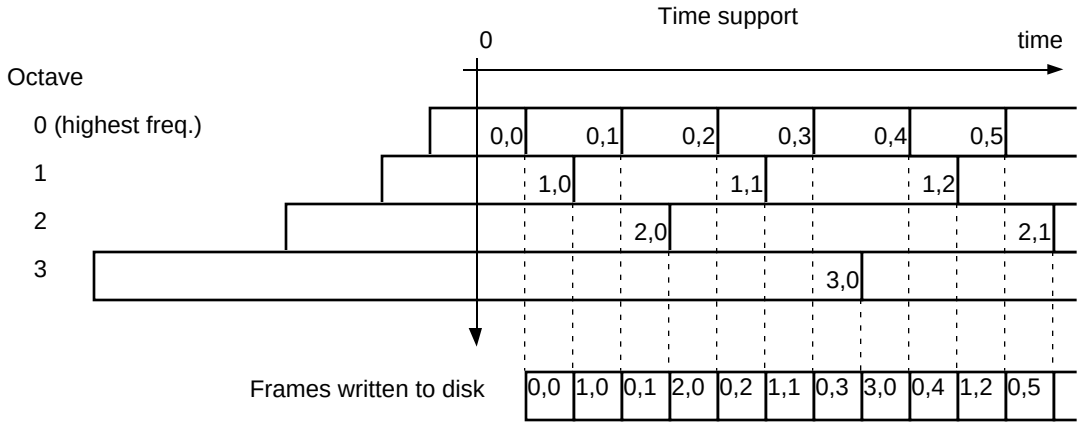



*figure 4.3 - The ordering of spectral data frames on disk. The upper diagram shows how each octave has a different region of time support. The lower boxes show the frames in order on disk. The order is derived from the order of conclusion of the frames. The pairs of numbers refer to the octave followed by the sequence index.*

## 4.7 RECOVERY OF A GIVEN SPECTRUM

To find the spectrum at a given instant in the sound, we need to interpolate the sampled streams for each subband. Each octave block has a series of time samples, and the interval between samples depends on the octave. Interpolation to calculate a close approximation to the exact filter output at a point between two samples is accomplished with a dynamically-calculated FIR filter applied to the eight nearest values; for each octave, these eight values for each bin in the octave can be obtained by grabbing the eight relevant coefficient frames spread variously over the disk as in fingure 4.3. Since a similar situation exists for each octave, but with different distributions of frames on disk, accessing the various frames can get complicated and indeed present quite a challenge to the disk controller. Therefore, a single, integrated function was implemented that caches the coefficients to reduce the expected disk accesses. This cover function handles all the interpolations and returns a complete spectrum precisely centered at the requested time. This makes using the filterbank output files very convenient, although potentially inefficient if the requests do not conform to the expectations of the caching logic.

# 5 Track Formation

## 5.0 INTRODUCTION

The output from the filterbank described in the last chapter is still a complete representation of the sound (albeit across two dimensions, frequency and time); with an appropriate set of inverse filters, we can completely reconstruct our input signal (within the frequency extent of our filterbank, and ignoring any aliasing that may have been introduced in downsampling the subbands). But one of our design goals was to build a representation that was *not* complete, insofar as certain details of a sound have negligible perceptual impact; these irrelevant features should not be encoded.

The next stage of processing implements such a perceptually-sufficient, imperfectly invertible transformation. The basic assumption is that at some low level the human auditory system is 'looking for' sonic elements which are stable in energy and frequency over some time interval. In our two-dimensional distribution, this corresponds to 'ridges' of locally maximum energy approximately parallel to the time axis. If we are correct to assume that such features form the basis of sound perception, then we can record everything we need to about a sound by characterizing these details alone. This is the approach we implement.

## 5.1 RELATION TO AUDITORY PHYSIOLOGY

One recurrent phenomenon in neurology is *lateral inhibition*, where the output of a bank of nerves mapping a perceptual dimension is made more selective by a positive feedback mechanism : the output of the most strongly stimulated element inhibits outputs from its neighbors and thus increases its relative prominence [Sham89]. We can imagine this occurring along the frequency axis of the auditory nerve. The expected result would be that the presence of a strong spectral peak would make the ear less sensitive to sounds of about the same frequency. This is indeed the case, and this experimental result is the well-known *critical-band masking* effect [Zwic90]. While there is no sufficient, direct model of critical band masking by lateral inhibition (in particular, the effect is not strictly local in frequency), the phenomenon certainly supports the general principle that if a spectral *peak* is accurately

reconstructed, one need not worry unduly about the adjacent spectral *valleys*, since they will fall under the masking envelope. (Critical band masking has been applied more directly to sinusoid models of speech by [Ghit87]).

This supports coding instantaneous spectra as simply a collection of peak frequencies and amplitudes. The second part of tracking is to group successive peaks along time into tracks. There is no compelling neurological evidence to support this association, but there is psychoacoustical evidence that two separate time-frequency events will be perceived as a single sound if they are joined by a frequency glide [Breg90]. We can also make an 'ecological' argument that, barring coincidences, continuity in time-frequency must arise from a common source.

## 5.2 PRAGMATIC MOTIVATIONS

The principle gain from this approach is the reduction in complexity. First, by coding the instantaneous spectra by its peaks alone, we are reducing the raw data. The exact proportion depends both on the signal and the intrinsic smoothness of the spectrum (arising from the window transform) but must be better than 50% (a peak needs a valley to separate it from its closest neighbor). Note that this is not strictly an argument for data compression in terms of raw bits (that possibility has not been pursued in the current work). Rather, we are simply observing a reduction in the number of parameters, regardless of resolution.

Secondly, by organizing these frequency peaks into tracks, or contours through time, and stipulating that these contours shall be treated as units through subsequent processing, we massively reduce the number of choices that have to be made, for instance by a source separation algorithm. We can imagine an algorithm that goes through every time-frequency sample of a complete representation and allocates all or part to each possible output source. A similar algorithm in the track domain is only required to make choices once for each track. There are very many fewer tracks than time-frequency samples, since each track can contain a significant amount of information. This processing advantage arises because we have assumed that each track comes from only a single source.

Let us consider the case when a track is formed by an erroneous amalgamation of signal energy from two sources. If this track is assigned wholly to one source, there will have been incorrect energy assigned to that source. If the captured regions are small and errors are few, we may be able to get away with it -- the degradation of the

reconstructed source may be minimal. But if a long track is formed by merging different harmonics significant to both sources, the seriousness of the impact of such an error should make us very careful about how much variation we allow within a track before ending it and starting afresh, and also perhaps make us provide some mechanism for breaking tracks later on in the processing if several other cues suggest it.

On the whole, we might feel tempted to make our system very conservative in the matter of building long tracks i.e. tracks representing a significant chunk of spectral energy that is going to be treated as indivisible henceforth. For instance, with an extremely narrow criteria for ending a track i.e. deeming a particular time-frequency energy concentration too irregular to be coded as a single object, we might end up representing a component that is really part of a single continuous event as three sequential tracks. It might be easier to successively allocate these three fragments to the same source than to recognize and segment three sounds that have been mistakenly amalgamated into a single track. But the disadvantages of unnecessary breaks in the tracks are significant. Firstly, it makes the subsequent processing more arduous (since the number of elements to be considered has grown). Secondly, it increases the possibility of error, since there may not be sufficient cues in a fragment taken alone to establish its true source; had it remained rightfully attached, its source might have been obvious. Thirdly, it turns out that the phase continuity algorithm used to resynthesize a given sinusoid component fares much better (in terms of artefact-free output sound) with continuous tracks than for those with frequent breaks. Phase continuity can only be applied *within* tracks, and the breaks may give rise to spurious noise from the uncoordinated offset and onset of the tracks on either side. In short, we have many reasons to be careful to get the track formation 'right' to the best of our ability.

## 5.3 IMPLICATIONS OF THE UNDERLYING CONSTANT-Q TRANSFORM

The multirate nature of the underlying time-frequency distribution, where each frequency has a different critical sampling rate, has a similar effect on the description of the tracks. For a track at a fixed frequency, it is necessary to store samples of the track's frequency and magnitude at the rate determined by the bandwidth of the filter passing that frequency -- i.e. at the sample rate of the underlying representation. This will of course avoid aliasing the amplitude-modulation information of such a track, since we are essentially copying the output of

the fixed-frequency filter onto the magnitude contour of our track. Potential aliasing of the frequency modulation contour is a more worrying problem -- it is not immediately obvious what the intrinsic bandwidth limitation of the peak frequency of a ridge will be. However, since a periodicity in the *frequency* contour will require the same periodicity in the *magnitudes* of the bands it crosses, we can argue that the bandwidth of frequency modulation will be limited to the (magnitude modulation) bandwidth of the associated subbands[1]. Note that for a track that moves in frequency, this means that the sampling rate for that track will vary between samples, i.e. it is *nonuniformly sampled*.

The qualitative nature of the tracks that are formed should be distinguished from those that would result from a fixed-bandwidth model. Since the filter bandwidths may vary by several orders of magnitude across the signal's spectrum, we see very different kinds of tracks. Consider pitched voice: At the low frequency, we see the slowly moving sustained fundamentals familiar from narrowband analysis. But at the high end we see pitch-pulse-aligned energy bursts at the formant peaks that would appear poorly modelled by sinusoids. (To some extent, this is a question of scale : when we expand the time dimension sufficiently, we see a picture broadly comparable to our previous low-frequency tracks). For the tracks modelled on the outputs of the broad filters, the time sampling is extremely rapid (to capture the full bandwidth) and a great deal of information is being carried in the amplitude modulation. This situation is different from similar models (such as the McAulay-Quatieri STS) which try to keep the amplitude contour relatively slow-moving (narrow bandwidth). However, the rapidly modulated tracks of the current model must have a correspondence to excitation of the similarly broadly-tuned hair cells in the cochlea. Later, we may speculate about how this wideband information is bandlimited by the neural channels.

---

[1]The magnitudes of the subbands must periodically show peaks each time they are crossed by the frequency contour i.e. each time that band contains the local maximum energy along the frequency axis. This argument implies that the upper band on the periodicity of subband magnitude (the subband bandwidth) will also be the upper bound on the *period* of any frequency modulation. However, to use it as a basis for the sampling rate ignores the possibility of harmonics higher than the fundamental in the frequency contour, which has not been precluded. Despite this, the assumption has been successful in practice.

## 5.4 DETAILS OF THE IMPLEMENTATION

This section describes the technical detail of how the tracks were grown on the constant-Q time-frequency transform.

## 5.5 PEAK PICKING

Picking peaks from spectral slices is a relatively common problem (see [McAu86], [Serr89]). However, the first obstacle is actually obtaining the spectrum itself, since each frequency band will, in general, have a different set of sampling instants. Fortunately, this is dealt with by the front end to the multirate spectral data files described in the previous section -- it handles the loading of the relevant data frames and the necessary interpolation between time frames at different octaves.

We then have a spectrum, sampled across frequency, of our input sound, centered at some specific instant. The sample points are exponentially spaced, but since the filter bandwidths are uniform in such a projection, the spectrum is 'consistently' sampled -- plotting the samples on a uniform axis gives a curve that is consistently 'wiggly' over its full extent. The use of Gaussian filters which fall off relatively slowly makes choice of sampling always difficult, and we have tended to err on the side of caution. Even so, the low Q factors of the filters (typically with center frequency four times the bandwidth measured at the $1/e^2$ or -17.4 dB points) make 12 samples/octave an adequate coverage, so each spectrum is only 72 or 84 points (6 or 7 octaves).[2]

The peak finding algorithm looks for a simple local maximum (magnitude larger than two neighbors). Although longer windows were tried (i.e. a peak as largest of five or seven local points), the simplest three-point approach gave entirely satisfactory results. A parabola is fit to the three points, and the exact frequency and magnitude

---

[2]The sampling argument here is to consider a spectrum of arbitrarily high resolution (requiring arbitrarily dense sampling for accurate representation) as having been convolved with the frequency magnitude envelope of the Gauss filters. (These envelopes will be uniform across frequency on the logarithmic axis of the spectrum under consideration.) This convolution will limit the frequency content of the function representing the spectrum thereby allowing it to be adequately represented by a finite sampling density in frequency. This concept of the frequency content of the spectrum H(w) treated as just another function, is reminiscent of cepstra, although it does not involve a logarithm term.

of the peak are taken as the maximum of this curve . Note that this is interpolating along the *frequency* axis.

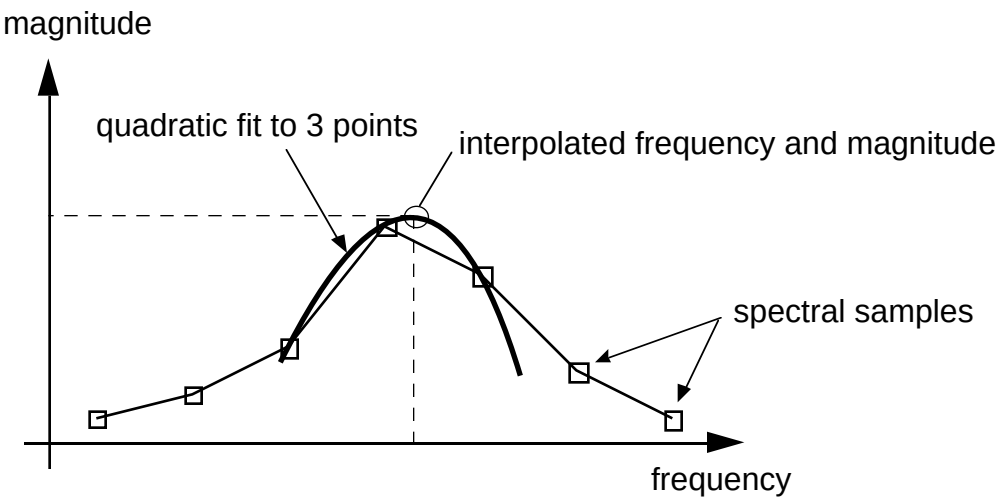



*figure 5.1 - Locating a peak by fitting a quadratic to three points, and finding the maximum of the fitted curve.*

We are also interested in the phase of the spectrum at the peak, since we wish to use phase to guide our resynthesis. This is a little harder to interpolate, since the phases of our spectral samples have been folded into +/-π (principle values). We make a linear interpolation of the phase between the two nearest samples, but record *two* phases being the two most likely unwrapping interpretations. This normally includes the 'correct' phase, or as close as we can get to it. The problems of phase interpolation and reconstruction are discussed further in the next chapter which describes track resynthesis.

## 5.6 TRACK FORMING

The process described above turns an instantaneous sampled spectrum into a list of peaks. The next job is to organize these peaks into tracks -- lists of peaks at successive times which appear to be part of a contiguous time-frequency energy concentration.

The basis of the algorithm is as follows, derived from [McAu86]. At any moment, we have a list of the tracks which are currently active (specifically, the peak records for the most recent sample in each track). We then advance our time pointer by one click (discussed below) and fetch a new peak list for this new time. We then go through a

matching algorithm that seeks to associate the new peaks with the peaks at the ends of the existing tracks.

The matching is achieved by a simple algorithm similar to that used by the 'diff' utility in unix. The 'distance' between to peaks is calculated simply on the basis of their frequency separation, although it might be desirable to introduce a magnitude-matching element as well for certain ambiguous cases. The existing and new peaks are sorted by frequency into two lists, and matches made by simultaneously descending both lists and making links between local minima in distance. While there are some situations where this will not generate the maximum number of links, it has proved sufficient.

After this matching, we are left with three lists : existing peaks which have been linked to new peaks, existing peaks for which no extension was found, and new peaks that were not allocated to any existing peak. The tracks for which no new peak was found are regarded as having finished and are removed from the list of currently active tracks. The new peaks that were unclaimed are seen as the beginnings of new tracks, and are added into the list of current tracks as such. This then completes one time tick, and after the peak trimming, which we describe below, the process repeats.

As described in section 5.3, the sampling interval required for a track is the sampling rate of the bandpass filter lying under the track's instantaneous frequency. This filter sampling rate is chosen as half that frequency, but the particular filter involved, and hence the sampling rate, will vary as the track moves in frequency. The sampling interval doubles for each octave descended in frequency, giving a factor of 64 difference between largest and smallest possible intervals in a six-octave system. But the single tick of the algorithm described so far will add peaks to each track at the highest possible sampling rate. To avoid unnecessary oversampling of tracks below the very highest band, a procedure is employed to remove excess samples : If the time interval between the previous and new peaks is less than the minimum sampling interval for that track at that time, the new peak is deleted, and the previous tail is left as the basis for the match in the next time frame. The process of deletion may repeat for several ticks until enough time has cumulated to warrant a new sample for that track.

Having a single time tick for the whole spectrum represents a considerable computational inefficiency in the creation of the tracks. Since the tick must be rapid

enough to allow complete sampling for the broadest frequency bin, it examines the narrowest bands much too often. It would be possible to compute the tracks far more rapidly by having several different basic sampling rates for tracks in different frequency bands. This process is of course complicated by the fact that each track has a different ideal sampling rate, and the gains in execution time were not of sufficient interest for this to be pursued.

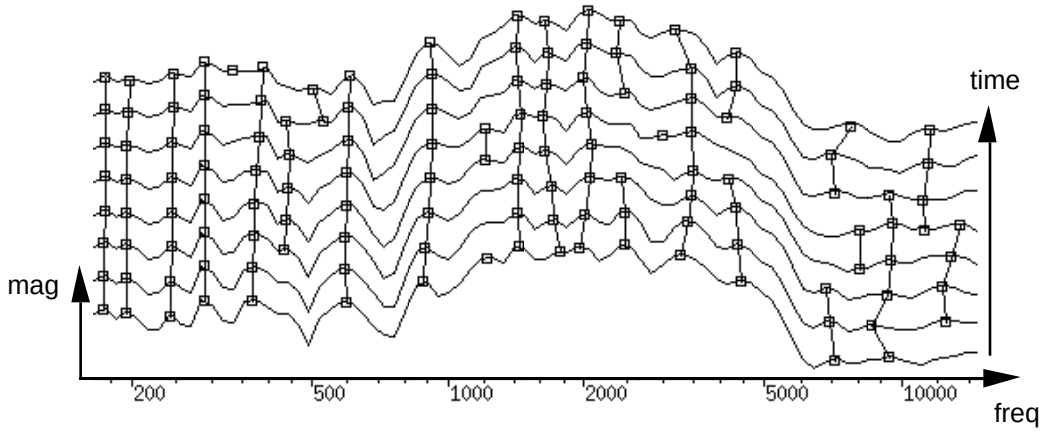
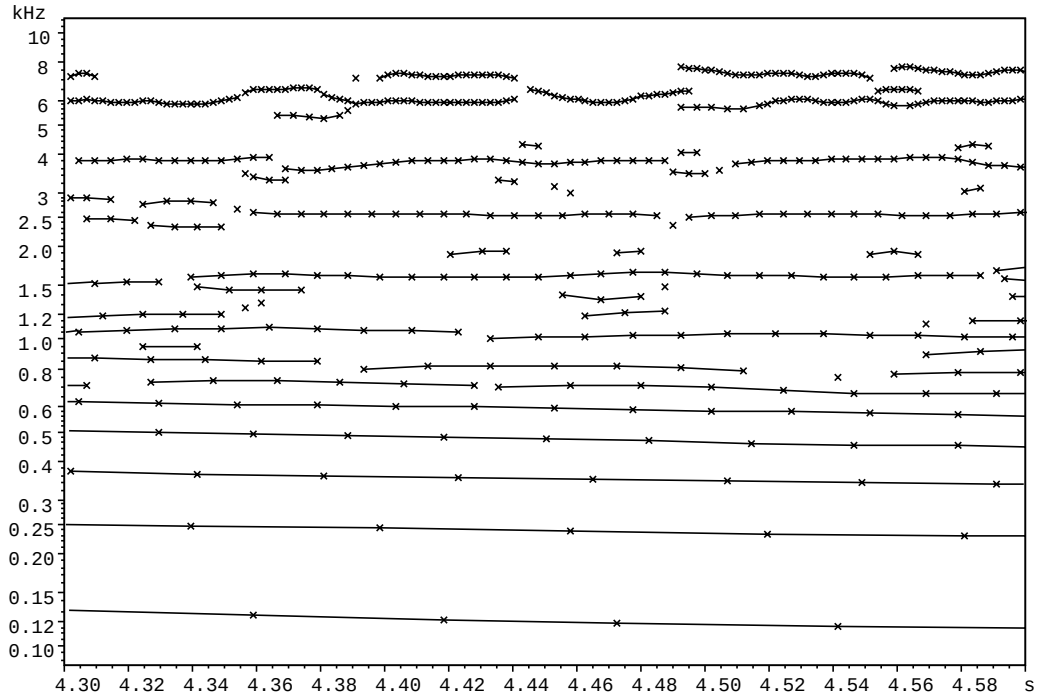



*figure 5.2 - Spectra from eight adjacent time clicks with the peaks marked, showing how tracks are formed by connecting similar peaks in successive frames. This diagram shows peaks for each track in every time frame; in practice, the peaks recorded for the lower frequency tracks will be many fewer since most will be deleted by the process described above.*

*figure 5.3 - Diagram showing several tracks at different frequencies for a fragment of a complex sound. Each track has a variable sampling interval depending on its instantaneous frequency, although there are no clear instances of variable sampling within a track; they are difficult to see on this scale. The sample points are visible as small 'x's on the tracks, and can be seen to be less frequent at lower frequencies.*



# 6 Track Resynthesis

## 6.0 INTRODUCTION

We have stated that this representation was to be invertible, and that even after modification we hoped to be able to work back to sounds that bore some significant perceptual relationship to the original source. A key component in achieving this is the method of converting a track-domain representation into a sound. This chapter discusses the choices available for this inversion, and the algorithm that we used.

## 6.1 ITERATIVE INVERSION

We have described an analysis method for transforming a sound into a time-frequency track representation that is highly nonlinear and deliberately lossy i.e. incapable of perfect inversion. Although we cannot expect to be able to work back to the exact input sound, it is still meaningful to try and reproduce the sound 'most likely' to have resulted in the representation we see. Further, since the information lost in our representation was specifically intended to correspond only to inaudible features, we should be able to invert what remains to a perceptually equivalent sound.

Given that there is no single 'correct' inversion, we must specify what exactly we are looking for in our inverted signal. The obvious choice is to produce a signal whose re-analysis matches the representation we are inverting. In general, there may be many signals that meet this requirement, but if our representation is perceptually adequate, they will all give the same acoustic impression, so any one is appropriate.

One common strategy when trying to work backwards through nonlinear systems is iterative approximation. If it is possible to get a sense of the error between two signals from the differences in their representations, then an estimate of the inverted signal can be improved successively until it is arbitrarily close to an optimal inversion. (An optimal inversion is one whose representation matches the target, or is maximally close in the case where the target representation does not correspond to a realizable signal). This is a minimum-assumption system : all that is required is

some way of converting an error in the analyzed domain into a modification to the signal domain. The better this loop is, the faster the convergence.

## 6.2 DIRECT INVERSION

However, if we are prepared to put stronger interpretations on the elements of a representation, we may postulate a closed-form inversion i.e. one completed in a single pass without going to the computational expense of re-calculating its representation. In any case, some first guess of this kind is required as a starting point for an iterative scheme. The representation we have is a set of time-frequency contours, so if we make the simplifying assumption that they have been independently and accurately measured by our analysis, we could form a resynthesis by superposing sinusoids generated according to each track.

This is consistent with what we have claimed for the representation - that it has identified regions of time-frequency energy concentration. Certainly by superposing sinusoids, we are putting energy 'under' each of our contours. The result is only an approximation because the peak energies do not, in fact, superpose independently - instead, adding two signals together will, in general, shift the energy peaks of both components. However, if the sinusoids are well separated in frequency compared to the relevant analysis filter bandwidths, the interaction is slight.

These arguments for the near independence of different frequency peaks are implicit in our treatment of the tracks as equivalent to individual time-frequency energy phenomena. To understand better how this independence is violated, consider a signal consisting of two sinusoids. Let us assume that in a particular analysis frame, they are of similar phase (since they are at different frequencies, their relative phase is constantly changing, and in a later frame they may have precisely opposite phase). In this case, the 'tails' of the envelopes of the two peaks will tend to bring both peaks closer together in frequency i.e. away from the actual frequency of the underlying sinusoids. When these tracks are resynthesized at the frequencies measured by peak-picking, the spectra peaks will again interact, so that peaks picked from the reanalyzed frame of resynthesis will be even closer together. If the original phases had been closer to opposite, the effect would be reversed.

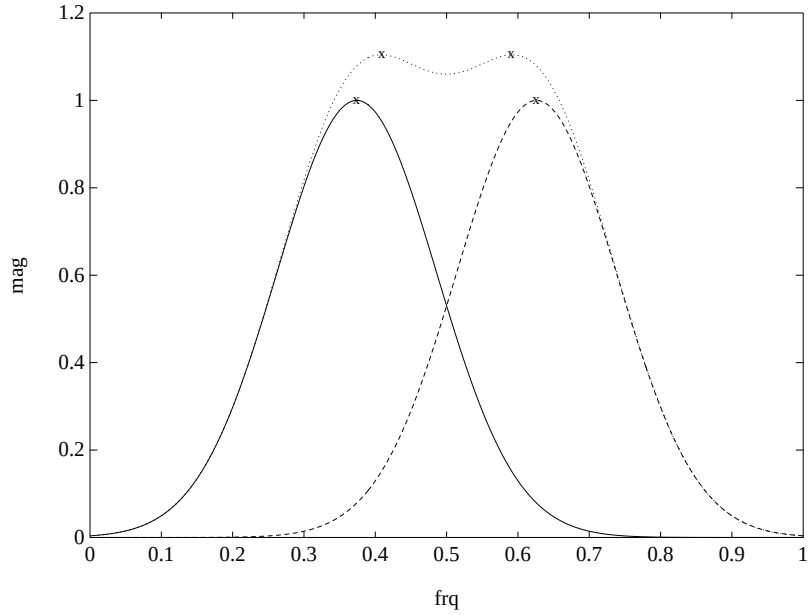



*figure 6.1 - Superimposing two peaks may affect the location of their peak amplitudes, shown by the x's above. In this case, the complex peaks are aligned in phase.*

It should be possible to compensate for these effects by consideration of possible interaction between peaks. There are two places where such compensation could be introduced. We could run our sinewave oscillators at frequencies slightly different from those of the contours so that the uncompensated reanalysis would match. But then we would be acknowledging that the track parameters were only approximately correct. If we want the values associated with the tracks to have a precise meaning, it would be better to compensate the frequencies *during the analysis*, such that an uncompensated resynthesis gives us what we want. This keeps the interpretation of our new representation simple.

There are two more points to be made in this regard. Firstly, we are not primarily concerned with a representation that can be simply inverted, but one that genuinely reflects perceived composition of a sound. If we believed that the kind of interaction that made our frequency peaks drift actually occurred in audition, we should *not* modify the frequencies in analysis, since we might expect that these distorted frequencies would accurately reflect the tones as perceived : to resynthesize a single tone from a mixture closest to how a listener would expect it to sound alone, we should use the distorted rather than the true frequency. Such a perceptual effect, where pitch of one tone is systematically modified by a near neighbor, has not been

observed, however. In any case, it would be a rapid modulation about the true pitch which would be difficult to measure.

The second point is that this entire discussion about the fine accuracy of frequency values is in many ways made irrelevant by the use of phase measurements in resynthesis, since the phase values have a very strong influence on the net frequency of the resynthesized tone, and in most cases serve to correct any 'noise' in the frequency estimates, as we shall argue below. Using phase has its own difficulties, discussed in the next section.

## 6.3 CONSIDERATIONS OF PHASE

Using the measured phase for reconstruction is a very powerful idea inherited from the McAulay-Quatieri Sinusoid Transform. The basic idea is this: given our samples of frequency and magnitude, how do we interpolate to generate the precise control parameters at each sample for our resynthesizing oscillator? Linear or 'sinc' interpolation is fine for magnitude, but for the frequency we realize that continuity of phase is crucial (to avoid artefacts) and thus we must integrate our frequency estimate (phase derivative). Implementing this integration discretely (as a phase advance for each sample) is numerically unwise, so we look for a closed-form solution to calculate the phase at arbitrary points between the two samples. We have frequency estimates at both ends, which correspond to boundary conditions on the slope of our phase curve. We could just fit a second-order polynomial and accumulate the arbitrary phase, but since we are dealing with the phase, we can get a more accurate resynthesis by storing phase samples during analysis and using them in our resynthesis, giving start and end values for our phase curve. This removes the need for an arbitrary phase origin, and gives one more parameter to match, providing a cubic function to generate our phase.

It turns out that using the measured phase in resynthesis greatly improves quality. Tones tend to be far steadier, and timbre seems less distorted. However, one could question the validity of suddenly invoking this expedient measure to make our resynthesis better. Everything so far has been perceptually oriented, whereas psychological experiments repeatedly demonstrate that relative phase of resolved partials has only the smallest impact on perceived quality [Patt87]; we have not proposed using phase in any other part of the processing chain, but suddenly we

bring it in to 'clean up' the resynthesis. We have an apparent contradiction between a very noticeable influence of using phase and its supposed insignificance.

The explanation for the beneficial effect of using phase is probably because it has a corrective influence on frequency estimates. Whereas the interpolated frequency value of the spectral peak suffers from several sources of error (such as the appropriateness of the parabolic fit), the phase is likely to be approximately constant in the vicinity of the peak for our zero-phase analysis windows. Thus the measured phase will not be sensitive to errors in the frequency estimate if it is varying little with frequency over that range. Both the frequency and phase measurements are contributing to the resultant frequency of the sinusoid i.e. the slope of the cubic phase function. Although fitting a polynomial constrains the instantaneous frequency at the sample instants to match the frequency samples, the phase samples define the average frequency over the interval (i.e. the exact fractional number of cycles completed between the sampling instants), and this longer-term average frequency is more likely to contribute to the pitch perception.

Measured phase is a more stable basis for the frequency also because noise and errors will not cumulate across frames; by contrast, phase errors derived from noisy frequency estimates will be integrated and go off on a ‘random walk’.

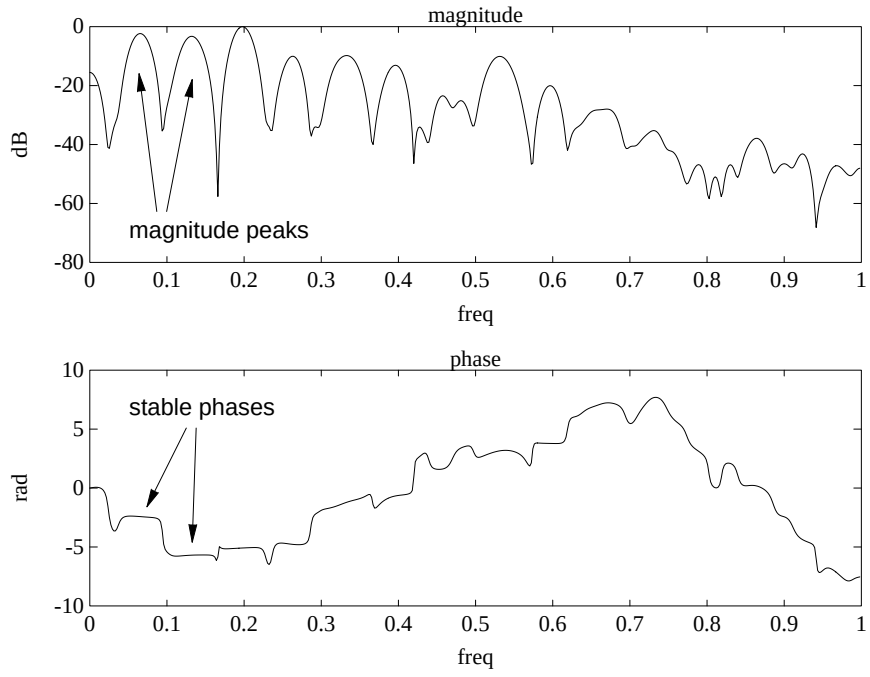



*figure 6.2 - The phase of the Fourier transform is often stable near a magnitude peak when a time-symmetric window is used in analysis*

We have presented some arguments to justify using phase as a more stable basis for the frequency of our resynthesized contours, we must consider why we have not used this estimate throughout. We are not suggesting that this phase is the basis for perceived frequency (although this is related to 'time' theories of pitch perception, and the 'synchrony' measures of [Cook91] and [Sene85]). Perhaps the differences are not so great as to make this a significant issue in track building or the kinds of correlation needed for source separation. However, we must acknowledge that this question of phase demands more serious consideration than it has received in this project.

## 6.4 PHASE EXTRACTION/INTERPOLATION

Given that we are using phase, it is important to remember that phase is not simply read off from a particular filter, but is, in general, interpolated along the frequency axis between spectral samples (which may themselves have been interpolated along the time axis, but in less confusing circumstances, so this need not worry us). This would be no more difficult than interpolating magnitude were it not for $2\pi$ ambiguity -- phase is only reported modulo $2\pi$ (this of course reflects a profound truth concerning the use of a continually increasing phase value to represent a cyclic function). When we have two adjacent samples of a spectrum, and we want to interpolate a phase value between them, we may choose to add an arbitrary number of $2\pi$'s to the difference between the two phases before interpolating - which will make a vast difference to our interpolated value. For instance, if we use linear interpolation the added-in phase difference will be weighted by the interpolation coefficient - i.e. it will no longer be a relatively innocuous multiple of $2\pi$.

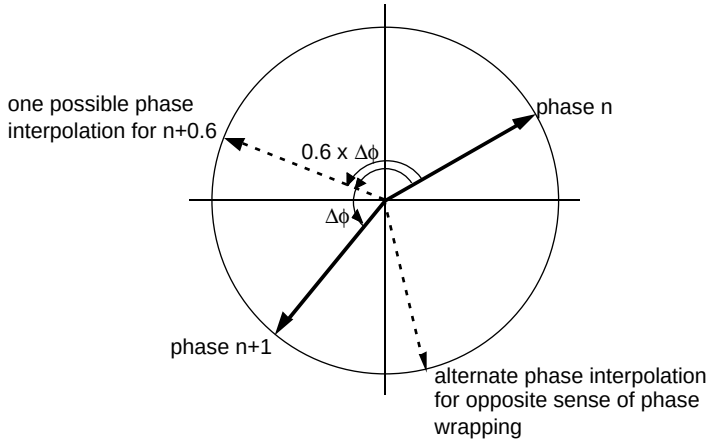



*figure 6.3 - Interpolating between two phase values can be ambiguous*

As we mentioned above, we do not expect large variation in phase between adjacent frequency samples around a peak, so normally the phase 'unwrapping' chooses the smallest angle between the phase samples to divide. However, when this distance approaches $\pi$, the unwrapping becomes ambiguous. Even with the (rather generous) frequency-domain sampling we were using, this ambiguity arose frequently enough to require special attention. If a phase difference was wrapped the 'wrong way', there would be one seriously incorrect phase value in the contour, which would result in a severe local distortion in the waveform, audible as a broadband 'click'.

To deal with this we could have implemented some rogue-suppression algorithm that only used a phase sample if it was consistent with other points and the frequency estimates. This would have problems in the situation where the very first sample on a contour had a rogue phase, since there would be no context against which to reject it. Instead, we made the assumption that the ambiguities are always at worst the result of one additional $2\pi$, and record two possible phase interpolations for each sample, one 'going round' each way. Then on resynthesis, to find the best path through these pairs of phase samples, we use dynamic programming and a cost metric based on frequency-derived expected phase increment to find the best set of phases to use. This always gives clean sounding results, and presumably finds the 'correct' answer if it is present.

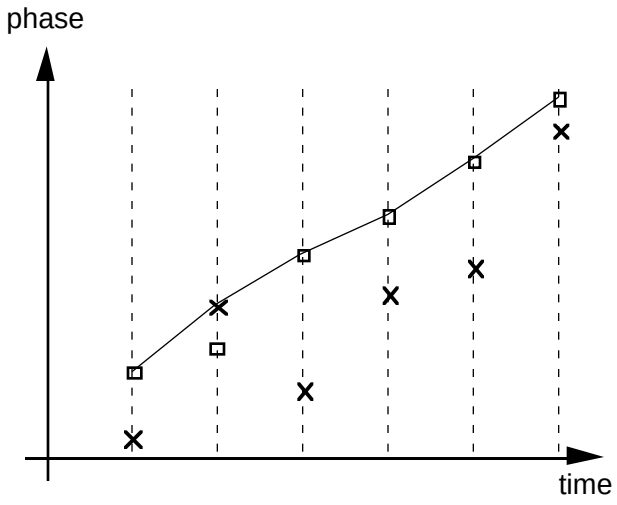
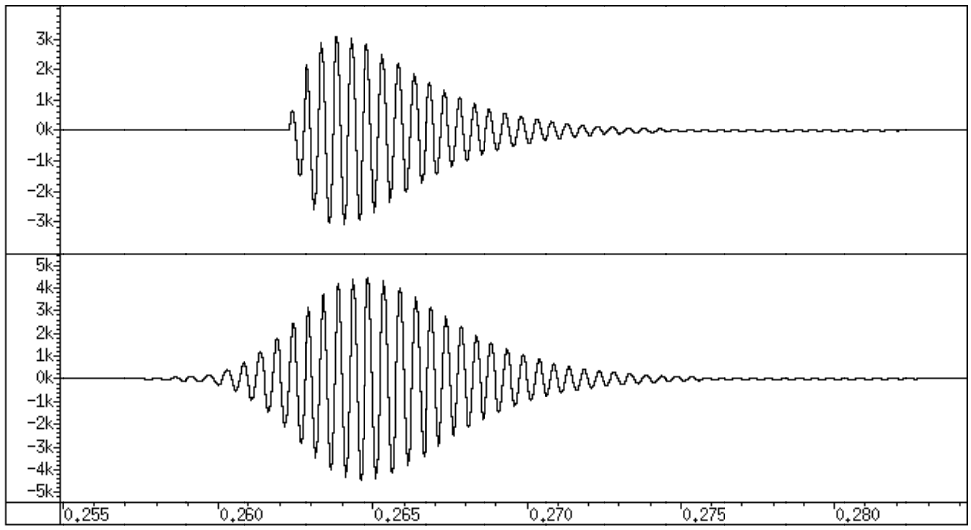



*figure 6.4 - Dynamic programming finds the 'best' path through pairs of alternate phase samples based on minimizing the deviation from the expected slope (frequency).*

## 6.5 TRACK ONSETS AND OFFSETS

One remaining problem in resynthesis occurs at the beginnings and ends of tracks. Since tracks have a finite extent, there is a first data point corresponding to when a particular peak was first resolved; when resynthesizing, we could just start our sinusoid at the frequency and amplitude held of that first point. However, in many cases this would cause an audible click, since the track may start at quite a large magnitude. It is preferable to 'grow' the new sinusoid from zero over some reasonable time interval. This was implemented linearly ramping the tracks from zero amplitude at their initial frequency over a fixed number of cycles -- i.e. a time that varied according to the frequency. Four cycles was used as being of the order of the sampling interval used for any subband (since the subband signals had an effective bandwidth up to a quarter of their center frequencies). Linear ramping was used rather than the more perceptually 'natural' logarithmic curve to avoid having to decide at what fraction of the final amplitude to start the fade (zero is not an option). A typical resulting sinusoid based on a very short track of three sample points is shown below:-

*figure 6.5 - Comparison between a short tone burst and its resynthesis from a single track, showing the phase matching and onset run-in. The resynthesis has added more than four cycles in front of the sound since the time uncertainty in the filter initiates the track marginally ahead of the actual onset.*



# 7 Results and Assessment

## 7.0 INTRODUCTION

So far we have described the analysis and resynthesis of converting between sound and the new representation; we have not looked too closely at how all this translates into practice. In this chapter, we consider the performance of the current manifestation of this system over a variety of sounds and by various criteria. We consider the qualitative nature of the tracks of the representation, and how well the reconstructed sounds match our intentions.

## 7.1 NATURE OF THE ANALYSIS OF SOME SOUNDS

What is the qualitative nature of the tracks which are the output of the processing so far described? Looking at a graphical representation of the tracks compared to the original sound at least begins to answer this question. In the following sections, our main visualization will be via a constant-Q spectrogram ('scaleogram') of the analyzed sound; this is the magnitude output of our filter bank, plotted with frequency vertical, time horizontal and value as black/white intensity (the dynamic range covered by black to white is about 60 dB and is logarithmic). The tracks representing the sound are overlaid as time-frequency paths. Although this does not convey directly the magnitude contour, the magnitude of a particular track is given by the intensity of the grayscale directly underneath. For each of the spectrograms, the time-domain waveform is shown at the top, on the same time scale.

### Sine tones

The first diagram, figure 7.1, shows the analysis of a simple sound file consisting of two stable, gated sine tones - first presented separately, then simultaneously. At 200 Hz and 2 kHz the sine tones are well separated in frequency, and are thus treated essentially independently. Looking at the grayscale, we see the expected dark horizontal bars around 200 Hz between 0.1 and 0.3 seconds, around 2000 Hz between 0.4 and 0.6 seconds, and at both places between 0.7 and 0.9 seconds. We also see tracks running down the centers of all these bands; these are the tracks representing the principle perceptual features of this sound.

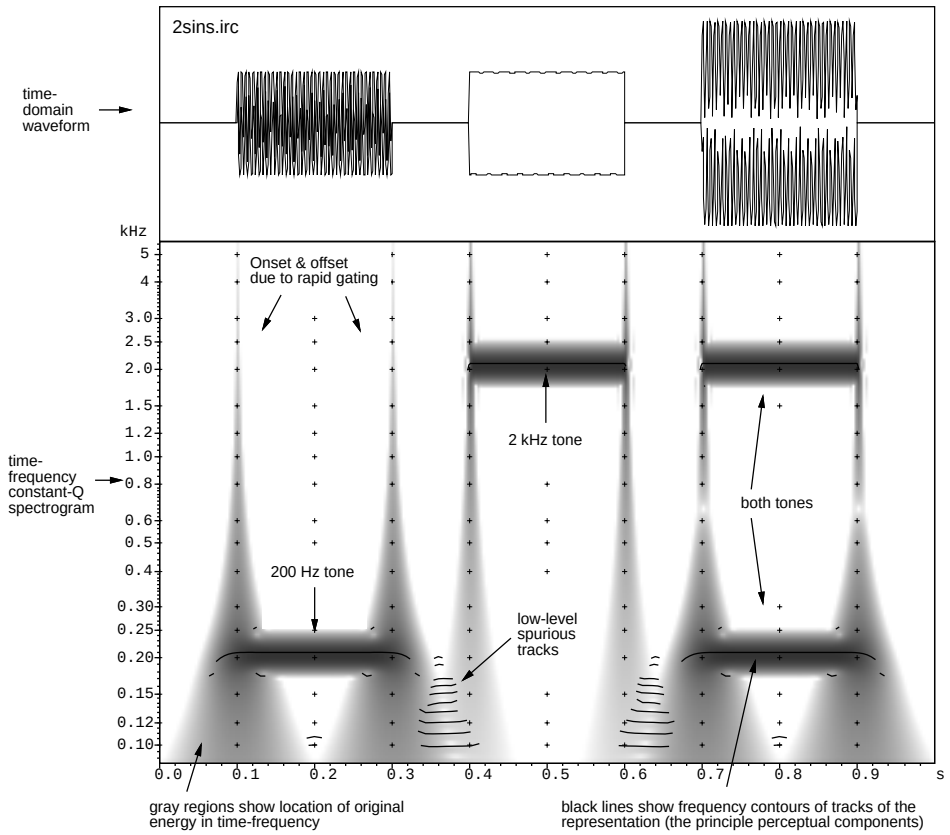



*figure 7.1 - Analysis of 200 Hz and 2 kHz gated sine tones*

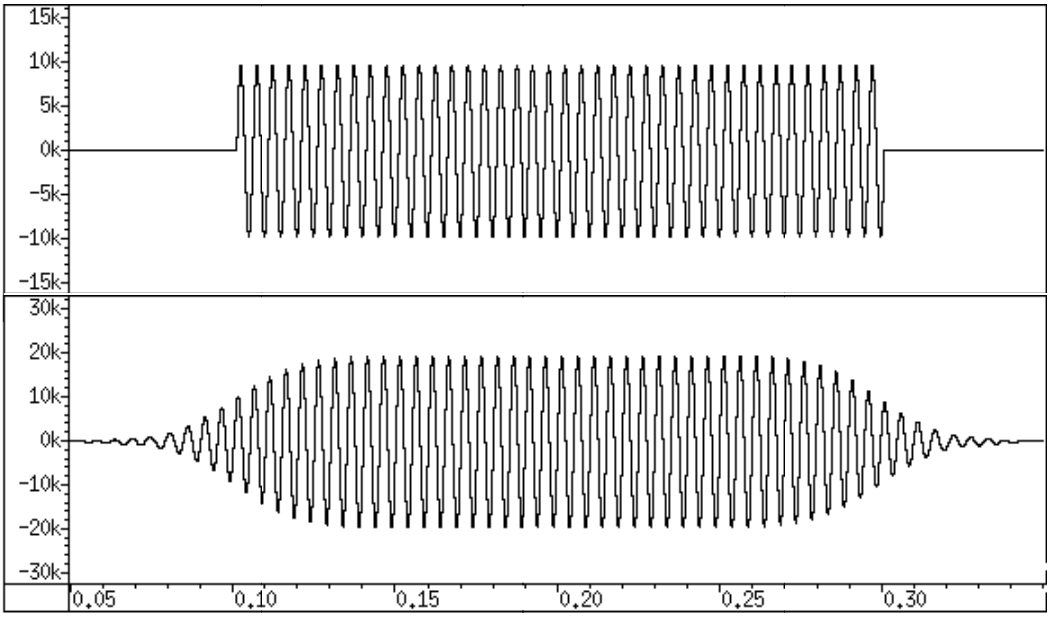



*figure 7.2 - Time waveforms of 200 Hz tone before and after processing*

The next most prominent features of the figure are the large vertical pointed spire-like shapes in the grayscale, aligned with the starts and ends of each tone burst. These are the spectra of the onset/offset transients[1] for the tones, which occur because the tones were started and stopped abruptly rather than being ramped up and down in amplitude. This abrupt change does indeed cause a broadband transient perceived as a click. But here we run into a weakness of the system, which, fortunately, only tends to arise with artificial signals such as this one.[2] Because the spectrum of the click is perfectly smooth, as only a synthetic signal could be, and because our tracking algorithm is constrained to represent regions of energy by local peak values, it cannot reflect the considerable high-frequency energy at 0.1 seconds because the spectrum simply has no peaks above 200 Hz at that instant, although there is considerable energy. In any 'real' sound we would expect natural coloration to introduce several peaks over this energy making it easier to record, but this exposes a weakness of the peak-picking strategy: it assumes that any peak will be sufficiently concentrated so that its slopes are essentially masked. The occasional smooth, broad peak violates this assumption. There are strategies to overcome this (breaking up the spectrum before peak-picking) but the problem is not common enough to warrant their implementation.

In fact, the only visible tracks apart from those directly associated with the sine tones are low-energy artefacts from interference of the skirts, essentially inaudible on resynthesis. Arguably, they should have been suppressed by a simple threshold mechanism to hide low-energy tracks.

The remaining comment to make on this diagram concerns the run-in and run-out regions of the 200 Hz track. The first sine starts at exactly 0.1 seconds, but the track starts almost 0.04 seconds before this and curves down to perhaps 20 Hz below its main frequency. The anticipation of the track starting arises because the filter bank

---

[1] It is useful to try and define transient. In general, the transient response of a system is the nonrepeating features of the output after a change in input, before steady state is regained. Here it is used more broadly to refer to a signal that is compact in the time domain, and hence distributed in the frequency domain. Time domain compactness is gauged relative to some characteristic time of a system. Thus what would be a transient to the lowest octave of our filterbank might be smooth and steady when viewed by the faster top octave.

[2] These test signals were generated using Barry Vercoe's *Csound* [Verc90].

has been made 'zero phase' - that is, each filter has an impulse response symmetric about time = 0, and thus each channel can 'see' a short time into the future. The slight frequency modulation occurs along with the amplitude build up as the track fades in (the modulation is from below because the lower filter has a longer inpulse response ; since this fade in is less than 8 cycles long, there is no perceptible detuning when the track is resynthesized. Instead, the resynthesis sounds very much like the original sine tone, but with the onset and offset transients removed : the tone is nicely tapered at each end. Figure 7.2 shows the time domain waveform of the resynthesized 200 Hz tone.

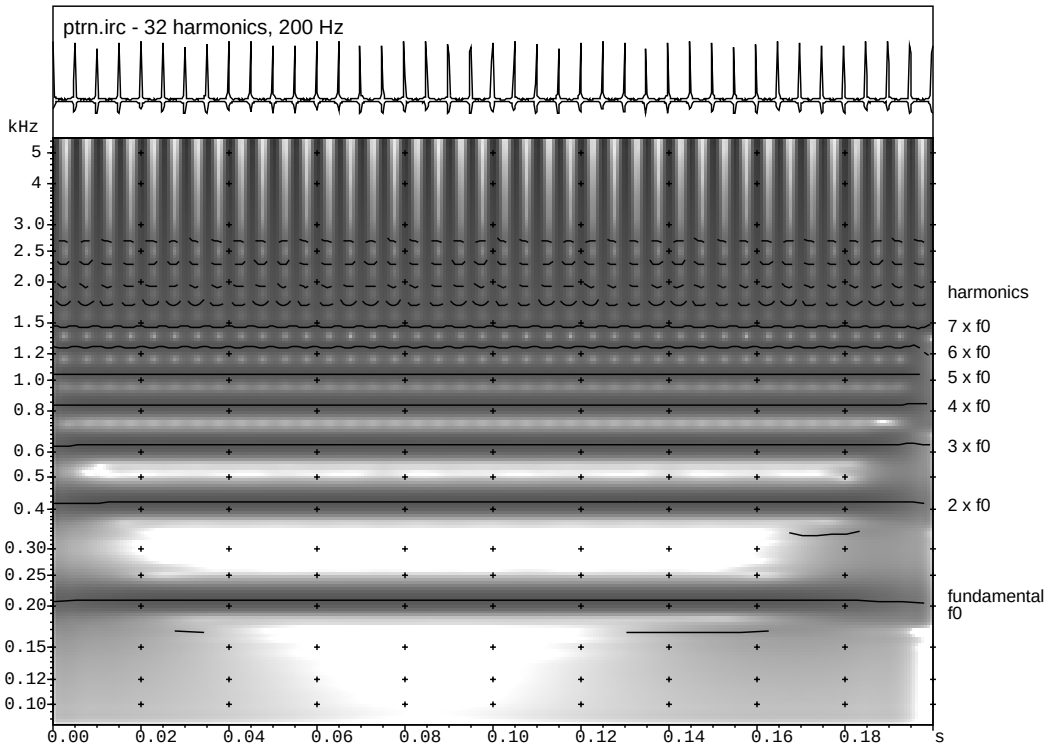



*figure 7.3 - Analysis of pulse train*

**Pulse train**

The next figure, 7.3, shows 0.2 seconds of a 200 Hz bandlimited impulse train formed by adding together 32 cosine-aligned harmonics. This sounds very buzzy and has a very flat spectrum. The constant-Q spectrogram shows how the narrowband, long-time, low-frequency filters resolve individual harmonics, while the faster, high-frequency filters show broad pulses at each impulse. This behavior is reflected in the tracks: the bottom seven harmonics are clearly represented as separate tracks, but above this we only see track fragments modulated at the pitch cycle. Interestingly, these occur *in between* the pitch pulses and are of relatively low energy. The broadband pitch pulse transients suffer from the same problem as the transients from the sine gating : their synthetic origin makes the spectra completely smooth, so that the tracking cannot find a single peak at which to sample and record the energy during the peak. As a consequence, the resynthesis from these tracks sounds very much duller than the original - it is as if the original signal had been low-pass filtered at 1.5 kHz. This is apparently in contradiction with the assertion that a constant-Q representation improves short-time resolution, but it is really a separate issue relating to the shortfalls of spectral peak picking. Again, any *real* sound of this nature would almost certainly have some peaks in the high frequency for the peak-picker to work with.

The other observation to make about this figure is low frequency energy spreading in from the edges at the low frequency. This is the 'edge effect' resulting from the filters effectively seeing silence surrounding the sound we see. The extent of this effect is strictly limited to within 0.1 seconds of the ends, and it is does not present practical difficulties.

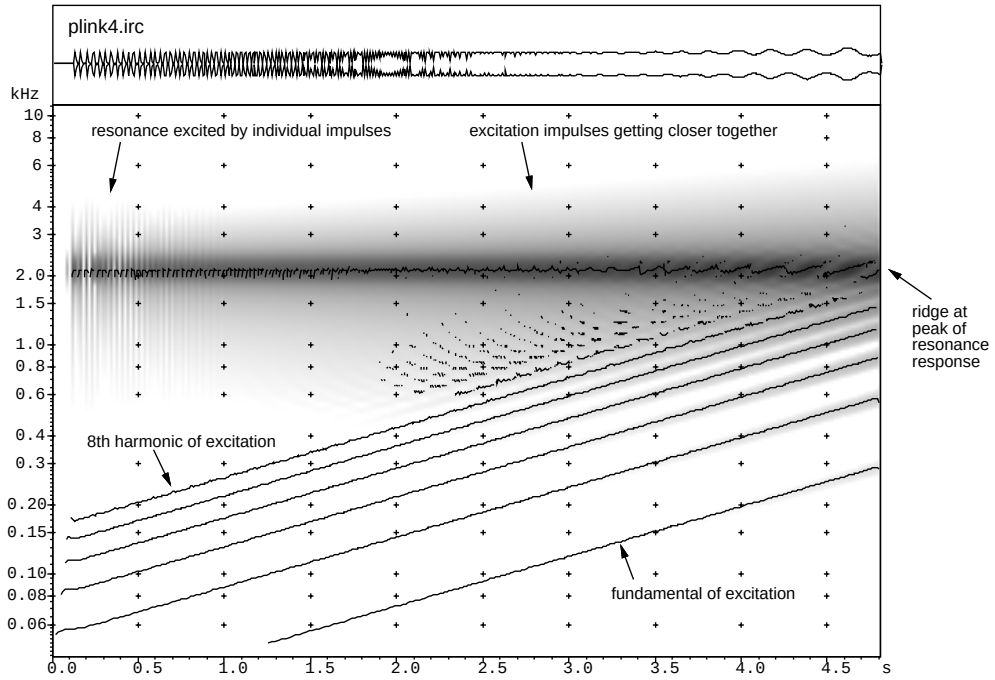



*figure 7.4 - Analysis of accelerating resonated impulse train*

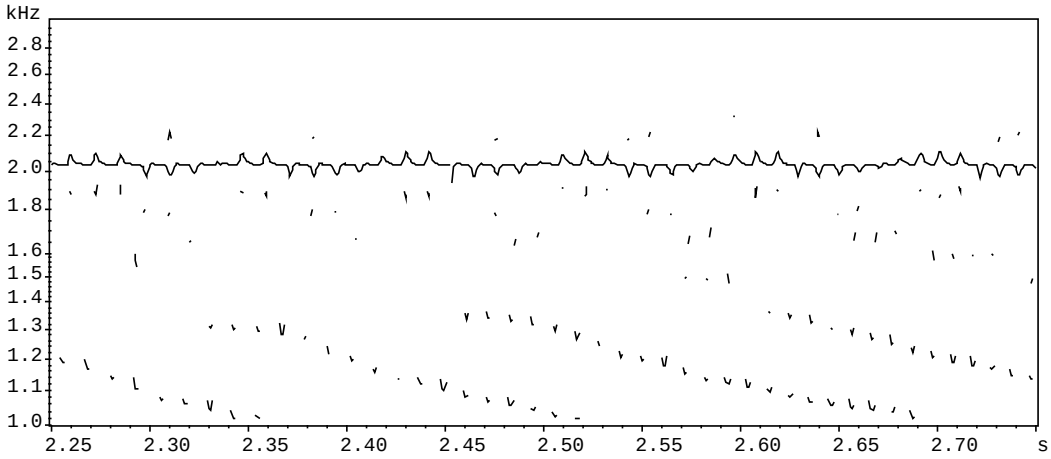



*figure 7.5 - Detail of resonance tracks showing variation around center of resonance*

**Resonant signals**

The third synthetic test tone shown in figure 7.4 is concerned with showing how the system treats a periodic signal with a strong resonant peak in its spectrum that does not necessarily lie directly over one of its harmonics (the formant peaks in voiced speech fall into this category). The signal labelled "plink4.irc" is composed of an accelerating impulse train that starts at a 30 Hz repetition rate and accelerates to 300 Hz over 4.7 seconds. These impulses are then passed through a sharp resonant filter centered at 2 kHz with a Q of around 10 (measured at the -3 dB points). As the impulse train varies continuously, the resonance will lie alternately over and between the successive harmonics.

At the left of the figure, the individual 2 kHz 'bursts' excited by the widely-spaced 30 Hz impulses show up as clearly distinct regions of gray in the spectrogram, summarized by a short, rapidly-modulated track centered on the peak energy at 2 kHz. These tracks remain distinct for at least the first two seconds, b y which time the resolution of the figure means we can no longer see gaps between the gray bursts or their overlaid tracks. At the same time, we see low frequency harmonics resolved, initially at 180 Hz and below; at around 2.0 seconds, we see other local maxima in energy between 600 Hz and 1.5 kHz. As the repetition rate accelerates, the fundamental appears on the graph at 1.25 seconds. (Note that the 50 Hz lower limit of the frequency scale is arbitrarily chosen as eight octaves below the highest filter band, and we could have carried on adding lower octaves as long as we liked). This identification of the lower harmonics is exactly as we expect from the pulse train above, but we are really interested in what is happening to the tracks in the vicinity of the resonant peak at 2 kHz. The scale is too compressed to make it clear, but essentially the result is that while the resonance lies far enough above the fundamental to avoid the region of separate harmonic resolution (say 3 or more octaves above), the track that is fit to the resonant peak will tend to be pulled towards the nearest multiple of the fundamental during the pitch pulse, but will relax back to lie on the true center frequency of the resonance during the rest of the cycle. This means that under these conditions, the track is actually spending a significant portion of its time conveying information on the *formant* location rather than the location of particular harmonics. Figure 7.5 shows a short portion of this time-frequency graph zoomed in to show this behavior.

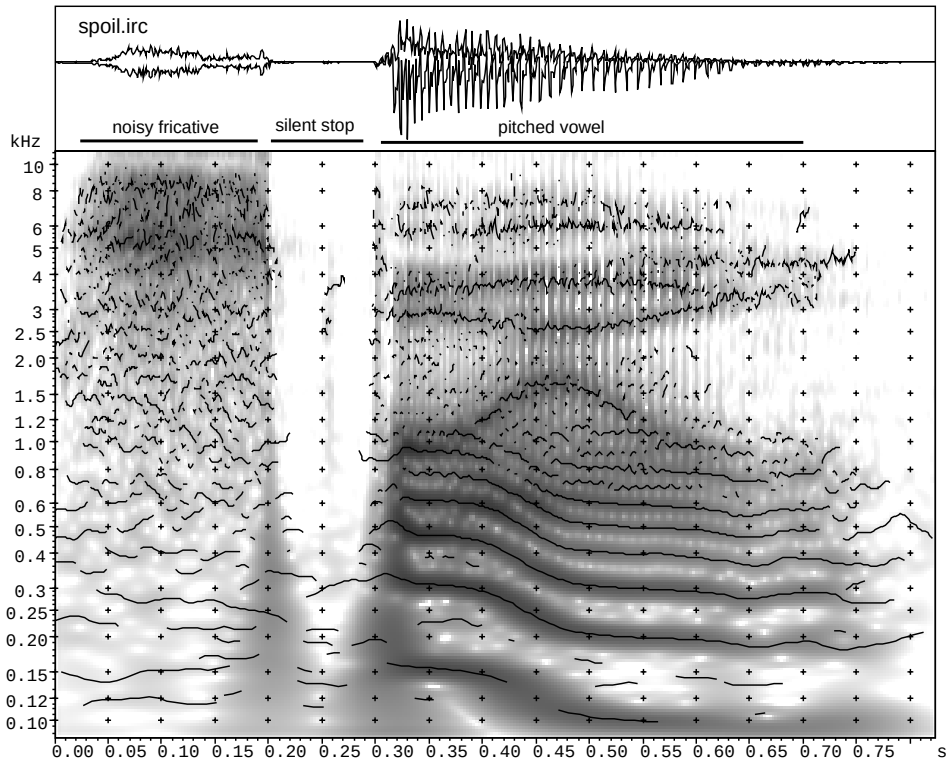



*figure 7.6 - Analysis of spoken word, "spoil"*

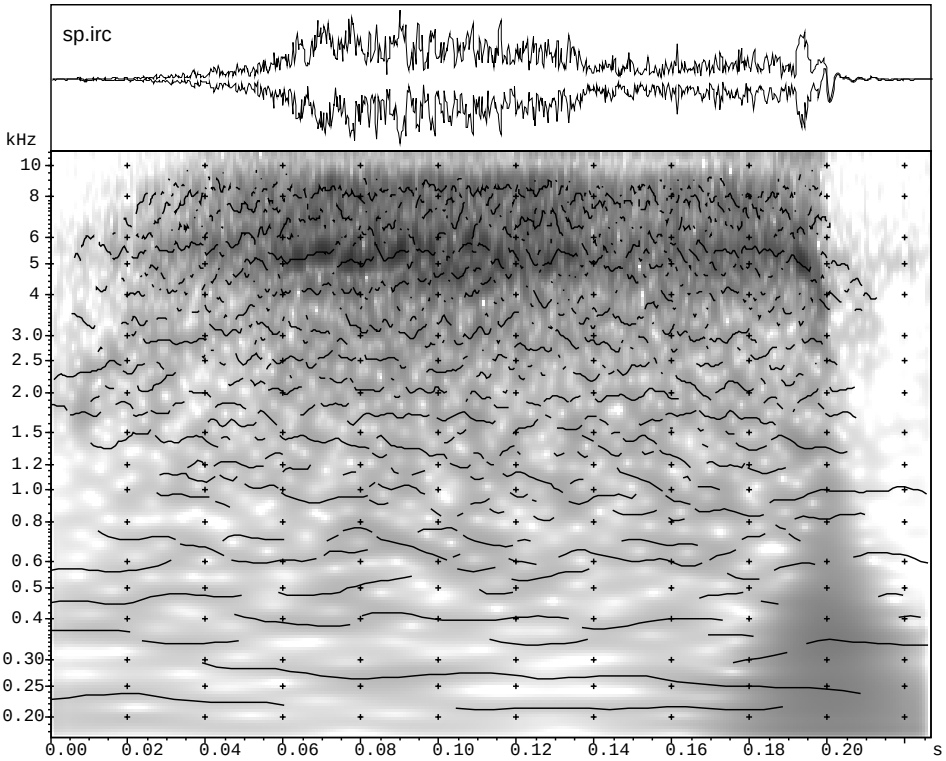



*figure 7.7 - Analysis of initial /s/*

**Voiced and unvoiced speech**

Figure 7.6 shows a portion of speech, the word “spoil”. The three segments of this sound are clearly visible : initially there is the unvoiced /s/ fricative (0.0 to 0.2 seconds), then a period of silence corresponding to the /p/ stop, then the moving formants of the pseudo-periodic voiced section (0.3 seconds onwards).

Looking first at the voiced portion, we see that the regions of greatest energy are covered by tracks picking up the first five or six harmonics of the voice pitch as it falls between approximately 150 and 100 Hz. Above these harmonics, we see tracks which are essentially picking out the formants of the vowels, which quite obviously are moving independently of the voice pitch. There is a short region of F2 around 1.5 kHz from 0.35 to 0.55 seconds. F3 and F4 around 2.8 kHz and 3.5 kHz respectively are also quite clear. Naturally, all these formant tracks are heavily modulated at the pitch rate -- most of them show short breaks between pitch pulses. Formants which fall onto the resolved harmonics do not have their own tracks, of course. We might expect less sensitivity to their location, since our auditory model is suggesting they are only perceived as emphasis on particular partials rather than as a partial-independent peak. This explicit coding of fundamental pitch and formant location in the track frequencies is encouraging : these are the features that seem to be most crucial to speech communication, so we would not be surprised to find them prominently featured in the brain’s representation of sound at some level.

The initial /s/ of this sound is essentially a noise signal, and gives us an opportunity to see how our model deals with this poorly-modelled case. There is a profusion of tracks, mostly of short duration (relative to the timescale of their frequency) showing little correlation in onset, offset or frequency modulation. The tracks in the high frequency appear to become increasingly jagged, but this a result of their faster time scale. Figure 7.7 is a zoomed-in view of just the noise burst, showing that the tracks in noise at different frequencies share many general qualities (modulation depth, density, normalized duration) when displayed on different timescales. This figure has the timebase expanded by a factor of 4.

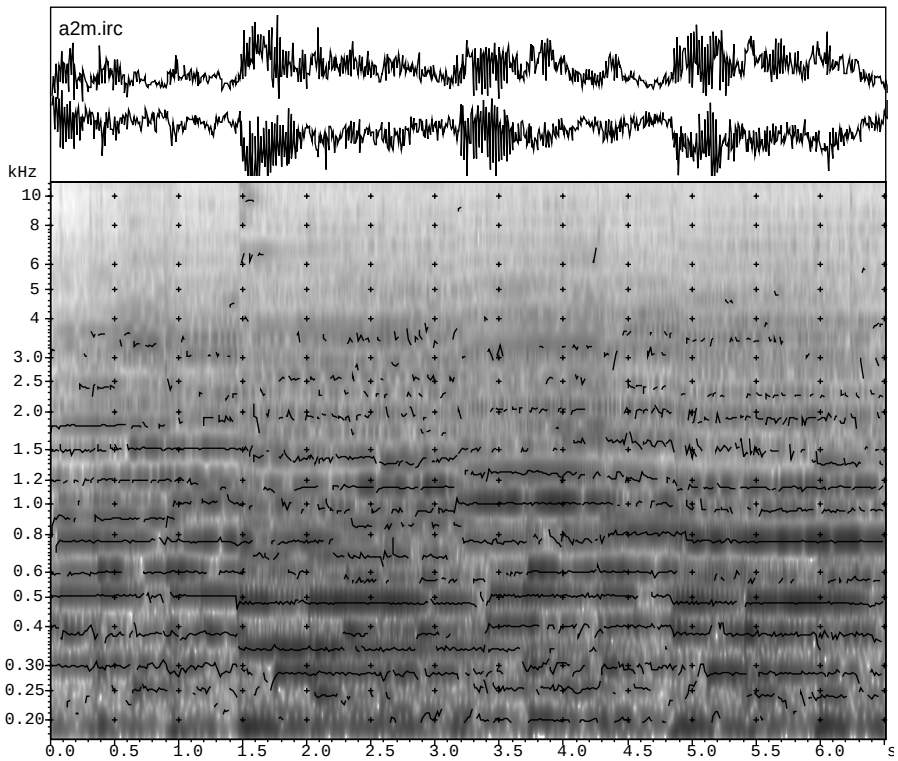



*figure 7.8 - Analysis of orchestral fragment (shorter high-frequency tracks have been suppressed for clarity).*

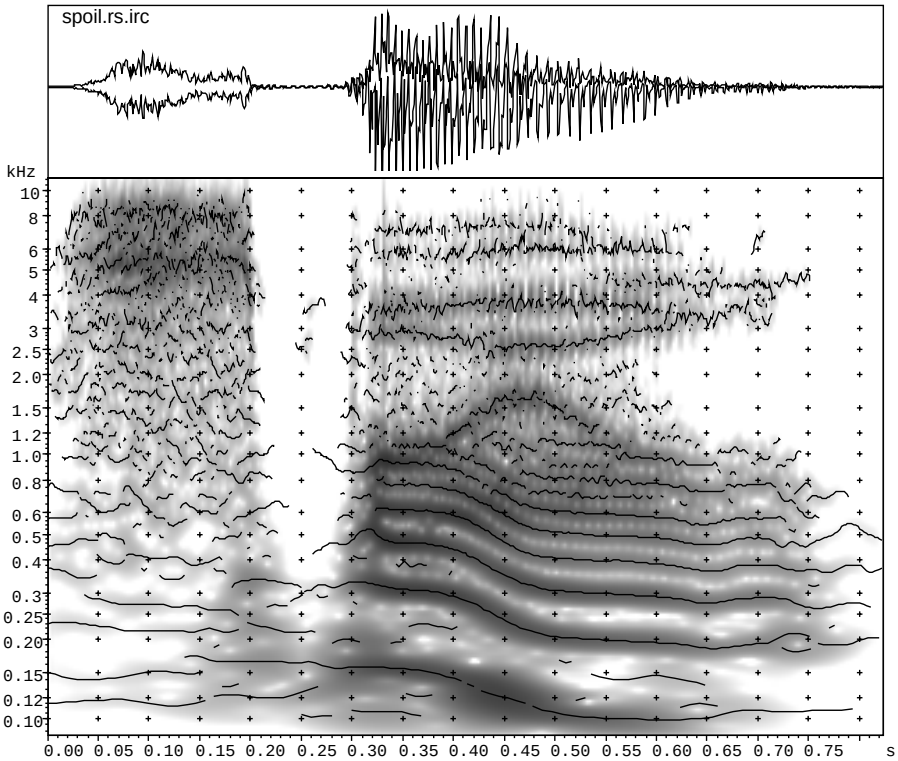



*figure 7.9 - Analysis of resynthesis of "spoil"*

**Complex sound**

The next figure, 7.8, is of a 6.5 second excerpt from an orchestral piece. The sound is composed of many sources, and the spectrum is correspondingly dense and convoluted. Although it is hard to interpret many of the tracks we see, the resynthesis they generate is in fact a good perceptual match to the original. Informal listening comparisons reveals a slight emphasis of high-frequency noise as the only noticeable artefact. The most conspicuous features of the diagram are the the long stable tracks below 2 kHz. Since this is medium-slow tonal music, these are frequencies important to the particular harmony at that moment. Some of them show rapid, shallow frequency modulation : this is most likely the result of interfering harmonics where several instruments are playing harmonically related notes.

Looking at a vertical slice through any given instant, there are only a relatively small number of tracks simultaneously active - in the region of three to four per octave, or 20 to 30 for the whole spectrum. This sounds like a big data reduction, but of course the bandwidth of the modulation information for the high frequency tracks is very high, so the amount of data implied by a single track can be very significant.

## 7.2 QUALITY OF RESYNTHESIS

The previous section focused on the analysis of sounds into tracks. We now consider the performance of the inverse procedure -- recreating sounds from sets of tracks. It is difficult to know how to assess these new soundfiles. In the case where we have not attempted to modify the sound in the track domain, we can compare the original and resynthesized sound files. But we have designed the whole system to preserve only the salient features of sound, and not to bother with any others, so we would expect deviation between input and output.

A sensible approach to assessing the resynthesis is to re-analyze the output and see if the tracks so formed match the tracks from which the output was constructed. This is a good definition of what we were trying to do with the resynthesis, and so is a sensible test. But as we discussed in the chapter on resynthesis, the very simple method used for resynthesis does not perform outstandingly well in this respect. We have tolerated this performance because on listening to these resyntheses, the perceptual match to the original sound is very good, much better than one might expect from the differences in analyses. This is consistent with a situation where our

representation is still too strict: while the tracks may embody all the information necessary to reproduce a perceptually identical sound, they may also contain extra perceptually irrelevant information, and if the differences between the original analysis and resynthesis-analysis lie along these excess dimensions, the distortion is of no significance.

Figure 7.9 is the reanalysis of the word “spoil”, generated from the tracks shown in figure 7.6. There is a clear maintenance of the general shape of the sound, although most tracks have been distorted to a greater or lesser extent. Regions of energy that failed to attract tracks in the original sound are now completely devoid of energy. There seems to be a tendency for tracks to get longer and join up -- the representation of diffuse energy by a single peak has ‘coagulated’ the energy around the tracks.

Informal listening comparisons were conducted. We have mentioned some specific examples in the previous section on analyses. The principle results are summarized below:-

- In most cases, resynthesis sounds nearly identical to the original signal, excepting the bandlimiting implicit in the filterbank we have used, which does not cover the whole spectrum.
- Speech in particular is extremely well reproduced, even by earlier less refined versions of the system. This leads us to suspect that some artefacts are better hidden in speech than in, say music.
- Some problems were encountered by the addition of a ‘washy’ background noise to music signals. Although these could normally be eliminated by careful adjustment of track parameters, their cause is not well understood, and may be due to problems with coding of nonperiodic signals.
- Resynthesis of subsets of tracks gave ‘sensible’ sounding results (for example, resynthesizing just the F2 and F3 formant tracks of speech retained intelligibility; extracting just the tracks corresponding to the fundamentals of notes in a piece of piano music converted it to a soft hammer glockenspiel rendition!). This is encouraging since we have been trying to position the tracks as perceptually significant objects.

# 8 Comparison with Sinusoid Transform

## 8.0 INTRODUCTION

As we have mentioned before, the current system draws many ideas from the McAulay-Quatieri Sinusoid Transform System (STS) [McAu86], the principle differences arising from the use of a constant-Q filter bank front end instead of the fixed-bandwidth FFT of the STS. In this chapter we will examine the qualitative differences between these two approaches. Since the STS has a fixed bin width, and hence a uniform time window length, this time window sets an upper limit on the time resolution available from the STS in certain circumstances. In the constant-Q sine wave model (CQSWM), the time window reduces with increasing frequency, so we will see that the principle advantage of our new system is in efficient and accurate encoding of rapid transients as distinct objects.

## 8.1 THE MCAULAY/QUATIERI SINUSOID TRANSFORM

One way of thinking about the STS is the following: when we look at narrowband spectrograms of pitched sounds we see what looks like a pattern of near-horizontal stripes (the individually resolved harmonics). Rather than encoding the whole spectrogram as a sampled two-dimensional surface (a conventional time-frequency decomposition), why not exploit this characterization by simply coding the center and intensity of the stripes at each sampling time? By exploiting continuity along the stripes, good quality resynthesis can be obtained. For sounds that do not have a strong harmonic structure (such as noise) it is not clear how the encoding scheme will work, but as it turns out the peak extraction algorithm generates a collection of short sine components that resynthesize to a perceptual equivalent to the original noise. Excellent data reductions for speech have been achieved through this representation.

Although this is an argument based mainly upon the *visual* characteristics of the narrow band spectrogram, it turns out to have excellent fidelity in the auditory realm. We may speculate that this arises from a good match between the strategy of concentrating on harmonics as entities and the processing in the human auditory system. So far, this discussion strongly resembles similar arguments for the CQSWM. After all, at the bottom of the spectrum the two systems are basically

equivalent. It is only in the high frequency where the CQSWM filters become much broader that differences become obvious.

## 8.2 CODING OF FORMANTS BY THE STS AND THE CQSWM

The STS has had great success with speech signals, but speech provides a good example to highlight the differences with the CQSWM. Our experience of voiced speech tells us that while sustained vocalizing does indeed have a prominent harmonic spectrum, the information in the upper part of the spectrum is carried by the formants i.e. broad magnitude characteristics of groups of harmonics, rather than precise amplitudes of individual harmonics. We also know that formants arise from excitation of damped resonant cavities in the vocal tract by the impulse-like glottal waveform, and when we consider the time-domain waveform of a vowel such as /ah/, we see that the glottal pulse shows up as a very abrupt transient once every pitch cycle followed by damped ringing from the formant resonances. This is exactly what we expect according to our all-pole, minimum-phase model of speech production. Now the STS encodes these periodic transients by calculating the strength of each of the many harmonics, as calculated from the Fourier transform of a windowed portion of the signal comprising 3 or so pitch cycles. Figure 8.1 illustrates this situation. In order to regenerate these time-compact formant transients, the synthesizer must generate each of several harmonics in the correct phase and amplitude relationships. Cancellation over the multiple-cycle window between these harmonics then regenerates the time-domain characteristics. But note how these features are indirectly coded as the net result of the phases and amplitudes of several tracks. In particular, the harmonics' phases, not normally considered perceptually significant, translate into fairly gross signal characteristics on the scale of the pitch period, perhaps 10 milliseconds. It is easier to believe that this 'phase' information is perceptually important.

By contrast, the CQSWM will typically pick up a formant with a correspondingly broad band-pass filter and encode it as a single track. Rather than having eight or ten adjacent harmonics with slowly changing phase and amplitude relations interfering to generate the pitch pulse transient, the transient can be recognized and resynthesized as a single, rapidly enveloped burst centered in frequency on the formant peak and in time at the pitch pulse. This representation seems the more 'direct' of the two equivalent forms when we look at the time domain waveform of a vowel with high formants, but we argue that it is preferable primarily because it

more closely matches our best understanding of what is happening in the actual auditory system.

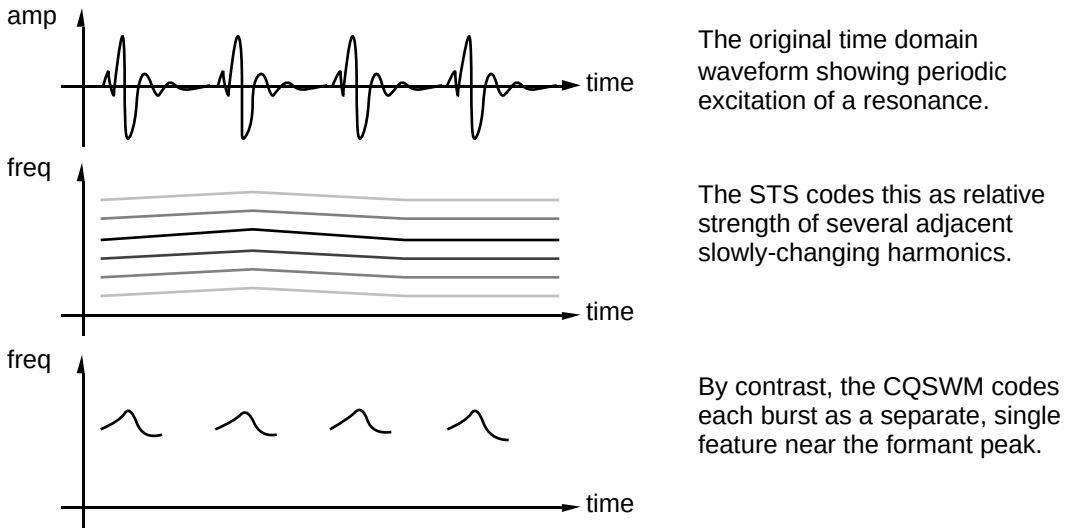



*figure 8.1 - Schematic comparison of treatment of formants*

## 8.3 COMPARISON OF TIME-DOMAIN WAVEFORMS

By way of an example, we can compare the performance of the two transforms on the synthetic-resonance signal used to examine the analysis (similar to that shown in figure 7.4). The figures below compare a short section of the time domain signal showing one of the short 2 kHz transients superimposed on the 300 Hz sine tone. It is from early in the signal, so the 2 kHz bursts are still well spaced in time, and only one is visible. We see that the transient starts very abruptly in the original signal, but in the resynthesis from the CQSWM the attack has been blurred by a couple of milliseconds, reflecting the time support of the filter at that frequency. However, the signal resynthesized from the STS with a 30 ms time window is blurred and delayed almost beyond recognition, and certainly with enormous perceptual modification. This is because the burst generated just one peak in one frame of the representation, but this includes a time uncertainty of tens of milliseconds. The STS is typically run with a window of this size to ensure recognizing the harmonicity of pitched sounds in the normal speech range. We can reduce this window down, as shown in the STS-6ms window figure, with a corresponding improvement in the reconstruction of the transient. However, at this stage we are on the verge of being unable to detect the 300 Hz sinewave as a stable harmonic; if the window becomes smaller, it shows up as a variable DC offset in successive frames with corresponding gross distortion.

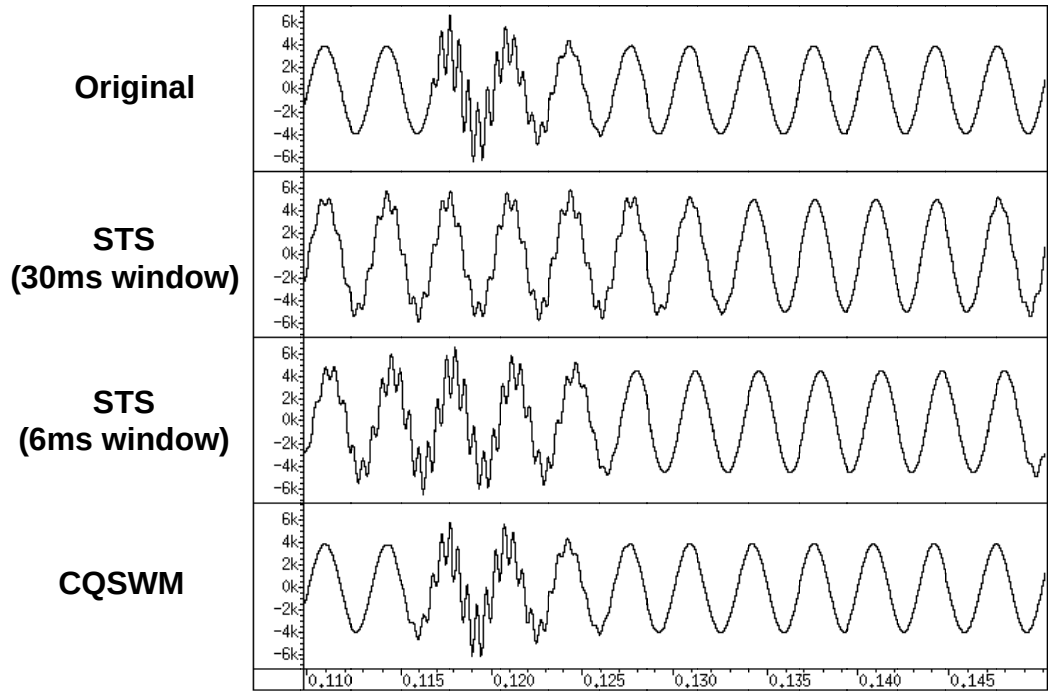



*figure 8.2 - Comparison of transient waveform with STS and CQSWM resynthesis*

Figures 8.3 and 8.4 show comparisons for two real sounds: a dropped can, and a stop release from a voice sample.

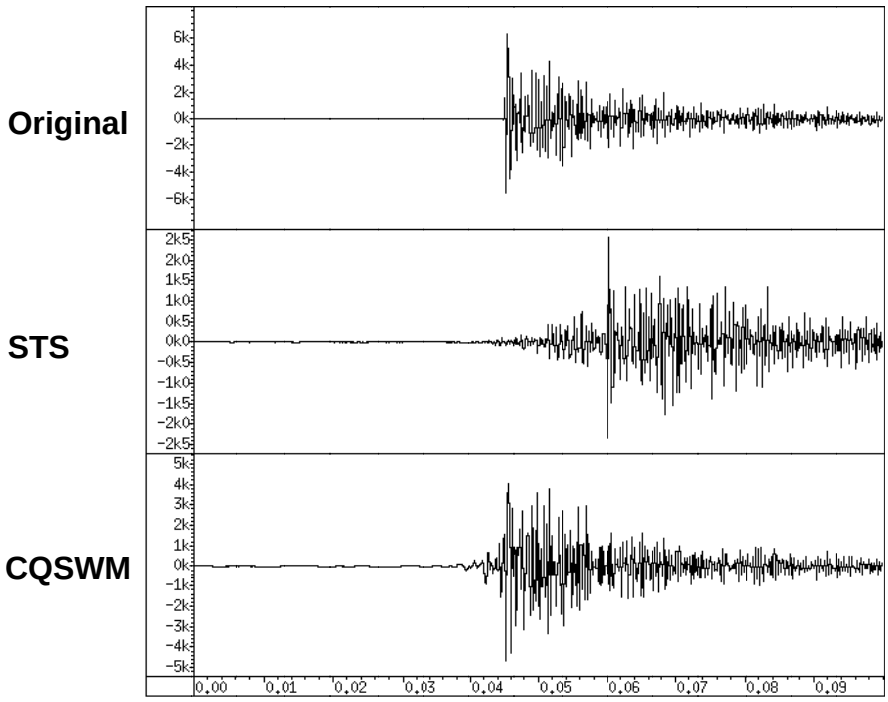



*figure 8.3 - 100 ms of the sound of a can hitting a hard surface. Original at top. The next waveform is the STS resynthesis, showing considerable pre-echo. The CQSWM resynthesis below shows less pre-echo, especially for the high frequencies.*

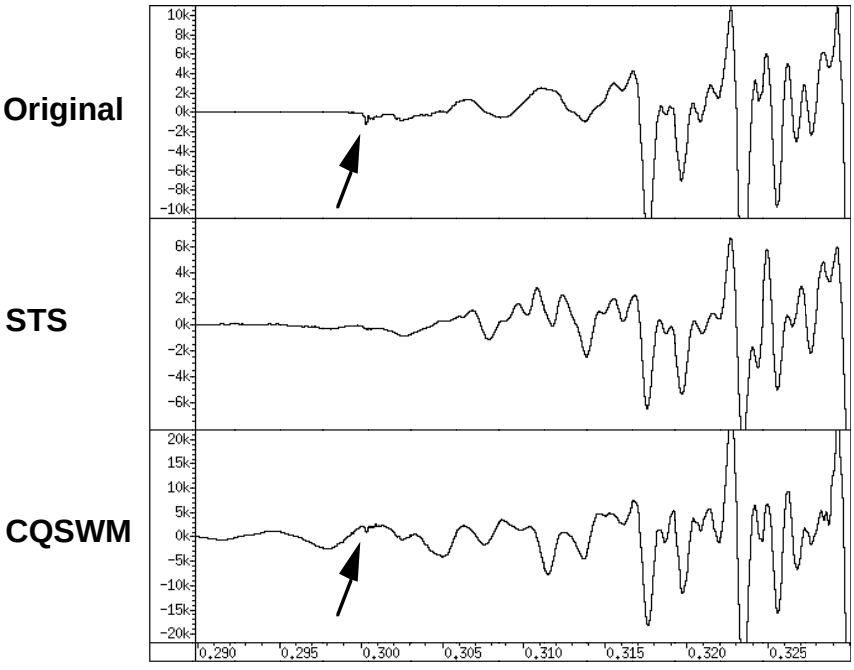



*figure 8.4 - A similar comparison for 40 ms around the stop release of "spoil". The release is a small spike at 0.300s in the original, indicated by an arrow. There is no corresponding spike visible in the STS, but it reappears in the CQSWM, again indicated. Note also that the CQSWM has worse pre-echo than the STS for the low frequency voicing fundamental, since its effective time window (about 160 ms for 100 Hz) is longer at these frequencies than the constant 30 ms of the STS.*

# 9 Track Domain Processing

## 9.0 INTRODUCTION

This thesis has described a transformation of digitized sound that we have argued will be useful in simulating the kinds of processing performed on sound by human listeners. Unfortunately, we have not yet demonstrated any such processing. This chapter describes our current ideas in this area -- some specific ideas about the kinds of analysis possible under the new system. This will serve to highlight the specific qualities of this domain -- that the sound has been segregated into a finite number of perceptually atomic elements, and that perceptually relevant and irrelevant features can be separated.

## 9.1 TRACK SMOOTHING

We made the point that modelling the sound as tracks essentially threw away much of the information present in the full waveform by ignoring any specific characteristics of the shape of spectral peaks or the valleys between them. However, informal inspection of the track representation for some sounds gives a strong impression that there is still more information reduction possible. In particular, a great many sample points are spent on coding the rapidly modulated tracks at the very top of the spectrum, and yet the amount of information carried by components from 8kHz to the limit of hearing seems very modest when one compares the experience of listening to full bandwidth and bandlimited recordings. There are doubtless many reasons for this reduced salience of the top end; can we exploit any of them to further reduce the information in our representation?

The auditory perception of relative phase is a fascinating problem that we have touched upon already; it also has a bearing on this question of the information in rapidly modulated tracks. An early but very telling experiment in this area, [Math47], concluded that phase was perceptible only insofar as it influenced the envelope of a set of narrowly spaced harmonics. They followed this through with a set of experiments on the perceived roughness of rapidly amplitude-modulated narrowband carriers, which demonstrated that carrier envelope was a significant feature, but that our sensitivity to it was rather limited. An interpretation of this

might be in terms of *neural bandlimiting*: amplitude modulation of a narrowband signal falling within a single critical band is apparently carried as a single percept between the ear and the brain. The effective bandwidth for this channel is finite, if we accept that it is carried on specific, dedicated neurons (a given nerve cell can only fire once every few milliseconds). If the allocation of neurons to critical bands is approximately constant (since their perceptual 'size' is equivalent) then we might expect there will be some point in the spectrum when the increasing bandwidth of the critical bands exceeds the capacity of the neural processing assigned to it, and the perception of these higher frequency channels will be undersampled or smoothed.

This is a rather speculative and untested hypothesis. However, it is consistent with the [Math47] observations, and it might be possible to confirm using our current representation. If these high frequency percepts are bandlimited, perhaps we can apply some kind of smoothing to the higher frequency tracks in our representation, thereby reducing the information needed to represent them. If the smoothing we apply is appropriate, and the data reduction turns out to be inaudible, the hypothesis will be strongly supported.

It remains to decide what 'appropriate smoothing' might be. We are confronted with similar issues to those involved in choosing the sampling rate for a track contour. It is reasonably straightforward to think about the bandwidth of the amplitude modulation in a given channel, and how it can be smoothed by low-pass filtering. But frequency modulation has a rather different manifestation - we would not expect it to be low-pass filtered, since this would, in general, alter its peak-to-peak excursion. But there is no apparent mechanism to limit the perception of this excursion - these frequencies are detected by separate hair cells. Our arguments of limited neural bandwidth merely suggest that at some later stage the precise pattern of modulation will not be too closely examined. Thus we might smooth the frequency modulation contour, but adjust it subsequently to maintain its peak frequency excursion over some time window.

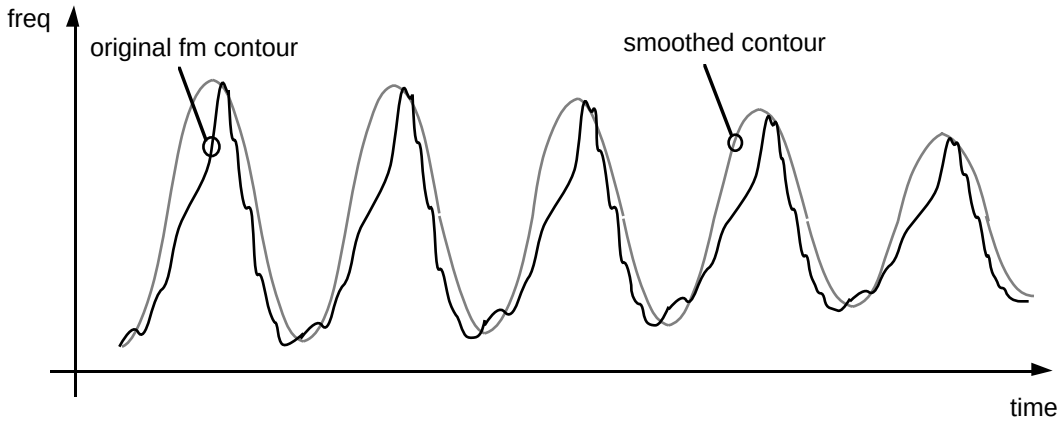



*figure 9.1 - A frequency modulation track that has been smoothed but renormalized to maintain its peak frequency excursion.*

## 9.2 NOISE LABELLING

We have said that the basis of the representation was a perceptual tendency to look for and focus upon stable time-frequency elements in sound, and thus we extract these as our tracks. However, there are many sounds for which this model is inappropriate -- we consider them as 'noisy' signals, underlining the fact that perceptually we cannot characterize them with any precision, although they are if anything a richer class of signal. When a noisy signal, such as a speech sibilant, is analyzed by our system, the resulting tracks have several distinct characteristics: they tend to be short in duration, and have a high degree of modulation (since the tracks are essentially formed from independent energy concentrations that happen to fall within the thresholds for forming a track). There are many tracks closely spaced over the time and frequency extent of the noise, but they do not exhibit cosynchrony or correlation.

These qualities are a problem for our system for several reasons. One strong motivation for forming tracks is to segregate the acoustic energy into relatively large regions that can processed as a whole. But with noise signals, we end up with a very large number of tracks related to a feature that is comparable in perceptual importance to a single harmonic in a pitched sound; our processing leverage has been lost.

Another problem arises in resynthesis. We have remarked that our synthesis strategy is assuming that different spectral peaks are not interacting, but this becomes less and less accurate as the peaks get closer together. Although that observation was in relation to the measured peak frequency, there is a similar departure from independence in the magnitude measurements. The effects of resynthesizing noise represented as a multitude of close, short-duration tracks include amplifying the noise feature (since the magnitudes have been overestimated) and a proliferation of any artefacts associated with the onset and offset of resynthesized partials.

The solution to these problems is to recognize such noise features and give them special treatment. This is strongly reminiscent of Xavier Serra's "stochastic plus deterministic" analysis [Serr89]. He picked out the well-resolved harmonics in a narrowband Fourier analysis of sounds, zeroed-out these portions of the spectrum, then modelled the residual sound with a low-dimensional broadband parameterization. On resynthesis, this general residual spectrum for which no clear harmonic peaks had been found was used as a modulating envelope for white noise, with extremely good perceptual fidelity.

Our time-frequency distributions are very different from Serra's, but the general idea of sectioning off the noise regions and describing them with just a few parameters is most appropriate. By building a feature detection layer that looked for sets of tracks exhibiting the characteristics of noisy regions we described above, and then summarizing them all into some kind of monolithic noise feature, we could regain our processing and data reduction advantages without sacrificing perceptual quality.

There is one catch to regions labelled as noise: since the definition has become more vague, it becomes very difficult to separate overlapping noise regions - a labelling scheme will tend to merge them. In recognition of this tendency, we should give subsequent stages of organization a preparedness to break up noise regions to extract subregions, where this is consistent with other cues. Naturally, the rules for this kind of analysis would need careful investigation and formulation.

## 9.3 PERFORMANCE EXTRACTION

There are a number of potential applications of our structured and part-interpreted representation of sound. We have developed the representation with a view to resynthesis of actual sound, but even without building a complete source separator, there is some information that could be very usefully derived.

By performance extraction we refer to recovery of some kind of constrained control stream from an audio recording. For instance, from a recording of piano music one might be able to use matched filters to reconstruct the sequence of key strike events. This would be very useful information, but it would not be a perceptually complete description of the sound; for that, we would need additional information such as the precise dynamics, the tone of the piano, and the ambience of the recording. Also, the limited domain of the output would be unable to describe anything on the recording that differed from its intended material - for instance, a fragment of voice. In this limitation, performance extraction can be viewed as a subset or cut-down version of complete perceptual parsing of audio which source separation approaches more closely (although the problems are different).

The nature of the tracked representation is clearly suitable for pitch-time transcriptions of tonal music. When the number of voices is small and the pitch fundamental is strong, the notes of the melody often have a direct correspondence to the tracks representing those fundamentals, so the transcription is a relatively simple matter of recognizing these tracks as such (perhaps by constraints of frequency and magnitude, with some mechanism for suppressing low harmonics), and recording their start times and modal frequency[1].

[1]This makes melodic extraction sound too easy. One of the frustrating things about tonal music from the point of view of audio analysis is that simultaneous notes very frequently are arranged to have many coincident harmonics - the worst possible arrangement for algorithms wishing to pretend that Fourier magnitudes superpose.

There are other useful applications of this kind. It should be easier to build a feature detector on the track representation than on the raw spectrogram, since the salient features have been isolated, and there is less data to search. Thus detection of complex audio events (such as percussion pattern extraction) should also be made easier.

# 10 Track-Based Source Separation

## 10.0 INTRODUCTION

The previous chapter described some potential developments and applications of the track-based system. However, the main design objective was a system useful for source separation. Although we have discussed the general strategy of simulating this ability on a computer, we will now describe in more detail how we expect this to be accomplished using the track-based representation.

## 10.1 BASIC SOURCE SEPARATION STRATEGY

Before talking about how source separation cues are applied to the track representation, it is necessary to set out the overall strategy for regenerating the separate sources from the original sound. We have striven to form tracks so that each track represents an indivisible portion of sound energy that can only have come from a single source. The problem of source separation is then reduced to sorting these tracks into non-overlapping sets, where all the tracks from a given source end up alone in their own set. We do this by establishing weighted links between tracks and groups of tracks by application of the various cue heuristics described below, and then applying a clustering algorithm to the resulting network to maximize the average within-cluster links and minimize between-cluster links. This algorithm will therefore partition the set of all tracks into as many sources as are required sensibly to represent the sound it is given. The clustering method has not been devised, but will be some kind of iterative approximation like the k-means algorithm.

## 10.2 CUES TO SOURCE SEPARATION

We have previously surveyed the results of psychoacoustic experiments to determine the cues necessary for source separation. Some of these rely explicitly on memory of what particular things sound like, but this kind of knowledge is outside the scope of the current separator design. Of the remaining cues, the following sections describe how they could be exploited in the track-based representation.

## 10.3 COMMON ONSET

This is probably the most important cue we are dealing with. Where harmonics are resolved as single tracks, it is easy to imagine sorting these tracks by start time and establishing links between those whose creation occurs within some local time window. The variable time resolution across bands means that this window of error needs to be variable across the spectrum, but that should not be a difficulty.

## 10.4 HARMONICITY

Harmonic relations between resolved partials is generally accepted as a very important cue to fusion. This information is readily extracted from the tracks; each resolved harmonic track could trigger a search for overtones located at integer multiples of its own frequency, and form links to any it finds. This would not directly link the second and third harmonic of a given pitch, although their linkage would be established if the fundamental if it were present, since they would both be linked to it. Of course, there is a well-known result in psychoacoustics that the common period of a sound is correctly identified even when the energy of the fundamental is completely removed or masked. This could be simulated by the computationally more expensive, but physiologically equally plausible strategy of searching the entire frequency space for sets of tracks falling into a harmonic pattern to a greater or lesser extent . Then the fact of a particular harmonic missing - fundamental or otherwise - would weaken the score of the match, but could easily be compensated by the presence of other harmonics.

## 10.5 PROXIMITY

The concept of proximity is that events that occur in a similar time-frequency locale may be related. In this case, a track starting shortly after the death of a predecessor at about the same frequency may well be the 'reappearance' of the same phenomenon. Looking at examples of the tracked representation of sounds such as voice shows that while the low frequency harmonics may be adequately linked by their common onsets, the higher-frequency energy is typically modulated onto separate tracks for each pitch pulse. Common onset would only group the first of these with the longer-duration harmonics without some mechanism to connect the successive pitch-pulse-tracks to the first. Proximity will serve this role; one appealing solution might be to go through looking for such 'trains' of short duration tracks, and use proximity to

associate them into a *meta-track*, a group of tracks that can behave as a single track for certain purposes, which could be linked as a whole to the lower, stable harmonics by common onset.

Another application of proximity and meta-tracks would be to correct small gaps in otherwise stable tracks, perhaps caused by interfering transients, by grouping together the tracks before and after the break.

## 10.6 COMMON MODULATION

Where spectral energy arises from a frequency-modulated periodic source, the modulation acts as a very powerful cue to the fusion of the relevant components. This is irresistibly demonstrated by the release from fusion of tones with interleaved harmonics as modulation of a few percent is introduced to one [McAd84]. Simulating this effect in the track domain was one of the design criteria - a more narrowband system might 'smear out' the necessary information - but still presents some problems.

For low frequency, resolved harmonics, common modulation is apparent as simultaneous, correlated disturbances on the frequency or magnitude contours. This can be detected by evaluating pairwise similarity between candidate tracks, although this seems expensive. A more efficient (and plausible) solution might be to to have some modulation feature detector run over the tracks and flag moments of modulation perhaps above some threshold. The instants of these reasonably sparse events can then be matched between tracks. This is a version of the primitive FM-detectors described in [Verc88] and [Mont89].

For the pitch-cycle-modulated higher frequency tracks, there is no explicit frequency attribute (it might be implicit in the temporal spacing of tracks in the meta-track to which they might belong). However, the kinds of period-to-period fluctuations we are primarily interested in detecting should be particularly obvious for these tracks since it will be manifested as systematic displacement from the expected pitch-pulse-time of *all* the formant peaks derived from a particular source. Thus, our common-onset machinery should do this job, of linking high-frequency peaks by their short-term frequency variations.

Common amplitude modulation is also important, and should be amenable to the same approaches of modulation-feature-detectors and simple common onset recognition.

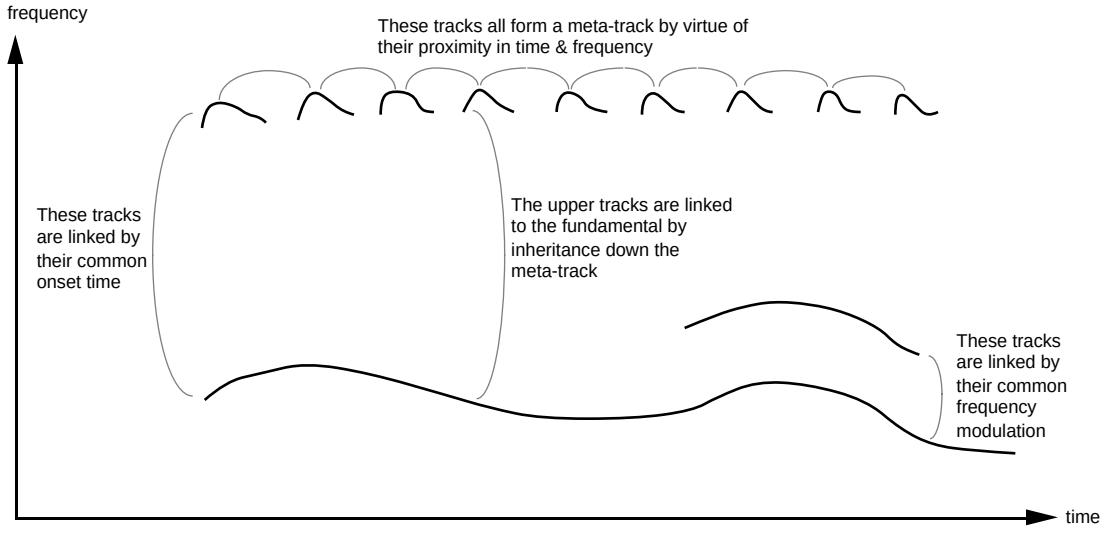



*figure 10.1 - Simplified diagram showing how tracks might be grouped by various heuristics*

## 10.7 SPATIALIZATION

Where binaural localization information is available, this has a profound influence on segregation. I have deliberately excluded it from my model by using but a single channel of sound. However, a two-channel binaural system could be extended to spatial cues by some kind of cross-correlation between tracks at the same time and frequency in the two ears. This would then generate location estimates for each 'stereo track' (as often as tracks from each ear could be linked), and proximity of perceived location could add another set of common-source links.

## 10.8 CONCLUSIONS

One function mentioned in chapter 2 that we have not addressed here is that served by memory - both in guiding separation to expected components, and in 'inventing' components that are most likely to be present, but are not explicitly detected through processes of masking. These functions are very important but very difficult in practice. The result of their absence is that in areas of interfering or masked energy, there tend to be audible 'holes' in the reconstructed spectra -- when separated from its original companion signal, regions that were previously masked become starkly

audible and conspicuously silent. This shortcoming must be addressed somehow, but we shall not speculate further here.

We have explained how the known significant cues to source separation could be detected from the track representation. Most of these methods can be implemented as simple operations or low-order searches, as was intended for this transformation. Although the prognosis is good, it must be recognized that the systems described are far from being implemented, and there is a big gap between something that sounds promising and something that actually works.

# 11 Conclusions and future work

## 11.0 INTRODUCTION

We have described our new representation based on auditory processing of sound. This representation has been motivated by the idea of finding a domain for signal processing functions similar to those performed by human listeners, in particular the separation of sounds from different sources. Such processing has not yet been implemented, but we have discussed in depth how this would proceed.

A key feature of the representation was that it should break the sound into elements that could be inverted, as a whole or in subsets, back to sounds with a clear perceptual relationship to the original sound. In particular, that the inversion of an unmodified representation should be perceptually equivalent to the original sound ("sound the same"). This has been essentially achieved : despite considerable information loss through the chain of processing, the resyntheses are in many cases indistinguishable from the originals (at least in informal listening tests), suggesting that there is a good match between the characteristics stored in the representation and those relevant to a sound's perceptual quality.

## 11.1 UNRESOLVED ISSUES FOR FUTURE EXAMINATION

As noted in the introduction, this is simply the first stage of a longer project. Chapters 9 and 10 have already described at length some of the next stages of development for the tracked representation of sound. However, there are also some other issues that need further examination.

There are several obvious weaknesses to our claim to be using a model of perceptual processing. Firstly, the filterbank bears only a schematic relationship to the cochlea: we have not used a Bark frequency scale, nor have we sought to match the asymmetric frequency responses observed for individual fibres. The mechanisms of automatic gain control have been omitted, and there is no explicit or accurate modelling of critical band masking. To all these, we can only make defence that we felt they were not utterly critical to the phenomena we were seeking to discover, that we made our model only as close as we thought necessary to reproduce the important

effects. [Marr82] talks about the issue of robustness in relation to implementations of (visual) perceptual mechanisms. Our belief was that we should be implementing mechanisms sufficiently robust to perform in a recognizable way even under these approximate conditions.

There have been passing references to data reduction. Since we are seeking to represent only salient features and discard irrelevant ones, we should be able to achieve some storage gain. In practice, data reduction requires very careful attention to *resolution* of all parameters, as well as a happily compact bandwidth of sampled streams. The more complex one's processing chain, the more exotic and unexpected the manifestation of quantization and aliasing distortion. Since this analysis is arduous, we have not pursued this route for this work; we have erred far on the side of caution in both resolution and sampling rates, showing huge *increases* in storage requirements compared to raw digital sound! None the less, there are compelling reasons to expect a gain in this area worthy of investigation.

Although we have used phase measurements to improve our resynthesis, the data is otherwise ignored. Given its importance in this final stage, it would be more consistent to use it throughout, and indeed to establish more systematically why it is useful. This would appear to tie-in with the use of 'synchrony' measures by [Cook91] and [Ghit87].

One matter we were frustrated to lack the time to address was the principle of 'old plus new' introduced in chapter 2. In theory, the ear could improve its bottom line sensitivity by 'subtracting out' expected spectral components from what was actually heard, to make low-energy components more prominent. The consistency of critical-band masking phenomena argue against this happening at a particularly low level, but we have been very tempted to employ some kind of 'expected spectrum feedback' based on current tracks to improve the discrimination on the output of the filterbank. If successful, this could give rise to an equivalent tracked representation (amenable to the same resynthesis method) but with much less interaction between the tracks, and much better scope for separation.

Overall there are many questions raised by this initial foray into perceptual signal processing. However, the time is ripe for modelling of human functions at this level, and we look forward with great excitement to the resolution of these issues in the near-to-medium future.

# Appendix A Filterbank Analysis

## INTRODUCTION

In this appendix we describe the filterbank mathematically and derive its overall frequency response to demonstrate that it is essentially flat over its pass band.

## THE INDIVIDUAL FILTERS

As we said in chapter 4, each bandpass filter in the bank is derived from a truncated Gaussian prototype. The filterbank is a constant-Q or wavelet array, so the different bands are generated by dilating the prototype. The prototype was designed to have a bandwidth of one quarter its center frequency at its -17.4 dB points ($1/e^2$). Thus our prototype filter, centered at $\omega_0$ and defined in the frequency domain,is:-

$$H_{\omega_0}(\omega) = \exp\left\{\frac{-k^2}{2}\left(\frac{\omega-\omega_0}{\omega_0}\right)^2\right\}$$

where k controls the width of the Gaussian envelope at $\omega_0$. To achieve our desired bandwidth, we want the response to be reduced to $1/e^2$ one-eighth of the center frequency each side of the peak, i.e.:-

$$H_{\omega_0}\left(\frac{7}{8}\omega_0\right) = H_{\omega_0}\left(\frac{9}{8}\omega_0\right) = \frac{1}{e^2}H_{\omega_0}(\omega_0)$$

so $$\frac{1}{e^2}\exp\{0\} = \exp\left\{\frac{-k^2}{2}\left(\frac{1}{8}\right)^2\right\} \quad \text{thus,} \quad k=16$$

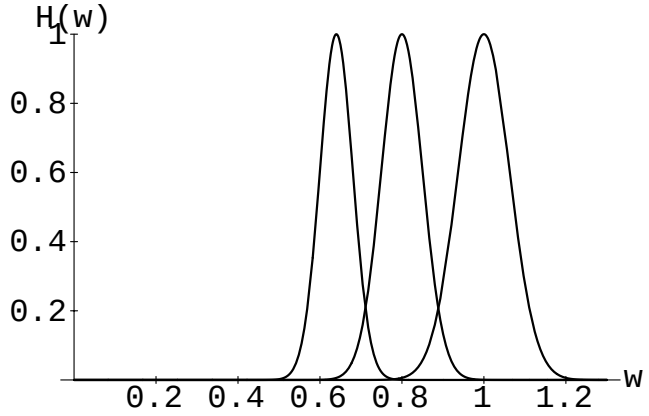
Now, the Fourier transform of a Gaussian is:-

$$h(t) = \exp\left\{\frac{-t^2}{2}\right\} \Leftrightarrow H(\omega) = \exp\left\{\frac{-\omega^2}{2}\right\}$$

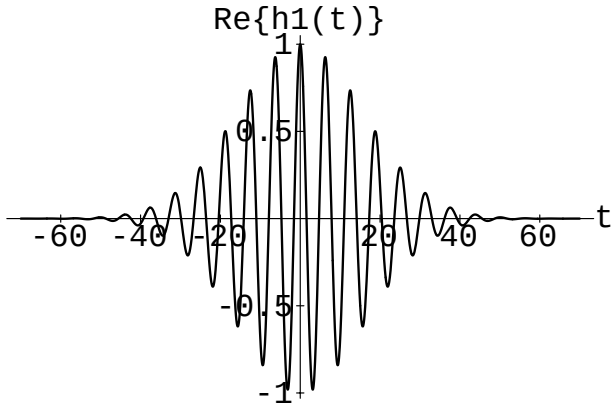
so impulse response of each filter is given by:-

$$h_{\omega_0}(t) = \exp\left\{-\left(\frac{\omega_0^2 t^2}{2k^2}\right) + j\omega_0 t\right\}$$

This frequency response is shown below for $\omega_0$= 0.64, 0.8 and 1.0:-




and the corresponding impulse response for $\omega_0$=1:-




We use a 256 point truncated FIR implementation of this time response, and the lowest frequency filter, which has the least compact time response, is at $f_{samp}/8$ i.e. 8 samples per cycle of the carrier. 256 points allows for 16 cycles each side of t = 0, taking us to t = 16x2π or approximately 100 on the diagram above, by which time the amplitude of the impulse response has fallen to below the noise floor intrinsic to the 16 bit sample values of our system.

## COMPOSITE RESPONSE

We have stated that the individual filters in the bank are formed by dilating a prototype. If we define our lowest filter at frequency $\omega_L$ by the prototype, $H_P(\omega)=H_{\omega_L}(\omega)$, then the $n^{th}$ filter in the bank will have a frequency response:-

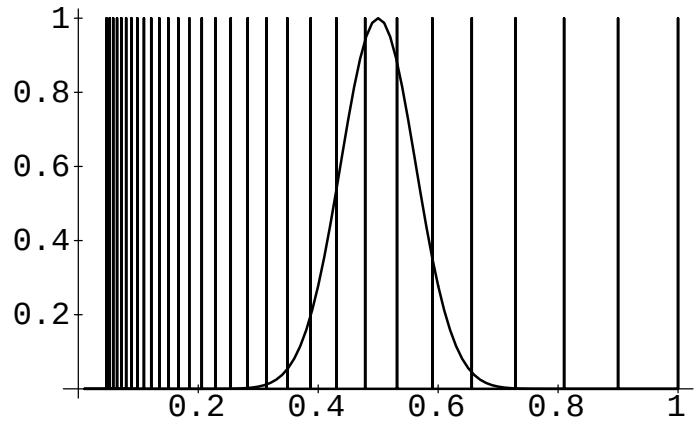
$$H_n(\omega) = A_n \cdot H_P\left(\frac{\omega}{\alpha_n}\right)$$

where $A_n$ is a weight applied to each filter, yet to be determined, and $\alpha_n$ is the dilation factor i.e. the factor by which the center frequency of this filter is increased beyond $\omega_L$. Since we have chosen to space our filters exponentially, and thereby obtain uniform overlap between adjacent bands, we must make $\alpha_n = r^n$, where r is the ratio of center frequencies between adjacent bins. In our system of 12 bins per octave, $r = \sqrt[12]{2} \approx 1.06$.

We are interested in the character of the overall frequency response of our bank of N filters, which we obtain by summing the individual responses, i.e:-

$$H_T(\omega) = \sum_{n=0}^{N-1} H_n(\omega) = \sum_{n=0}^{N-1} A_n \cdot H_P\left(\frac{\omega}{r^n}\right)$$

We can understand this equation with the help of the following diagram:-




The curve represents $H_P(\omega)$, and the lines on top represent terms of the summation for $H_T$ , evaluated at a particular frequency value $\omega_x$. (in the diagram, $\omega_L$ is 0.5 and $\omega_x$ is 1.0). The rightmost vertical line shows where the term for n = 0, $A_0 \cdot H_P\left(\frac{\omega_x}{r^0}\right)$, is sampled -- $H_P(\omega)$ is very close to zero here, so this term does not

contribute significantly. It is only the samples that fall over the raised part of the curve that contribute to the overall sum.

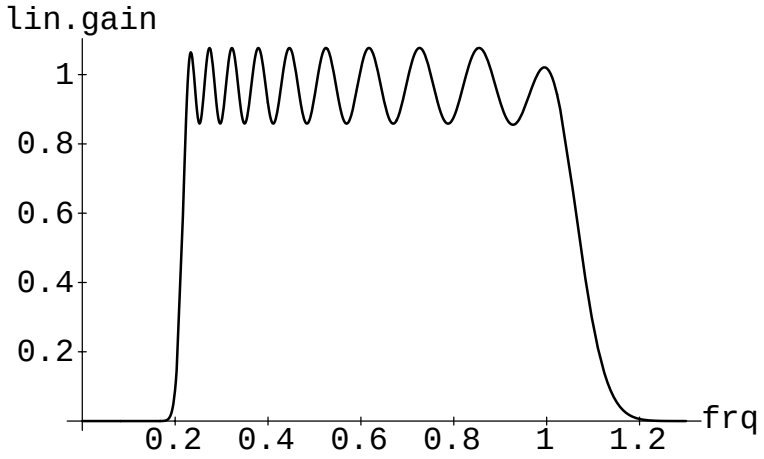
The vertical lines occur at $\omega=\omega_x/r^n$ for n = 0 to N-1, the values in the summation. Starting from the right, each line divides the space remaining between its predecessor and the origin by a factor of 1/r i.e. by a constant ratio. Now, ignoring the upper limit on N, the pattern would look the same for larger frequencies $\omega_y$ as long as one of the lines fell upon one of the lines in our picture, i.e. $\omega_x = \omega_y/r^n$, since all the remaining lines must fall into the same pattern. Thus if we use a constant weighting, say $A_n=1$, then our system will have uniform gain at a set of frequencies $\omega_x$ , $r\omega_x$ , $r^2\omega_x$ etc, which is on the way to becoming the flat frequency response we are seeking.

In between these points of equal response, the sampling pattern will be slightly different, which will result in ripple in our frequency response. This can be kept small by making the spacing between the sampling points small relative to the rate of change of the underlying curve so that the value does not change much between adjacent samples, which is equivalent to having a closely spaced or highly overlapped filterbank.

The remaining question is what happens at the limits of the frequency response. The lower limit occurs when $\omega_x$ begins to approach $\omega_L$ , the center frequency of the prototype from above. As $H_P(\omega_x)$ begins to differ significantly from 0, the terms in the summation giving the total response start dropping out and the gain falls, eventually to zero.

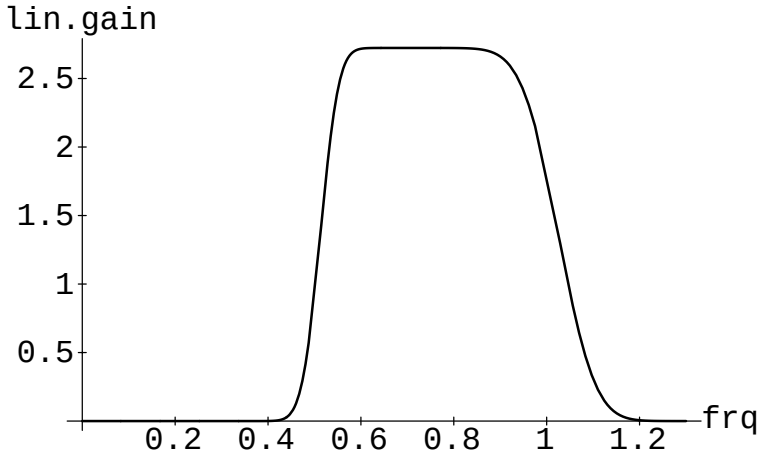
At the high frequency, we are constrained by the size of our filterbank, i.e. the upper limit N on our summation. The furthest *left* sampling point has a frequency $\omega_x/r^{N-1}$ ; when this begins to approach $\omega_L$ from below, we again start to lose terms from our summation, and the gain again falls to zero.

The overall response of a 10 element filterbank using our prototype Gaussian with a spacing of r = 1.18 is shown below. The relatively coarse spacing makes the ripple rather severe. We can also see the asymmetric slopes at high and low limits.




Note that uniform filter weights are used for convenience. Careful selection of filter weights can improve the flatness at the band edges, as the above diagram suggests. For more details on such weightings, see [Musi83].

The next diagram shows the expected response of one octave of the same filters at the spacing actually used (r = 1.06). The ripple is no longer visible. This should be compared with figure 4.2 which shows the actual impulse response of our filterbank implementation on a logarithmic magnitude scale.



# Appendix B Peak Picking Analysis

## INTRODUCTION

This appendix is intended to provide a little more rigor to our assertion that the result of picking peaks from among the broad filters of the constant-Q filterbank is able to track the actual frequency of resonances such as speech formants, rather than just the harmonic closest to the resonance.

## RINGING OF RESONANCES

Perhaps the quickest argument to support this is to look at the time domain waveform of a simple resonance excited by an impulse train with a repetition rate longer than the time it takes the ringing of the resonance to die away. Figure B.1 shows a 1 kHz resonance with a Q of 10 being excited by a 95 Hz impulse train.

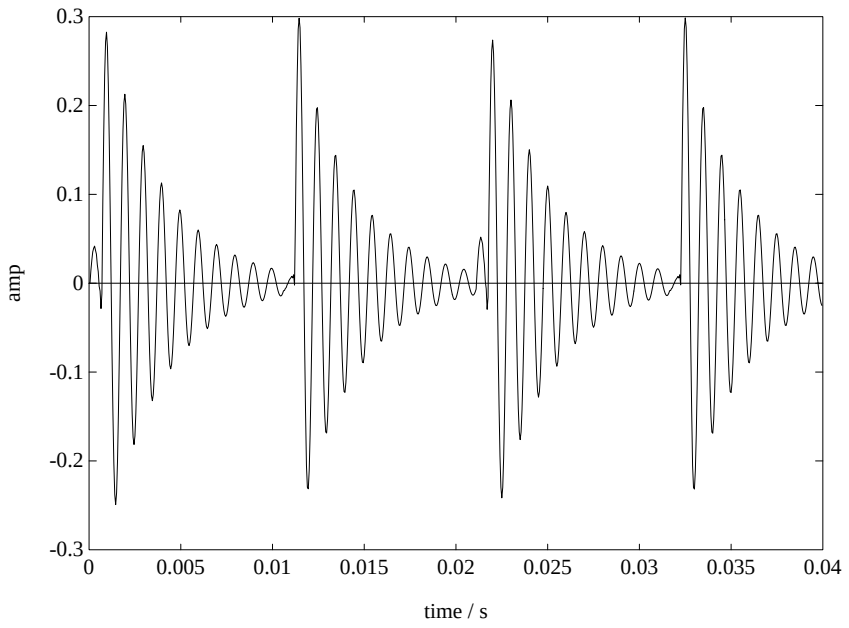



*figure B.1 - a 1 kHz resonance excited by a 95 Hz impulse train shows clear 1 kHz ringing even though there is no harmonic of the excitation at that frequency.*

Although the nearest harmonics of the impulse train are at 950 Hz and 1045 Hz, it is clear that the ringing occurs with an exact 1 ms period since it is essentially a copy of the impulse response of the resonance. Thus a filter with a time window short

enough to only extend across the ringing from one pulse at a time will report maximum energy for the filter with peak gain at 1 kHz, at least in-between the impulse instants. The interpolation along frequency to find spectral peaks is essentially simulating an infinite bank of bandpass filters spanning every possible center frequency; a peak in our smoothed spectrum indicates that one of our bandpass filters centered on that frequency would have a larger output than its immediate neighbors.

## SUMS OF HARMONICS

A more mathematical investigation of grouped harmonics follows. Consider the position illustrated in figure B.2:-

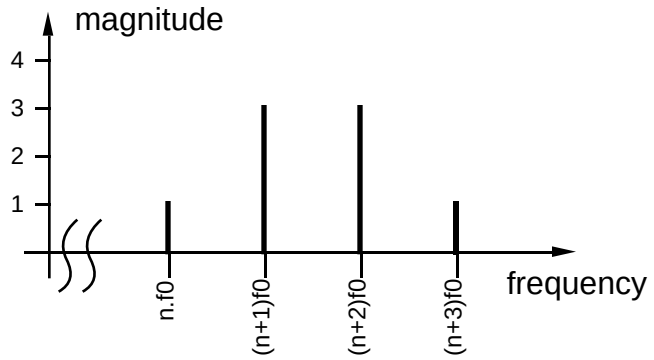



*figure B.2 - four harmonics shaped by a resonant filter whose peak lies between harmonics (at (2n+3)f0/2 )*

What we have is four harmonics of a fundamental of frequency f0; our harmonics range in frequency from nf0 to (n+3)f0. The harmonics are relatively weighted 1, 3, 3, 1 as shown (these numbers are chosen carefully, but we will generalize slightly below). This could have resulted from a locally flat spectrum (as if from an impulse train) passing through some kind of resonance. By symmetry, we see that the peak of such a resonance appears to lie midway between the two strongest harmonics, and in any case not directly over any harmonic.

If we consider the time-domain waveform of the signal drawn, assuming sine phase alignment, we have:-

$$x(t) = \sin(n\omega t) + 3\sin((n+1)\omega t) + 3\sin((n+2)\omega t) + \sin((n+3)\omega t) \quad (1)$$

(where $\omega = 2\pi f0$). We can rearrange this as (for reasons that will become clear:-

$$\begin{aligned} x(t) = \quad & \left(\sin(n\omega t) + \sin((n+1)\omega t)\right) \quad + \left(\sin((n+1)\omega t) + \sin((n+2)\omega t)\right) \\ & + \left(\sin((n+1)\omega t) + \sin((n+2)\omega t)\right) \quad + \left(\sin((n+2)\omega t) + \sin((n+3)\omega t)\right) \end{aligned} \quad (2)$$

Now, we recall the trig identity:-

$$\sin A + \sin B = 2\sin\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right)$$

so
$$\sin(n\omega t) + \sin((n+1)\omega t) = 2\sin\left(\frac{2n+1}{2}\omega t\right)\cos\left(\frac{\omega t}{2}\right) \quad (3)$$

We can apply this to each pair in (2) to get:-

$$x(t) = 2\cos\left(\frac{\omega t}{2}\right)\left\{\begin{array}{l} \sin\left(\frac{2n+1}{2}\omega t\right) + \sin\left(\frac{2n+3}{2}\omega t\right) \\ + \sin\left(\frac{2n+3}{2}\omega t\right) + \sin\left(\frac{2n+5}{2}\omega t\right) \end{array}\right\} \quad (4)$$

We can further apply (3) to each row inside the brace of (4), and then again to the resulting two terms to get:-

$$x(t) = 8\cos^3\left(\frac{\omega t}{2}\right)\left[\sin\left(\frac{2n+3}{2}\omega t\right)\right] \quad (5)$$

We can consider this as a sinusoid at the center frequency of the resonance $(2n+3)f0/2$ amplitude modulated by the carrier $\cos^3(\omega t/2)$. Since we assume $n\omega >> \omega/2$, the carrier and modulation will be relatively independent. Below we see the case for n = 8:-

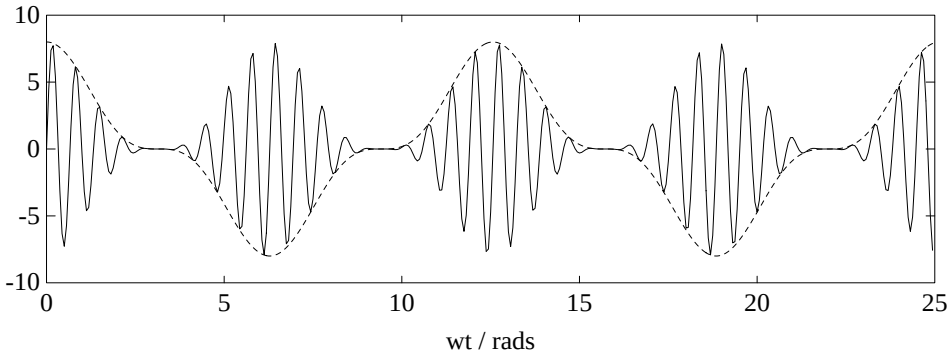



*fig B.3 - the weighted sum of harmonics 8 through 11 (solid) with the $cos^3$ modulation envelope (dotted).*

A short-windowed bandpass filter passed over this signal will show maximum output if it is tuned to the carrier frequency, (2n+3)f0/2, so this is where the track will be located. Also notice that the carrier amplitude falls to zero once every cycle of f0 (twice each cycle of ω/2), so we would expect the pitch-cycle modulated trains of tracks we have described before.

This kind of simplification will work for sums of sines with weightings drawn from Pascal's triangle (which approaches a Gaussian since it is the repeated convolution of a two-element pulse). If the amplitudes of the harmonics do not fall into this pattern, we can model them as such a pattern plus excess terms from the difference between the modeled and actual magnitudes. If the phases result in a modulation pattern like the one above, we might see the carrier picking up the track when its amplitude was strong, but the track migrating to one of the excess residual terms when the modulation envelope made the carrier weak. This might result in the kind of periodic modulation between resonance peak and nearest harmonic that we saw in figure 7.5.

# Appendix C Computing Environment

## INTRODUCTION

This appendix describes some technical details of the hardware and software used in the project.

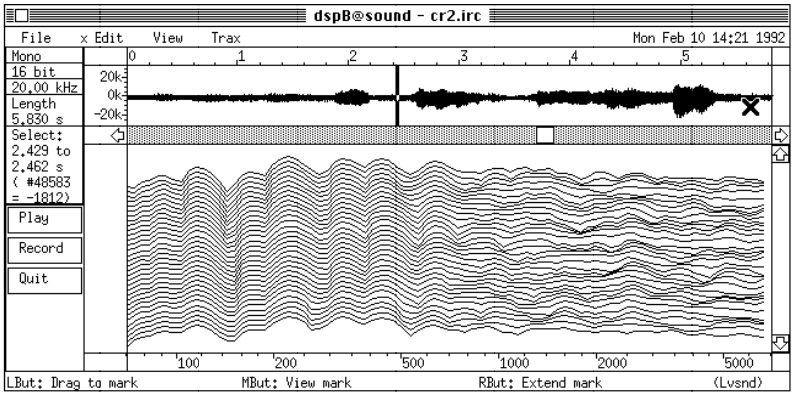
## HARDWARE

All work was done on unix workstations - primarily a DECstation 5000/200, and latterly an SGI Iris Indigo. The DECstation had CD-type sound output via the DECaudio interface unit, and the Indigo has excellent sound input and output capabilities built in.

## SOFTWARE

The system was developed under unix in C, using the ansi-compliant Gnu compiler, gcc. It is intended to be portable. The main components are listed below. This software is in the public domain and is available from the author who may be reached via email as dpwe@media.mit.edu.

| | |
|---|---|
| cqt3, cqi3 | Convert between 16 bit PCM soundfiles and constant-Q time-frequency analysis files. Allows specification of Q of Gaussian filters, number of octaves, bands per octave etc. |
| trk | Build a track model of a constant-Q analysis file. |
| tksyn | Resynthesize a track model directly into PCM sound. |
| dspB | Graphical environment for inspecting all the above data. Includes time domain, gray-scale spectrogram and perspective plots. Allows interactive selection of groups tracks overlaid on spectrogram for partial resyntheses. A typical display is below. |



# References


[Breg90] AS Bregman (1990) *Auditory Scene Analysis*, Bradford Books MIT Press

[Blau83] J Blauert (1983) *Spatial Hearing*, MIT Press

[Cook91] MP Cooke (1991) “Modelling auditory processing and organisation” PhD thesis, University of Sheffield Dept of Computer Science

[Dols86] M Dolson (1986) “The phase vocoder : a tutorial” Computer Music Journal 10(4)

[Dove86] Webster P Dove (1986) “Knowledge-based pitch detection” PhD thesis, EECS dept, MIT

[Duda90] RO Duda, RF Lyon, M Slaney (1990) “Correlograms and the separation of sounds” Proc Asilomar Conf on Sigs, Sys & Computers

[Ghit87] Oded Ghitza (1987) “Auditory nerve representation criteria for speech analysis/synthesis” IEEE Tr ASSP 35(6)

[Glas90] BR Glasberg, BCJ Moore (1990) “Derivation of auditory filter shapes from notched-noise data” Hearing Research 47

[Hand89] S Handel (1989) *Listening* Bradford Books MIT Press

[Mahe89] RC Maher (1989) “An approach to the separation of voices in composite musical signals” PhD thesis, University of Illinois, Urbana-Champaign

[Marr82] D Marr (1982) *Vision*, W.H. Freeman

[Math47] RC Mathes, RL Miller (1947) “Phase effects in monaural perception” JASA 19(5)

[McAd84] S McAdams (1984) "Spectral fusion, spectral parsing and the formation of auditory images" PhD thesis, CCRMA, Stanford University

[McAu86] RJ McAulay, TF Quatieri (1986) “Speech analysis/resynthesis based on a sinusoidal representation” IEEE Tr ASSP 34

[Mell91] DK Mellinger (1991) “Event formation and separation in musical sound” PhD thesis, CCRMA, Stanford University

[Mins86] Marvin Minsky (1986) *The Society of Mind*, Simon and Schuster

[Mont89] BM Mont-Reynaud, DK Mellinger (1989) “Source separation by frequency co-modulation” Proc 1st Int Conf on Music Perception and Cognition, Kyoto

[Musi83] BR Musicus, JC Anderson, JP Stautner (1983) “Optimal least squares short time analysis/synthesis” MIT-RLE monograph

[Pars76] TW Parsons (1976) “Separation of speech from interfering speech by means of harmonic selection” JASA 60

[Patt87] RD Patterson (1987) “A pulse ribbon model of monaural phase perception” JASA 82(5)

[Patt91] RD Patterson, J Holdsworth (1991) “A functional model of neural activity patterns and auditory images” in *Advances in speech, hearing and language processing vol 3*, ed WA Ainsworth, JAI Press, London

[Quat90] TF Quatieri, RG Danisewicz (1990) “An approach to co-channel talker interference suppression using a sinusoidal model for speech” IEEE Tr. ASSP 38(1)

[Sene85] S Seneff (1985) “Pitch and spectral analysis of speech based on an auditory synchrony model” PhD thesis, EECS dept, MIT

[Serr89] X Serra (1989) "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition" PhD thesis, CCRMA, Stanford University

[Sham89] S Shamma (1989) “Spatial and temporal processing in central auditory networks” in *Methods in neuronal modelling*, MIT Press

[Slan88] M Slaney (1988) “Lyon's cochlea model” Apple Technical Report #13

[Veld89] RNJ Veldhuis, M Breeuwer, RG Van der Waal (1989) “Subband coding of digital audio signals” Philips J Res 44(2/3)

[Verc88] BL Vercoe (1988) “Hearing polyphonic music on the Connection Machine” Proc Special Session on Music and AI, AAAI

[Verc90] BL Vercoe (1990) “Csound: A manual for the audio processing system and supporting programs” MIT Media Lab Music & Cognition

[Zwic90] E Zwicker, H Fastl (1990) *Psychoacoustics facts and models* Springer-Verlag