Auditory Perception and Speech Demodulation

Robert H. McEachern

Abstract: This paper presents a model of the human auditory system's front-end signal processing. The model is biologically plausible and provides simple explanations for a wide variety of psychoacoustic effects. It is proposed that the auditory system evolved as a threat-warning receiver, long before the development of speech. This threat-warning receiver was subsequently expropriated for use as a communications receiver. It functions primarily as an AM and FM, multi-tone demodulator. FM information is derived from the AM in a manner similar to that by which the eye derives color information. Many of the peculiar characteristics of speech signaling evolved in response to the problems encountered while attempting to use this demodulator to process information transmitted through communication channels exhibiting high-levels of both multi-path and multi-source interference. Similar problems are encountered in designing communication systems that exploit high-frequency (HF) ionospheric channels. It is not surprising then, that many analogies exist between the structure of speech signals and certain types of HF modem signals. It is proposed that these analogies are not coincidental; they reflect a common set of solutions to a common set of problems. Computer simulations have confirmed that good quality speech can be reconstructed from the model's outputs; the model may be thought of as a special form of a harmonic speech coder, designed for use in high noise/interference environments.

1. Introduction

Human eyes and ears are not designed to analyze all the light and sound that happens to enter them. For example, you cannot see the absorption lines in the solar spectrum, nor can you reconstruct a fax image in your head, by listening to a fax modem signal as it is transmitted over a telephone line. The reason why our sensory systems cannot perform the above tasks, is quite different from the reasons why they cannot perceive ultraviolet light, or high frequency ultrasonic sounds. In these latter cases, our sensory receptors are simply insensitive to frequencies outside of a limited bandwidth. But the solar absorption lines and modem signals occur within our receptor's bandwidths. The problem is not that our senses cannot detect such signals, but that they are not designed to extract all, or even most, of the information contained within the frequency bands to which they are most sensitive.

Stated in this manner, this seems like a fairly innocuous conclusion. However, restated in a somewhat different manner, it appears rather astonishing; our sensory apparatus "knows" what it is looking for, and is deliberately ignoring everything else. Consequently, our senses are not designed to inform us about "what is out there", but to tell us if what is "out there" matches what they are looking for. What are they looking for?

Viewed from this perspective, our senses seem to act more like signal demodulators, rather than as some sort of general-purpose signal analyzers. A demodulator is a device that exploits an a priori model of some signal of interest, in order to exclude from its output, signals that do not behave like the signal of interest. A central premise of the auditory system model presented in this paper, is that the original signal of interest was not speech. The receiving system was built to demodulate signals other than speech. Speech evolved much later, and was constrained to use an auditory system, ill-suited to high-speed data transmission.

In section two of this paper, we describe the proposed process, by which the auditory system extracts information from acoustic signals. In section 3, we discuss psychoacoustic evidence, which supports the proposition that the auditory system employs this process. We also discuss the process in relation to the harmonic coding of speech. In section 4, we discuss the likely evolution of this process and its biological plausibility. In section 5, we consider the structure of speech, in light of the limitations imposed by this receiver processing. We also discuss the implications of analogies that exist between the structure of speech and certain types of HF modem signals.

2. An auditory system model

The human auditory system contains a pair of channelized receivers. Within the cochlea of each ear, groups of hair cells combine to effectively form a set of bandpass filters, or channels. Each filter is tuned to a specific center frequency and has a characteristic shape and bandwidth that depends upon that center frequency. The output from each of these channels is encoded into neural firing patterns, that seem to encode the filter's output "voltage", half-wave rectified, then lowpass filtered. At frequencies above the lowpass cutoff frequency, this structure acts as a bandpass filter followed by an envelope (amplitude) detector; it is an AM demodulator. At frequencies below the lowpass cutoff, "phase locking" occurs, such that output neural firings are highly synchronized with the alternating, half-cycles of a sinusoidal input signal, responding to what we shall call the positive half-cycles, but not the negative.

Although we have ignored many details, this brief description is rather uncontroversial. However, the nature of the subsequent processing has been the subject of long debate. This subsequent processing is the primary subject of this paper. We shall describe a model that addresses (1) what this processing does and (2) why it does it. The latter topic will consider both the origins (evolution) of the processing as well as its optimization for certain functions. We shall also present evidence to support the validity of this model, in the form of comparisons between the model's behavior, and that of humans subjected to various psychoacoustic tests. There is a large literature describing and analyzing such effects. Most analyses conclude that the effects must result from unconventional signal processing; conventional signal processing does not exhibit all the same effects. To account for these many unusual effects, it is frequently assumed that many different ad hoc processes must be employed, such as using one frequency measuring technique at low frequencies, and a different technique at higher frequencies (below and above the neural phase-locking frequency, respectively). It is argued here that there is only one basic process at the heart of most of these unusual effects. That process is illustrated in Figure 1. It is a multi-channel, frequency diversified, AM and FM demodulator. Figure 1 depicts only a portion of one receiver channel. The auditory system contains many such channels, each tuned to a different fundamental frequency, F_n. Each channel consists of pairs of bandpass filters tuned near each harmonic (only three are shown) of the channel's fundamental frequency. The "A" filter is tuned to a frequency slightly less than that of the harmonic, and the "B" filter is tuned slightly above the harmonic. The boxes labeled "R" in the figure, perform a special type of FM demodulation, based on the ratio of their paired inputs.

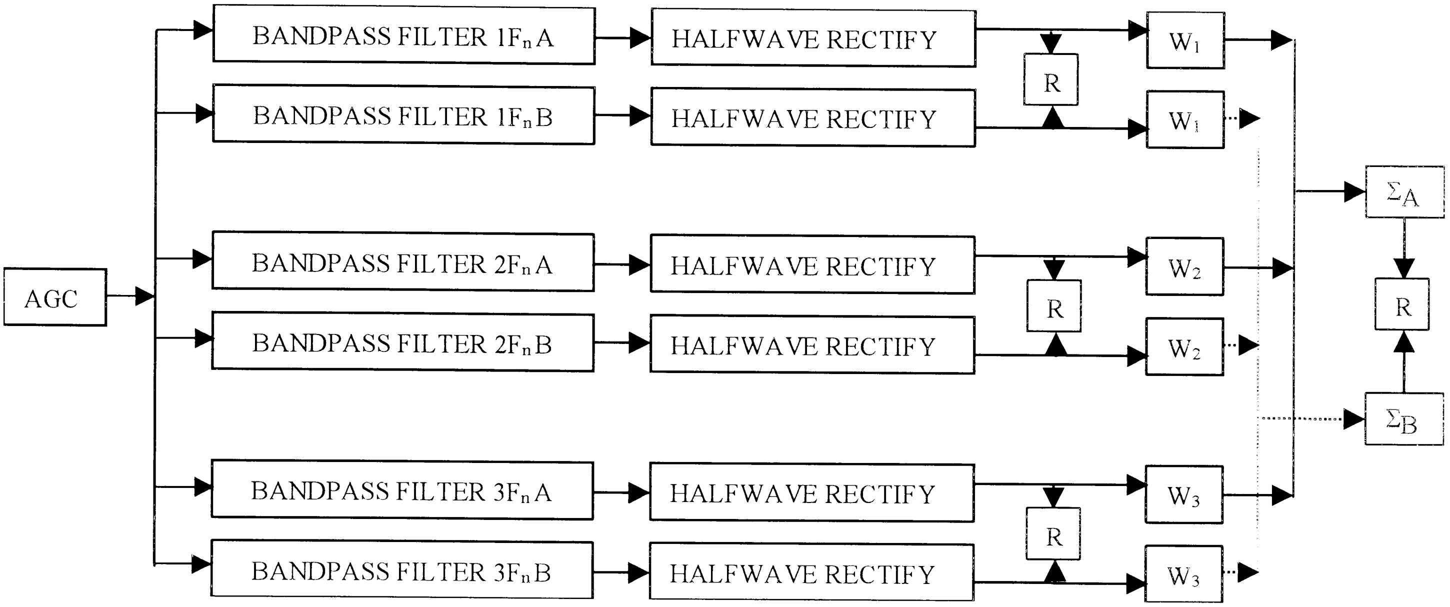

Figure 1: Block diagram of an auditory harmonic demodulator

Weighted combinations of the harmonics (weight factors W_m) are used to produce a weighted estimate of the input's fundamental instantaneous frequency. The principles of operation are described below, beginning with the nature of frequency diversity processing.

2.1 Frequency diversity signaling

The concept of frequency diversity signaling is well-known in the communications field in general, and in high frequency (HF) data transmission in particular. It deserves to be much better known in the auditory function and speech processing fields as well, since it appears to be the answer to the question posed above; What types of signals of interest is the auditory system looking for? The auditory system appears to be built, from the ground up, to look for signals that exploit a particular type of frequency diversity signaling; signaling that produces frequency and amplitude modulated harmonics.

Frequency diversity signaling evolved (both in communications engineering and in the auditory system) in response to difficulties encountered in attempting to receiver data transmitted through channels exhibiting high levels of frequency selective fading. This type of fading commonly occurs when several copies of a transmitted signal arrive at a receiver, from different propagation paths, such as a direct path and one or more reflected paths. These multipath signals interfere with each other, resulting in substantial, time-varying reductions in received signal amplitudes (fading) at various, time-varying frequencies. Consequently, the receiver may be unable to recover any information emitted at a faded frequency, for the duration of the fade. However, if the same information had been transmitted at several different frequencies, chosen so that they were not all likely to fade simultaneously, then the receiver would usually be able to recover all the transmitted data, from one frequency or another, or from some combination thereof. The transmission of the same information at multiple frequencies, in order to overcome frequency selective fading, is known as frequency diversity signaling. Frequency modulated harmonics are a special form of frequency diversity signaling.

If the logarithm of the instantaneous frequency of a fundamental is labeled log(f(t)), then the logarithm of the instantaneous frequency of the n'th harmonic is simply log(n)+log(f(t)). In other words, on a log-scale, the fundamental and the harmonic are identical functions of time, except for a known constant. Consequently, when a transmitter frequency modulates harmonically related carriers, the same information is being carried by each harmonic. Hence, a receiver that "knows" that the signals of interest contain such harmonics, may employ frequency diversity processing techniques to recover the frequency modulated information in the presence of frequency selective fading, caused by multipath interference. The peculiar nature of pitch perception, such as hearing a pitch at the frequency of a fundamental, even when the fundamental is not present in the input, is a direct result of the auditory system's frequency diversity processing. This capability is not an artifact of some ad hoc process. The ability to recover a fundamental's modulation from any "reasonable" combination of time-varying, selectively fading harmonics, is what the auditory system is all about.

2.2 FM derived from waveform ratios

Before one can understand how the auditory system processes multiple harmonics, in order to extract frequency modulation information, it is first necessary to understand how a single, sinusoidal tone is processed. The proposed principle of operation at the heart of the cochlear filter bank is that when a single, slowly modulated sinusoid is passed through a pair of bandpass filters, one output is merely a scaled copy of the other (except for a possible delay, which we shall ignore, for the moment). Furthermore, the scale factor, the ratio of the two, output "voltages" at any time, is directly related to the input instantaneous frequency. This will be true even if the outputs are rectified, or rectified and then lowpass filtered. How the scale factor is related to the input frequency depends on the shape of the filters. As we shall demonstrate below, filters can be shaped to exactly "compute" virtually any desired function of frequency. This enables a system to transduce pairs of voltage, amplitude or power measurements into a linear frequency estimate, a log-frequency estimate, a mel-scale frequency estimate etc., depending on the filter shapes. The filter shapes can even be designed such that all harmonically related inputs will yield the same scale factor, and hence, the same frequency estimate (the frequency of the fundamental). This seems to be the process underlying pitch perception.

Several investigators have discussed the possibility that the auditory system may transform pairs of voltage, amplitude or power measurements, into frequency estimates (McEachern, 1992, 1994a; Dai et al., 1995; Quatieri, et al., 1996, 1997). To illustrate such a process, consider any adjacent pair of Gaussian bandpass filters, within an evenly-spaced filterbank, as shown in figure 2. (Other filter shapes may be used. The significance of a Gaussian shape is that it yields an exact expression for frequency, as in equation (2) below, whereas other filter shapes may only yield an approximate expression, which may only be valid near the center frequency of the filter pair. A Gaussian filter also possesses the minimum possible time-bandwidth product.)

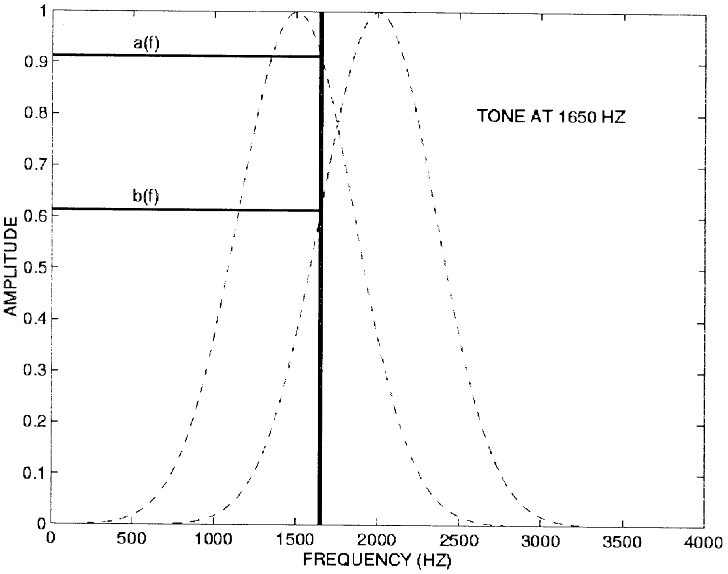

Figure 2: A pair of Gaussian bandpass filters enable frequency estimation; the ratio of their two outputs, a(f) and b(f), is a function of input frequency

When a sinusoidal signal of amplitude A and frequency f is passed through these filters, each filter's
output is an attenuated copy of the input signal. The two output amplitudes, a(f) and b(f), are given by (1):

$$(1) \quad a(f) = Ae^{-(f-N\Delta f)^2/c\Delta f^2}$$

$$b(f) = Ae^{-(f-(N+1)\Delta f)^2/c\Delta f^2}$$

"c" is a constant which effects the filter bandwidth. N stands for the N'th filter in the set of evenly-spaced filters. Assuming that the gain "A" is the same for both filters, then taking the natural logarithm of the ratio a(f)/b(f), and solving for f yields (2):

$$(2) \quad f = N\Delta f + \Delta f/2 - (\Delta f/2)c(\ln(a(f)) - \ln(b(f)))$$

This equation says that a tone's frequency, if it varies slowly, is a simple, linear function of the difference between the logarithms of successive filter pairs' output amplitudes or voltages. Hence, a tone's instantaneous frequency may be determined by a simply linear transformation of a pair of instantaneous amplitude, voltage or power measurements, from a pair of adjacent filters. Furthermore, if the abscissa in figure 2 is relabeled with some other function of frequency, such as log(f), equation (2) will still be valid, provided we make a change of variables from "f" to whatever new variable is specified on the abscissa. In the case of changing the variable from f to log(f), figure 2 would depict a filterbank with filters that are equally spaced along a logarithmic frequency axis. This is a "constant Q" filterbank, in which the filter bandwidths are proportional to frequency. For such a filter bank, the modified version of equation (2) would compute log(f) rather than f itself. A change of variables to a mel scale would result in a filterbank with a bandwidth versus frequency characteristic similar to the auditory system's, and the pitch as computed via the modified equation (2) would correspond to a mel rather than linear frequency scale. This is the primary function of the box labled "R" in figure 1. It effectively computes a ratio (or difference of logarithms; log(a/b) = log(a)-log(b)).

2.3 Harmonically tuned filterbanks

A wide variety of different filterbanks may be designed, by merely choosing different changes of variables. But all such filterbanks employ filters that are identically shaped and uniformly spaced, when plotted against an abscissa corresponding to the changed variable. However, it is also possible to design filterbanks that are not identically shaped in this manner. Some of these have special properties that are useful for processing harmonically related inputs. *Physically*, there is only one cochlear filterbank. But the *logical* architecture may consist of more than one filterbank. In particular, it is interesting to consider a logical architecture in which the physical filterbank is subdivided into many different logical filterbanks, each one consisting of harmonically tuned pairs of filters (a detector), tuned to a different fundamental frequency. One such logical filterbank is depicted in figure 3, with four harmonic detectors tuned to a 500 Hz fundamental.

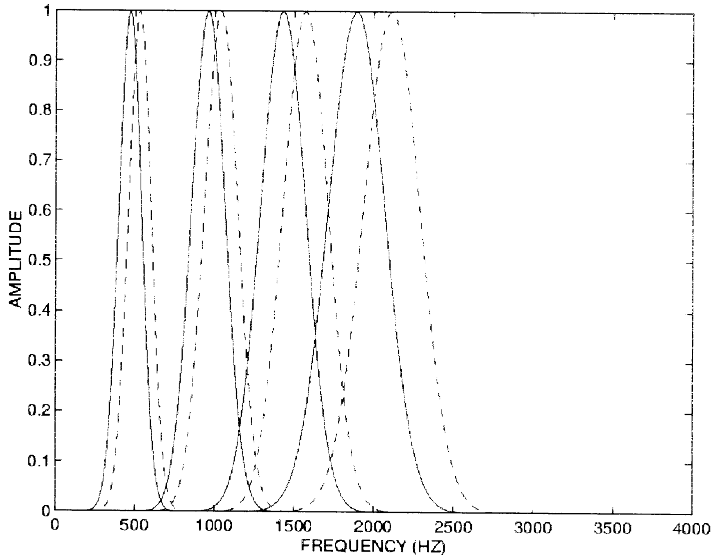

Figure 3: Four dual frequency, harmonic detectors, tuned to a fundamental at 500 Hz

We shall describe two techniques for designing such filterbanks. First, each logical filterbank could be designed individually, with each of its filter pairs being exactly tuned to a harmonic of a specified fundamental frequency. The entire physical filterbank is then constructed by combining all the logical filterbanks, each tuned to a different fundamental. A second, more economical way, to design the filterbank, is to design a set of detectors, closely and evenly spaced on a logarithmic frequency axis. Each detector consists of a pair of suitably designed bandpass filters. This may also be thought of as two filterbanks; The "a" filterbank, contains the lowest frequency filter from each detector, and the "b" filterbank contains the highest frequency filter from each detector.

If the latter type of filterbank is designed with certain "magic" spacings, then it will always be possible to select any detector in the filterbank, as the fundamental frequency detector, and then find other detectors that are accurately (but not exactly) tuned to most of the lower harmonics of the fundamental. The twelve notes per octave, used in piano tuning, is a familiar example of such a magic spacing.

However, the auditory system probably uses much finer spacing, to ensure that a detector can always be found very closely centered upon any desired fundamental frequency. In the computer simulations described below, the filterbanks were designed using the second technique, with 53 filters per octave. With this spacing, the first two dozen harmonics will be accurately tuned, for any selected fundamental detector.

Note that although the detectors are evenly spaced on a logarithmic frequency axis, this is not a constant-Q filterbank. The bandwidths of the detectors may be set to any desired function of frequency. An interesting feature of this type of filterbank is that each harmonic of every detector is always a fixed number of detectors above the fundamental detector; for example, the second harmonic is 53 detectors above the first. Harmonically tuned, logical filterbanks, as in figure 3, may be constructed by combining a selected subset of detectors from the physical filterbank. Because of the magically-spaced, logarithmic tuning, the pattern for selecting the subset of required detectors, is independent of the fundamental frequency. The auditory system need not exploit this property in order to function as described. But it does simply the bookkeeping within a computer simulation; only a single "comb" filter structure is required.

2.4 Pitch estimation

Let "$a_n(f)$" be the amplitude versus frequency response of the lowest frequency filter in the pair of filters tuned to the n'th harmonic, and "$b_n(f)$" the higher frequency filter response. It is possible to design these filters such that they obey an equation like (2), have almost any desired bandwidth, and to simultaneously control their shapes such that equation (3) will be true for any value of "n", and any value of the weight factors "w_n", which may even vary with time.

$$(3) \qquad \frac{a_n}{b_n} = \frac{w_1 a_1 + w_2 a_2 + w_3 a_3 + w_4 a_4}{w_1 b_1 + w_2 b_2 + w_3 b_3 + w_4 b_4}$$

In other words, all the harmonics (assuming, for the moment, that only one harmonic exists within each filter pair) will yield the same scale ratio "a/b", and any weighted combination of harmonics will also have the same scale ratio. Consequently, when these scale ratios are used in equation (2), with the appropriate change of variables, they will all yield the same frequency estimate; any weighted combination of harmonically related inputs can be simply made to yield an estimate of the fundamental (pitch) frequency. In our model of the auditory system, this capability is exploited as a form of graphic equalization, to perform frequency diversity processing. All true harmonics should yield the same scale ratio. Hence, all the filter pairs that do yield the same ratio may be combined to yield a single pitch estimate as in (3). However, any filter pair whose scale ratio differs from the others (due to noise, interference etc.) may have its weight factor reduced, to adaptively de-emphasize inharmonic contributions to the pitch estimate. Forming the weighted sums of the "A" and "B" filters, appearing in the numerator and denominator of equation 3, is indicated in Figure 1 by the blocks labeled "ΣA" and "ΣB".

Note that the auditory system may form these sums by directly adding the neural "spike-train" representations. Weighting may be accomplished by simply scaling the number of spikes per second before adding spike trains. (We shall discuss neural encoding effects further in section 3.3.) Note also that it does not matter whether or not the spike trains are phase-locked. The same simple ratio process will work in either case, because it is merely extracting the relative scale factor between the "A" and "B" channels. The overall scale of the data, determined by the automatic gain control (AGC), shown in figure 1, also does not matter. Since it will cancel out when the ratio is taken, the overall gain need not be communicated to the process in blocks "R".

This process, when implemented on all the logical filterbanks tuned to different fundamentals, contributes to the auditory system's ability to separate multiple input signals. At any one instant of time, signals that consist of harmonically related tones, will produce a strong output, only from those filterbanks with filters approximately matching the frequencies of one of the input sets of harmonics. A harmonic speech coder may be implemented by extracting the voltage or amplitude outputs from the filterbank "best tuned" to an estimated pitch, at any given instant of time. Computer simulations have verified that good quality speech, including unvoiced speech, can be reconstructed from these outputs.

To design filter pairs that exhibit the above properties, consider equation (4):

$$(4) \qquad a(f) = Ae^{-(\ln(f/f_a))^2/\Delta^2}$$
$$b(f) = Ae^{-(\ln(f/f_b))^2/\Delta^2}$$

This is just equation (1) with a change of variables from "f" to "ln(f)", and a change of notation such that the filter center frequencies are f_a and f_b, and the filter bandwidth is determined by the parameter Δ. With these changes of variables and notation, equation (2) can be transformed into equation (5):

$$(5) \qquad \ln(f) = \frac{\ln(f_a) + \ln(f_b)}{2} + \frac{\ln(a(f)/b(f))}{D}$$

where $D = \left(2\ln(f_a / f_b)\right)/\Delta^2$

The first term in equation (5) is simply the center frequency of a filter pair, on a logarithmic frequency axis. These center frequencies are the values that must be tuned to the harmonics in the filterbanks described above. Note that the bandwidth parameter Δ may be set to any value, without changing the value of ln(f), provided that the values of f_a and f_b are selected to keep the value of D constant. Consequently, even though the center frequencies of the detectors may be exactly evenly spaced (53 per octave in the simulations) on a logarithmic frequency axis, the bandwidth of the detectors need not be proportional to their center frequencies. By judicious choice of f_a and f_b, the bandwidths may be selected to yield almost any desired bandwidth versus frequency function.

The significance of this result is as follows. The instantaneous frequency of the n'th harmonic may be written as $nf_pM(t)$, where f_p is a constant value equal to the fundamental frequency (pitch) of a harmonically tuned filterbank, and M(t) is a time-varying modulation. The logarithm of this frequency is given by equation (6):

$$(6) \qquad \ln(f(t))=\ln(n)+\ln(f_p)+\ln(M(t))$$

But from equation (5), with detectors tuned to harmonics of f_p, we have equation (7):

$$(7) \qquad \ln(f(t))=\ln(n)+\ln(f_p)+\ln(a_n(f)/b_n(f))/D$$

Subtracting (6) from (7) yields (8):

$$(8) \qquad D\ln(M(t))= \ln(a_n(f)/b_n(f))$$

Since the left hand side of (8) is independent of the harmonic number, n, the right hand side must also be independent of n. Thus, for harmonically related inputs, this type of filterbank produces outputs with identical scale ratios, a/b, for all harmonics. Consequently, the output ratios can be averaged over the harmonics, as in equation (3). An averaged estimate of the pitch frequency may then be obtained by substituting the averaged scale ratio into equation (5), and using the center frequency of the fundamental detector for the first term in the equation. The instantaneous frequency of any given harmonic may be obtained using the center frequency of the detector actually containing the harmonic, as the first term in equation (5), rather than the center frequency of the fundamental detector. In this manner, the same basic process may report either an individual tone's actual frequency (when only one tone is present) or the frequency of the fundamental, when multiple, harmonically related tones are detected. Additional details of the design of such filterbanks can be found in (McEachern, 1994b, 1994c)

Quatieri et al. (1997) have discussed the performance of this type of FM detection process in the presence of noise, when the two inputs to the process have been amplitude (AM) detected. Here, we briefly note two other points to be considered. First, when the two inputs to the process have not been AM detected, as appears to be the case at low frequencies within the auditory system, the instantaneous input signal-to-noise ratio (SNR) is continuously changing. Obviously, for a halfwave rectified, sinusoidal input, there is no signal during each half cycle for which the neurons produce no spikes. During the half cycles for which spikes are produced, the SNR is highest at the peak of the input sinusoidal voltage. It is not known how the auditory system deals with this continuously varying SNR. It may simply perform a "hold maximum" to capture the frequency estimate at the highest input voltage peak. But it probably performs a more sophisticated weighted average. Second, one of the important characteristics of frequency diversity processing, as embodied in equation (3), is that signal averaging (summing the harmonics) occurs before the FM detection process. FM detection exhibits a "threshold effect", in which the output SNR (the quality of the frequency estimate) drops precipitously once the input SNR falls below a certain threshold. Averaging the harmonics prior to FM detection increases the input SNR, which effectively lowers the FM detection threshold. In other words, the frequency diversity processing not only enables the system to deal with harmonics that are constantly fading in and out, but it also enables the receiver to estimate the fundamental's instantaneous frequency at a lower SNR than would otherwise be possible.

2.5 Sorting and assembling acoustic jig-saw puzzles

Returning to the overall architecture of the auditory system, we are now in a position to discuss auditory scene analysis. The auditory system evolved to provide creatures with information about the other creatures they are likely to interact with in their environment. (A corollary to this statement is that one does not need information about entities with which one is unlikely to interact. Humans, for example, have little use for detectors responding to the ultrasonic hunting signals emitted by insectivorous bats. The same cannot be said with regards to the moths the bats are hunting.) Are these creatures friends, foes, or food? How many such creatures are within earshot? Where are they located? What type of environment exists between them and me? The typical acoustic scene contains multiple sources of sounds, and each sound is likely to arrive at the receiver along multiple paths. Moreover, many sounds of interest consist of discrete, harmonic tones. The harmonics from different sources may be interleaved in frequency so that it is not immediately obvious that some of the tones are in fact harmonics. That is, there may not be any immediately obvious correlation between the tones, which might serve to indicate that they came from the same source. In effect, the problem faced by the receiver is like having all the pieces from several jig-saw puzzles all jumbled together in one pile. The system must sort the pieces into an initially unknown number of puzzles, and then reassemble each puzzle into a coherent picture of its source.

However, the accurate frequency measurements of each tone, obtained from the frequency diversity processing described above, may reveal that some subsets of tones exhibit a precise harmonic relationship, that is maintained even as the tone frequencies are modulated. This correlation is indicative of the tones originating from the same source. Origination from the same direction also provides evidence for a common source, as does a correlation between the amplitude modulation present on different harmonics. The fundamental frequency (pitch) and the distribution of power among the harmonics (vocal tract resonant frequencies) provide information useful for identifying sound sources by correlating such extracted parameters against stored memories of previously extracted and identified parameter sets.

Amplitude and frequency modulation information is the primary information being extracted from the input, in order to characterize each identifiable signal source. The recovery of phase (timing) information is of a secondary nature, and is used mostly to characterize the multipath transmission channel, as opposed to characterizing the acoustic source. It has often been remarked in the literature that the ear is rather insensitive to phase. This observation has perplexed many investigators for two main reasons. First, nerve firings are often phase-locked. So phase information is obviously available. Second, because the instantaneous frequency of a signal is usually defined as the derivative of the instantaneous phase, many investigators have assumed that that must be how the auditory system estimates frequency. How can a system that is insensitive to phase, compute the derivative of the phase? As described above, the auditory system does not estimate instantaneous frequency from the phase, so the second problem does not exist. The resolution of the first problem is more interesting.

The reason the auditory system is relatively insensitive to phase is that phase modulation is one of the many types of possible signal modulations that the system is simply not looking for. The auditory system is a demodulator. It is a device that attempts to actively reject signal modulations that do not match what it is looking for. And it is not looking for phase modulations. Or more correctly, phase modulations are being primarily attributed to "channel encoding" rather than "source encoding". Phase information is being used to characterize the multipath environment traversed by a sound, rather than characterizing the modulation emitted by a source. Phase information may serve to indicate the direction of arrival of a sound, and to characterize multipath induced reverberations. But this is all information pertaining to the nature of the transmission channel rather than the source. Why should this be so?

It has been well established in the analysis of severe multipath environments, such as HF ionospheric channels, that it is much more difficult for a receiver to recover absolute (non differential) phase modulation information than either amplitude or frequency modulation information. The reason for this is that for these types of transmission channels, phase modulations induced by the channel may be hard to distinguish from phase modulations originating at the source. The same is not usually true of amplitude and frequency modulations. For example, amplitude modulation induced by the transmission channel (fading) may reduce the SNR to the point that the source modulation may no longer be recovered. But the two types of modulation are usually very different in appearance, which is to say that they can usually be readily distinguished from one another, by the receiver.

But when a system cannot reliably determine if a signal's phase modulation originated at the source, or in the transmission channel, it is safer to assume it originated in the channel. This phenomenon can be easily observed in harmonic speech coding based on the process shown in figure 1. The AM for individual speech harmonics was extracted, together with an averaged FM pitch. Phase was ignored. The speech was then synthesized by amplitude modulating a set of zero-phase harmonics, which were frequency modulated with the FM pitch signal. A number of informal listeners were then asked to describe the difference between the sound of the original and the synthesized speech. Typical responses were that it sounded "like the original with but with an echo" or "like someone speaking through a tube". In other words, the difference between the sounds was usually attributed to familiar transmission channel effects, and not to a difference originating at the speaker. It sounded like a normal voice being heard through a transmission channel that produced a vaguely familiar distortion; familiar enough that most people could assign a label to the distortion ("echo", "through a tube" etc.).

This is in sharp contrast to the labels listeners assign to speech synthesized in a manner that fails to preserve the AM and FM modulations, that the auditory system is looking for. Linear predicted coding of speech, for example, is usually labeled with terms like "artificial". Note also that even the unvoiced speech, which does not contain any harmonics, can be successfully reproduced as modulated harmonics. This fact has been noted previously (Kohata, 1999; McAulay and Quatieri, 1992; Macon and Clements, 1997). However, the explanation for this effect, as given here, differs somewhat from previous explanations. The reason given here is that the system is attempting to interpret all input signals as though they are "what it is primarily looking for", a set of harmonics. The fact that the input may differ significantly from the model the demodulator is trying to fit to the data, does not change the fact that the demodulator is trying to force such a fit. Hence, what we hear when we listen to any sound, is not the sound itself, but how our "harmonic" demodulator responds to the sound. Spectrally rippled noise and inharmonic chimes notes, have distinct pitches, because the harmonic demodulator is force fitting them into its preferred representation for its primary signal of interest. That is what sophisticated demodulators do. They attempt to "undistort", or adaptively equalize the input, by forcing all inputs into a best fit with an a priori known, desired output. We shall discuss this equalization process in greater detail in section 5. The reader who doubts that sensory systems attempt to fit a preferred model to their inputs, might reflect upon the system underlying color perception. Color (frequency) seems to be estimated from ratios of paired filter (cone cell) outputs, as in the technique described above. Colors mix the way they do (red+green=yellow), because the system assigns a single frequency (yellow), to any spectral power distribution that yields the same power ratio from a given pair of filters.

3.0 Evidence in support of the model

In a conventional spectrum analyzer, the frequency of a sinusoid may be estimated by simply noting the center frequency of the filter having the largest output response. When the analyzer has many, finely spaced, narrow, bandpass filters, this technique can result in fine (precise and accurate) frequency estimates. Such a system may even work when the bandpass filters are not equally spaced and have different bandwidths, provided that the wideband filters are highly overlapped. The "place" theories of pitch perception are based on this type of spectrum analysis. But this cannot be the process by which the auditory system obtains fine frequency estimates. It is well established that the auditory system's frequency estimates do not always correspond to the center frequency (place) of the filter with the largest output (Warren, 1999a), even when the input consists of a single sinusoid.

3.1 Pitch versus Place

In the model proposed here, precise (but not necessarily accurate) frequency estimates are derived from the relative scale factors (voltage, amplitude or power ratios) between outputs from adjacent filters in the cochlear filter bank. The behavior on this process is subtly different from the behavior of a place theory. These differences are significant, because they are, for the most part, the same differences that exist between place theory behavior and the behavior of the auditory system, as deduced from psychoacoustic testing. As an example of one such difference, consider a filterbank in which the two lowest frequency bandpass filters have center frequencies of 50 and 55 Hz. Since there is no "place" or filter below 50 Hz, a 20 Hz input signal will produce its greatest output response from the lowest filter, at 50 Hz, far above the correct input frequency. But the process described above would correctly deduce that the input must have been at a frequency of 20 Hz, far below the center frequency of any filters within the filterbank. Because the 20 Hz tone is detected far from the peak of any filters' frequency response, it is highly attenuated. Such low frequency tones would thus have a much higher auditory threshold, than tones at higher frequencies. This is consistent with the known threshold of hearing versus frequency in humans. In the model proposed here, the sharp loss of sensitive to tones at the highest and lowest frequencies that can be heard, is caused not so much by the low sensitivity of filters tuned to those frequencies, but by the absence of any such filters. The highest and lowest tones are being detected on the "skirts" of filters well below and above the frequencies of such tones, respectively.

3.2 Pitch biases and shifts

Another interesting phenomenon peculiar to the process of deriving FM from AM, is the existence of systematic biases in the FM estimate, when an interfering signal is present, together with a tone, within the pair of filters used to derive the FM estimate. Since the interference may be stronger in one filter than in the other, its presence may change the ratio of the amplitude outputs from the filters. That, in turn, will change the frequency estimated from that ratio. Indeed, any process that changes the amplitude ratio will cause a frequency shift. For example, an automatic gain control process might change the gain within one filter slightly more than the other. As a result, changes in signal amplitude may result in slight changes in the estimated frequency. Such amplitude dependent pitch shifts are well documented within the auditory system (Zwicker and Fastl, 1999). But more conventional forms of FM estimation do not exhibit such biases and shifts.

It recent decades, FM estimation techniques have been dominated by techniques that exploit the fact that instantaneous frequency may be obtained as the derivative of the instantaneous phase. Such phase-based FM estimation techniques exhibit zero-bias, in the presence on small interferers. Furthermore, the phase-based FM estimates are insensitive to input amplitude. Indeed, these characteristics are the primary reason why the phase-based techniques have become popular. They are good FM estimators. But for that very reason, they are poor auditory models.

When the input consists of a signal other than a single sinusoid, even more peculiar effects may be observed. For example, an inharmonic complex of tones, may be perceived by a listener as having a pitch that does not correspond to the frequency of any of the tones in the input, and also does not correspond to the difference in frequency between any of those tones (Warren, 1999a). But the model presented here has no difficulty in reproducing such behaviors. A pitch estimate, derived from the averaged amplitude or voltage ratio, given in equation (3), behaves in this fashion.

Unfortunately, it is difficult to make definitive comparisons between the performance of the model and that of the auditory system. The model's pitch estimates depend on a large number of parameters, which are not well known. These include the exact shape and center frequency of each filter, the time-varying weights used to combine multiple harmonics, and frequency, amplitude and time-dependent gain variations across the filterbank. On the other hand, there are probably enough adjustable parameters to make the model fit a very wide variety of psychoacoustic effects, assuming that it is fundamentally correct. Rather than belaboring such details, we shall instead examine some larger issues; phase effects and speech analysis and synthesis.

3.3 Phase Effects

It has often been said that the auditory system is rather insensitive to phase. At first glance, this seems rather surprising, given that auditory nerve firings are highly phase sensitive, as evidenced by the phase-locking behavior noted earlier. But upon further consideration, there is no mystery unless we misconstrue sensory perception as being primarily concerned with analyzing everything that is "out there", rather than merely seeking to determine if what is out there matches what the system is looking for. In the model presented here, the primary form of information being extracted from auditory signals is amplitude information, which is subsequently transduced into frequency information. Since this combined amplitude and frequency information constitutes the majority of information being extracted, phase plays a limited role in auditory perception.

In the case of speech perception, it is now well established that phase information contributes almost nothing to the intelligibility of speech, and only a little to the sound of the speech. Harmonic speech coders, such as those described here, can reproduce good quality speech without encoding any information about the harmonics' relative phases. However, this does not mean that the decoder ignores all phase effects; rather, it means that phase effects sufficient to reproduce good quality speech can either be supplied to the decoder as a priori knowledge, or can be derived from the encoded harmonic amplitudes. We shall discuss these coding techniques in section 3.4.

Two phenomena, that do depend upon a signal's phase, and that have attracted much attention in the literature, are aural harmonics (Plomp, 1967) and binaural beats. To address the origins of these and other phase related phenomena, we must first examine the nature of auditory neural encoding. The details of neural encoding are complex, with various saturation and nonlinear effects occurring. But to a first approximation, the temporal density of nerve firings (spikes/second) is proportional to the halfwave rectified, instantaneous "voltage" of the input waveform. This can be observed in period histograms, such as those studied by Rose et al (1971) and Teich et al (1993). In the case of a single, pure tone input, Teich et al model period histograms as a convolution of three functions; (1) a periodic train of delta functions, with period equal to the tone period (2) a pulse, corresponding to a half-cycle of the input sinusoid and (3) a pulse whose shape and width corresponds to the distribution of errors in nerve-spike measurement times.

As noted earlier, the frequency estimation technique proposed here, is based on comparing the relative scale of the outputs from pairs of suitably shaped, bandpass filters. These scaling factors may be preserved regardless of whether of not the outputs are amplitude detected, half-wave rectified, phase-locked, or encoded as neural spikes. The same cannot be said about the preservation of phase information. Amplitude (envelope) detection will clearly remove phase information. This has a bearing on the details of the ratioing process labeled "R" in figure 1. Clearly, the auditory system extracts amplitude information. But does it perform this amplitude detection before the ratioing process or after? Most investigators that have considered such frequency estimation techniques, have assumed that amplitude detection occurs first, and the frequency estimate is subsequently derived from the amplitude ratio. But the existence of phase effects, such as binaural beats, suggests that phase information is being preserved until after frequency diversity processing occurs. That in turn suggests that amplitude detection may occur after the process "R". But apart from having to contend with a time-varying, signal-to-noise ratio (there is no signal during half of the half-wave rectified neural signaling) the ratioing process can still yield precise frequency estimates. Note also that the process "R" occurs in two locations within figure 1; (1) for estimating the frequency of each individual harmonic and (2) for estimating the frequency of the weighted sum of harmonics. Hence, the issue of where amplitude detection occurs, before or after "R", must be considered separately for each of the two locations of "R".

3.4 Harmonic speech coding

To investigate the proposed model of the auditory system, a computer simulation of the process was developed and used to process speech. There were two reasons for being particularly interested in speech processing. First, the model is a demodulator and demodulators only pass a selected portion of an input signal's total information content through to its output. Consequently, if good quality speech cannot be perceived from the demodulator outputs, then the demodulation process cannot be a good model of the auditory system. Second, speech coding is a topic of considerable interest. A demodulator that extracts the same information that is extracted by the auditory system and that rejects the same information that it rejects, is a good candidate for an efficient, low bit-rate, speech coder.

A filter bank was constructed with 350 detectors, with 53 detectors per octave, spanning the audio range from approximately 60 Hz up to the Nyquist frequency of the 16 kHz sampling frequency. Each detector consisted of a pair of filters, as described previously. The bandwidth of the filters was selected to approximate the bandwidth of the cochlear filters, and the shape of the filters was designed such that the ratio of the logarithm of the outputs from each pair yielded an estimate of the logarithm of the instantaneous frequency of any tone within the detector's bandwidth. The filters were designed to have linear phase FIR responses, such that they all exhibited the same signal delay. Rather than implementing the filters as real bandpass filters, they were implemented by complex basebanding of the signal, followed by lowpass filtering. This was done for three reasons; it made it possible to implement the filterbank via an efficient FFT algorithm (Appendix 1), it enabled a simple computation of both the amplitude and phase of each filter output, and it enabled amplitude detection to be performed without the need for applying a lowpass filter, as would be required in a rectifying detector. The latter point avoids the problem of having to specify the poorly known lowpass filter characteristics, appropriate for each channel; the AM bandwidth is determined solely by the channel's bandpass filter response.

For simplicity and also for low bit-rate speech coding considerations, it was decided to forego the simulation of the half-wave rectification and neural encoding processes, to ignore the phase information, and to only use the amplitudes detected at the filter outputs to perform subsequent frequency estimation and frequency diversity processing. Filterbank outputs are illustrated in figure 4.

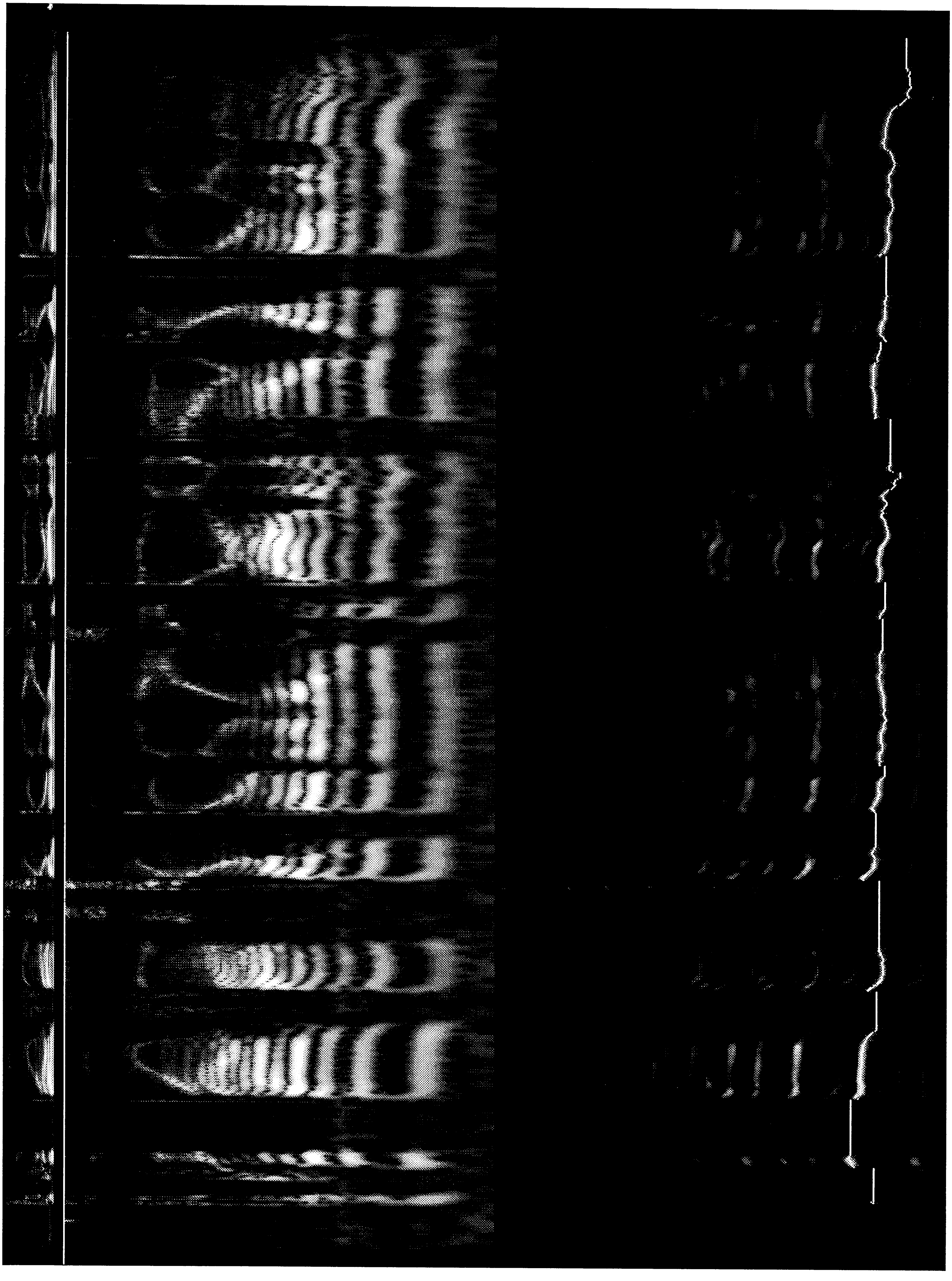

FIGURE 4: Three spectrograms: (top) amplitude of 42 selected harmonics, (middle) amplitude of 350 filter channels, 53 filters/octave, with "cochlear" bandwidths, (bottom) amplitude of 350 weighted and summed channels, with overlaid "best pitch" estimate

The figure depicts three "spectrograms" of approximately 5 seconds of speech, consisting of the author reciting "The quick brown fox jumped over the lazy dog and the cow jumped over the moon." The middle spectrogram depicts the amplitude output from the "a" filter (the lower frequency filter) from each of the 350 detector pairs. The frequency scale is logarithmic, due to the logarithmic spacing of the detectors. Speech harmonics are readily apparent, with a pitch frequency on the order of 110 Hz. The lower spectrogram depicts a weighted sum of the harmonic amplitudes (equation 3 numerator). A peak picker selects the detector with the largest weighted output (usually the fundamental), and traces a bright line through the selected peaks. When the peak power falls below a fixed threshold, the peak picker simply holds the last value exceeding the threshold, resulting in the straight line segments during unvoiced speech segments. Pitch estimates are then derived from the ratio of the paired filter outputs from the detector pair corresponding to the picked power maximum (usually the detector whose center frequency is nearest to the pitch frequency). The time-varying pitch frequency is then used to compute harmonic frequencies, and the amplitudes of the first 42 harmonics are estimated as being equal to the amplitude outputs from the detectors whose center frequencies are closest to the computed harmonic frequencies. These harmonic amplitudes are shown in the small, uppermost spectrogram. The speech formants are readily apparent.

Even using the FFT channelizer algorithm, this speech analysis is computationally expensive; there are 700 unique filter outputs to compute. But subsequent speech synthesis and/or modification is simple; 42 harmonic sinusoids are amplitude and frequency modulated, using the amplitudes and pitch frequency extracted (and then modified, if desired, to shift the pitch or alter the effective length of the vocal tract) from the filterbank analysis, and then summed together to recreate the speech signal. In the synthesis performed here, the initial phase for each harmonic was set to zero, and a small, random phase jitter was also added to some harmonics. The resulting synthesized speech sounded like the original speech, played through a channel with a very slight reverberant character. This reverberant character could be greatly exaggerated, by simply lowpass filtering each harmonic's time-varying amplitude, prior to synthesis. (The desirability of AM detection without lowpass filtering was noted previously.)

The reason these reverberation effects occur seems to be due to the manner in which the brain interprets "pulse stretching" as being correlated with multipath interference. To see this, consider two identical pulse trains, arriving at a receiver with a relative delay somewhat less than the pulse width. The sum of the two is a single pulse train, with the same pulse repetition rate, but with longer pulse durations. Hence, received, AM detected, multipath signals will have stretched out pulses. The lowpass filtering, noted above, stretches out the pitch pulses occurring in speech, in a manner similar to that caused by multipath. Since multipath distortion is a familiar acoustic effect, but the lowpass filtering is not, the brain misinterprets the similar effect has having the more familiar cause; stretched pulses are caused by multipath induced reverberation.

Likewise, the vocal tract slightly delays some harmonics relative to others, changing the shape of pitch pulses. By neglecting these delays (phase shifts) in the synthesized speech, the sum of the harmonics will not precisely reproduce the original pitch pulse shapes. The resulting slight distortion is perceived as being caused by multipath reverberation. Other investigators (Kohata, 1999) have noted that this distortion can be reduced, without having to encode any measured phase information, by setting the initial harmonic phases to non zero values. Replacing the zero-phase initialization with either a minimum phase (derived from the AM) or a "Rosenburg pulse" phase, resulted in improved speech quality. Both of these approaches simply embed a priori information (about how the vocal tract typically distorts a pulse propagating through it) into the synthesizer, thereby enabling it to do a better job of reconstructing pitch pulses without requiring actual phase measurements. Another source of pulse stretching, which contributes to the perceived distortion, is the double filter effect. Whenever a pulse is passed through a bandlimited filter, it will be stretched. Speech is ordinarily passed through only one such filter, at the receiver, the cochlear filter. But the synthesized speech is passed through two; the analysis filterbank, prior to synthesis and then the cochlear filterbank when a listener hears the synthesized speech. Hence it is stretched twice.

In addition to using a non zero phase initialization, it is desirable to avoid exact harmonic relations between the summed sinusoids used to synthesize the speech. Quality can be improved by adding a slight phase jitter to the summed sinusoids (Macon and Clements, 1997). But it can also be improved by removing every other upper harmonic from the synthesized speech. The former technique impairs the ear's ability to measure a harmonic's instantaneous frequency. The latter improves it.

To elucidate why this works, we shall consider several variations for synthesizing speech. First, consider synthesizing speech from each harmonics' measured AM and FM. The filterbank is able to resolve the lower harmonics, and thus make precise measurements of their AM and FM. But the upper harmonics are not resolved. Due to their wide bandwidths, filters tuned to upper harmonics suffer from interference from adjacent harmonics, so their AM and FM measurements are distorted. When these distorted measurements are used to synthesize speech, the distortion is evident in the sound of the speech. This distortion can be alleviated by performing a sample-by-sample test to determine if a harmonic's instantaneous frequency differs from its harmonic number multiplied by the estimated pitch, and then substituting the latter for the former if the difference exceeds one or two percent. This eliminates the worst FM distortion, to which the ear is rather sensitive. But if one replaces all the measured FM values with multiples of the pitch, rather than just the worst ones, the synthesized speech typically has a buzzy sound.

A similar buzzy sound occurs when synthesizing purely artificial tones with many harmonics. Consider the case in which a listener hears a signal synthesized with several perfectly harmonic tones, that are frequency modulated with a sinusoidal modulation of 0.5 Hertz, and a frequency deviation of the fundamental equal to about ten percent of the fundamental's center frequency of 100 Hertz. As long as there are less than about a dozen harmonics in the signal, the timbre of the sound is not buzzy. But when several dozen such harmonics are synthesized, the listener typically perceives a sound with a buzzy timbre.

In the model presented in figure 1, presumed harmonics, from each separate signal source, are summed together via a weighted summation with time-varying weights. The value of each weight may be made to depend upon several factors, one of which may be the difference between the fundamental frequency estimate derived from each individual harmonic and the averaged fundamental frequency derived from all the harmonics. Tones exhibiting much larger than average frequency differences, may be deemphasized, within a single, reconstructed output (and thereby made to contribute to a different output), by reducing their weights, on the assumption that they may in fact not actually be harmonics. Consequently, which tones get summed together as though they are all harmonics, emanating from a single source, depends on the distribution of the fundamental frequency estimates from the individual tones.

In the case of the speech synthesized with perfect harmonics, the auditory system seems to combine the low harmonics into one perceived entity, the speech, and some of the upper harmonics into a second perceived signal, the buzz. It seems as though several of the upper harmonics "match" each other well enough (perhaps by matching a multiple of the fundamental, as is evident in the lower part of figure 4) that they are combined into a single, buzz perception, but that perception does not match the fundamental frequency of the lower harmonics well enough for the system to combine the two perceptions.

By manipulating which tones will yield precise frequency estimates, one may, to some degree, influence which tones will be perceived by a listener as emanating from a common source. Phase jittering the upper harmonics more than the lower ones seems to prevent the upper harmonics from matching one another enough to be combined into a separate buzz perception. Eliminating every other upper harmonic reduces interference from adjacent harmonics, enabling more precise frequency estimates, which then match the lower harmonics well enough to enable the system to combine them all into a single perception.

Phase jittering the upper harmonics seems to prevent any of them from matching amongst themselves well enough to be perceived as a separate entity from the lower harmonics. Hence, they are combined with the lower harmonics into a single signal, but given such low weights that they are almost deemphasized out of existence. Using an adaptive threshold, which depends on the error distributions (deviations from the average fundamental) may also explain how the auditory system combines inharmonic tones, such as chime tones, into a single perception. In effect, when few tones accurately align as harmonics, the system broadens its "search window" around the "best fit" harmonic tone frequencies.

A spectrogram of speech synthesized with only every other upper harmonic, looks obviously artificial. But it does not sound artificial. Since the upper harmonics are now more widely spaced in frequency, there is less mutual interference caused by having multiple harmonics within the bandwidth of individual detectors. Hence, the auditory system can measure the upper harmonic frequencies more accurately and thus conclude that they do in fact match the lower harmonics, so that they are all combined into a single perception. Since such effects depend on whether or not upper harmonics are isolated within individual detectors, they also depend on the fundamental pitch frequency, since a higher pitch results in more widely separated harmonics.

Deleting every other upper harmonic from the synthesized output obviously alters the spectral power distribution of the speech. But the auditory system seems to be much less sensitive to this amplitude distortion than to frequency distortion. (Note also that substituting a multiple of the pitch for the actual distorted FM measurement, may "correct" a tone's frequency distortion, but not its AM distortion.) At first, this might seem rather odd, given that the frequency is being derived from the amplitude. But it must be remembered that a demodulator is not attempting to characterize its input, it is merely seeking to find what it is looking for, within that input. Techniques such as adding phase jitter and deleting every other harmonic make it easier for the demodulator to separate out components that match what it is looking for, from those that do not. The former makes sure poorly resolved components do not mismatch. The later improves the resolution of those same components so that they correctly match. Furthermore, in the model proposed here, frequency selective fading and frequency diversity processing, often produce nulls throughout the spectrum. So adding a few more judiciously located nulls, may not produce an unnatural input signal, as far as the demodulator is concerned.

Several speech files, and their sonograms, are described in Appendix 2. These illustrate some of the effects noted above, for harmonically synthesized speech, as well as some of the phase vocoder speech described below.

3.5 Relationship to phase and channel vocoders

There are obvious similarities between the harmonic speech coder described above, and other voice coders, such as phase and channel vocoders. It is therefore instructive to examine why the quality of the speech reproduced by such methods is so much poorer than the harmonic coder. The name "phase vocoder" is a misnomer. A phase vocoder is basically just an AM and FM detected filterbank, with the FM being derived from the derivative of each channel's instantaneous phase. Apart from minor effects noted previously, deriving the FM from a phase derivative rather than an amplitude ratio makes little difference to the performance of the coder. However, using filterbanks with inappropriate bandwidths, and no frequency diversity processing, causes significant degradation in the synthesized speech; the speech is perfectly intelligible, but it sounds quite artificial.

Phase vocoders typically employ filterbanks with only a small number of relatively wide bandwidth filters. As a result, the low harmonics occurring in speech are not isolated into individual channels, so the measured AM and FM on the channel outputs does not correspond to the modulation of any harmonic. Worse still, because of the "capture effect" of an FM demodulator, several channels within the filterbank may respond, at least partially, to the same input harmonics, particularly if one is much stronger then its neighbors, as might be the case for a harmonic close to a formant. When such similar responding channels are resynthesized, they create similar, but measurably different, copies of some signal components. Like multipath, these copies interfere with one another and result in easily detectable distortion.

The problem caused by using filters with bandwidths wider than one harmonic, is that the resulting modulation (such as beating AM caused by two harmonics) is much more erratic than that of a single harmonic. For example, the measured modulation will change somewhat if the center frequencies of the filters are shifted slightly, or if their bandwidths are slightly altered. Thus one has the highly undesirable result that slight modifications to the measuring apparatus result in notably different measurements, even when the input has not changed. Such behavior is quite different from the harmonic coder. The large number of overlapping filters, combined with frequency diversity processing, enables the latter to carefully pick and choose only the "best measurement" channels for use in synthesizing the speech. Hence, it ensures that it has one, and only one good copy of each harmonic, rather than some peculiar combination of harmonics, that don't quite add up to the real thing.

Channel vocoders are similar to phase vocoders, but force the synthesized output to exhibit harmonics, by filtering a pitch pulse generator's output, while synthesizing the speech. But like the phase vocoder, it does not ensure that AM measurements correspond to individual harmonics, so it cannot precisely reproduce the correct amplitude modulation on the reconstructed harmonics. Both phase and channel vocoders typically reproduce speech by remodulating all their measurement channels, rather than picking and choosing only the appropriate ones. Because of this, they are damned if they do have the correct filter bandwidths, and damned if they do not. If they do not, then the measurements will not correspond to individual harmonics. If they do, they will have several channels partially responding to one individual harmonic, and will consequently reproduce multiple interfering copies of the same.

Since the filterbank simulator was constructed to enable easy modification of the filterbank's characteristics, it was a simple matter to employ it to construct a variety of phase vocoders with differing characteristics. Phase vocoders were constructed with one, three, four and twelve constant-Q filters per octave, and also with a mel-scale type structure similar to the cochlea. All of them produced intelligible speech, and the quality improved as the number of filters employed increased and the structure more closely matched that of the cochlea. But the quality never approached that obtained with the harmonic coder, and the sound of the synthesized speech could be changed by minor adjusts in the filterbank structure, such as slight shifts in the filter center frequencies.

4.0 Evolution and biological plausibility

The previous sections provided evidence that the proposed auditory model exhibits many behaviors similar to the known behavior of the auditory system. It was also demonstrated that the information being extracted from complex input signals, such as speech, preserves the information being extracted by the auditory system; good quality speech can be reconstructed from the extracted modulation information. Here we consider the biological plausibility of the model, and its probable evolution from simpler, more primitive systems.

A likely sequence of steps in the evolution of the auditory system is:

1) Development of a single, wideband acoustic power detector, based on cells responding primarily to the power within a relatively wide frequency band. The system estimates amplitude (power), but not frequency or phase. Simple "direction finding" is based on time-differences-of-arrival of AM envelope features, from two such detectors.
2) Development of a power spectrum analyzer, based on the differentiation of the wideband detector into numerous, narrowband detectors, tuned to different frequencies.
3) Development of a threat warning receiver, with filter bandwidths matched to the bandwidths of signals of interest (cochlear filterbank). FM detection (based on AM ratios) retrofitted to the AM detected spectrum analyzer.
4) Upgrades to the threat warning receiver exploit the FM estimation capability to perform frequency diversity processing and acoustic source "reconstruction".
5) The threat warning receiver is expropriated for use as a communications receiver; speech development.

When did phase processing enter into this picture? The evidence suggests that effects such as the "phase locking" of neural firing patterns has existed from a very early time. But the subsequent processing has never figured out how to effectively exploit the phase information encoded into neural firing patterns, except in cases in which a difference in phase results in time shifts in the AM envelopes from the various receiver channels. For example, half-wave rectification enables a shift in a tone's phase to be detected as a temporal shift in the lobe structure of the AM envelope of the half-wave rectified waveform. Similarly, the preservation of phase information from signals from both ears, enables binaural beats to be detected by the time-varying amplitude of the combined signals. Phase shifts seem to be detected indirectly, as time shifts in a component's AM envelope. This seems to be the case even in situations in which psychoacoustic testing has confirmed that listeners can detect the difference between two signals that have identical power spectrums, but differ in phase.

At first glance, one might suppose that identical power spectrums imply that the AM envelopes produced by the cochlear filterbank must also be identical. But such is not the case. If a Fourier spectrum analyzer is viewed as a filterbank, then all the filters have identical impulse responses (merely being tuned to different frequencies), that integrate the input over an infinite duration. But each of the cochlear filters integrates its input over a different, finite duration. Consequently, the behavior of the AM detected cochlear outputs is fundamentally different from a Fourier analyzer. This can be seen in the spectrogram in figure 4. At low frequencies, the cochlear filters integrate over many pitch pulses, so the output harmonic amplitudes exhibit no pulse structure. But at high frequencies, the duration of a filter's impulse response is less than the pitch pulse period. Hence, fine striations in the harmonic amplitudes are visible, at the pitch period. The short integration time of the filters combined with half-wave rectification enables the system to encode considerable phase information into the AM envelope to the filterbank's outputs. But as discussed previously, much of this phase information pertains to the communications channel linking the source to the receiver rather than the source itself. Consequently, it is of comparatively little use to a threat warning receiver attempting to characterize a sound source. Therein lies the reason that natural selection has not produced a receiver more attuned to a signal's phase.

4.1 From spectrum analyzer to threat warning receiver

It is well documented that the auditory system responds directly to an input signal's amplitude variations. It is argued here, that most frequency and phase information derived, from a signal, is also detected via amplitude variations. Amplitude variations from one filter to the next can easily be converted into precise frequency estimates, and time shifts in half-wave rectified AM envelopes convey phase information. Thus, it would seem probable, that the present auditory system evolved from a system that was primarily an amplitude detector.

An AM detecting filterbank probably first functioned as a simple power spectrum analyzer. It evolved into a sophisticated threat-warning receiver. As a simple detector of acoustic power from signals of interest (SOI), an auditory filterbank, consisting of several narrowband AM detectors, probably outperformed either a single wideband AM detector that spanned the same bandwidth, or a smaller (narrower total bandwidth) filterbank, such as a single narrowband detector. For the ancestors of humans, the signals of interest were signals emanating from creatures that interacted with those ancestors; creatures they ate, those that ate them, and those they mated with, for example. Many such creatures employ vibratory mechanisms for producing sounds. Such mechanisms often produce narrowband, tonal type signals. Consequently, a receiving system with narrow bandwidths, matched to the bandwidths of those tones, would provide better signal-to-noise and signal-to-interference ratios than wider bandwidth detectors, and better probabilities of detection than narrower bandwidth filterbanks, whose total bandwidth was less than the SOI bandwidth. These advantages alone may have provided the natural selection forces that initially drove the development of AM detecting filterbanks, with individual filter bandwidths approximately equal to the bandwidths of the tones occurring within the SOI.

It is well known that the cochlear filterbank has filter bandwidths that remain approximately constant up to center frequencies of about 500-1000 Hz. At higher center frequencies, the filter bandwidths become proportional to the center frequency. It is interesting to note that such a bandwidth versus frequency relationship could result simply from AM detectors evolving as described above. Optimal power detection in noise demands that filter bandwidths match the bandwidths of the SOI. As noted previously, the SOI are likely to have been the tones, particularly harmonics, produced by other creatures. Such harmonic tones are often frequency modulated. According to Carson's Rule, the bandwidths of these tones are approximately constant for the lower harmonics, and proportional to the harmonic number (center frequency) for higher harmonics. Carson's rule-of-thumb relationship states that the bandwidth of a frequency modulated signal is given by 2D+2M, where D is the frequency deviation and M is the modulation bandwidth. For frequency modulated harmonics, D is proportional to harmonic number and M is the same for all harmonics. Hence, for low harmonics, the first term, D is small, and the sum approaches the constant value of 2M. But for high harmonics, the sum is dominated by the 2D term, so the bandwidth becomes proportional to center frequency.

Thus, the bandwidths of cochlear filters are matched to the bandwidths of FM harmonics. We do not believe this to be a coincidence; the cochlea appears to be specially designed for harmonic power detection. It appears as though the cochlea's peculiar bandwidth versus frequency characteristic may have resulted from the natural selection of filter bandwidths optimized for simple harmonic power detection in noise. What shape would optimal filters have? Since the problem at hand is to obtain a good, bandlimited power estimate as quickly as possible, to provide threat warning, filters that have a short temporal response would be desirable. Since Gaussian filters have the minimum possible time-bandwidth product, they would be a good candidate for meeting these requirements. As noted previously, Gaussian filters have a second optimal property; a pair of them, tuned to different center frequencies, can be used to construct a perfectly linear FM demodulator. Furthermore, as was also shown above, if the bandwidths of the two filters are not equal, the demodulator may compute a non linear function of frequency. In particular, bandwidths like those found in the cochlea may result in a non linear frequency function matching a mel-scale function, like that known to be used by the auditory system.

A plausible evolutionary sequence is thus: a system evolved to detect acoustic power. Over time, the detector bandwidths became optimized to detect the power in frequency modulated harmonics (the SOI). That resulted in the observed cochlear bandwidth versus frequency curve. An FM detection system was then retrofitted to the power detection system, by simply exploiting pairs of existing power detectors. That resulted in the mel-scale encoding of frequency information. The recombining of multiple harmonics into a single acoustic entity was then accomplished by frequency diversity processing, that merely forms simple, weighted combinations of the outputs from the existing filterbank. The model is simple, plausible from an evolutionary point-of-view, and exhibits many of the peculiar performance characteristics known to occur within the auditory system.

5.0 Speech communication – the form and function of speech structures

In the scenario described above, the auditory system's front-end, consisting of an AM and FM detecting filterbank, evolved as a threat-warning receiver, long before the development of spoken language. Furthermore, there is little evidence to suggest that the characteristics of this filterbank, such as filter bandwidths and shapes, have changed significantly since the development of speech. Rather, it seems as though the characteristics of speech have evolved in order to best accommodate the capabilities and limitations of this preexisting receiver architecture, and the nature of the impairments found on typical speech communications channels. The main impairment is multipath interference.

The problem that multipath represents to a communications receiver is depicted in figure 5. The top line shows the original message being transmitted. The middle line illustrates two copies of this message arriving at a receiver, with a relative travel-time delay that is a small fraction of the length (duration) of one of the transmitted symbols. The bottom line depicts two copies arriving with a relative delay that is large compared to the symbol duration. Because of the multipath, the received signal is a distorted copy of the transmitted signal. In general, when the relative delay is large compared to the symbol duration, the resulting "intersymbol interference" makes recovering the message more difficult then when the delay is small .

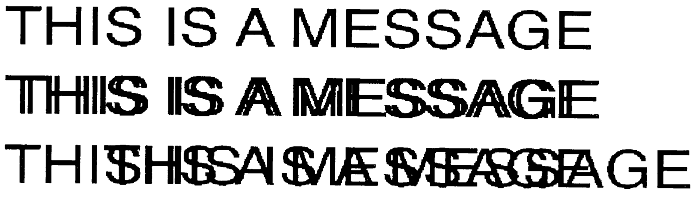

FIGURE 5: Illustration of multipath induced, intersymbol interference: (top) original message, (middle) two copies of message received with a short delay, (bottom) two copies of message received with a long delay

The simplest technique for alleviating intersysmbol interference, is to employ symbol durations that are long, compared to typical multipath delays. This occurs in speech, and was also employed in most early modem designs. Since the speed of sound in air is approximately one foot/millisecond, each additional foot of path length difference adds about 1ms relative delay between the multipaths. In typical conversational speech environments, path length differences are usually on the order of a few feet. So the multipath delays are on the order of a few milliseconds. This is short compared to the duration of speech symbols (Syllables appear to be the shortest speech segments that are directly detected by the auditory system. Shorter units, like phonemes, are inferred from previously perceived syllables and words. (Warren, 1999b)).

Another obvious characteristic of speech is that it interleaves vowel and consonant sounds on a regular basis. Most of the information content in speech resides in the latter. But the former are easier to detect (particularly when the receiver was specially designed for detecting harmonics!). This too is analogous to techniques used in modems designed to operate in high multipath environments, with symbol durations that are short compared to the relative delays. Such modems periodically interleave easily detected, low information content symbols, with harder to detect, high information content symbols. The former are known a priori by the receiver, making them easy to detect (by matched filters, for example). They are used to estimate the time-varying channel impulse response (multipath characteristics), which in turn is used to adaptively equalize (reduce the intersymbol interference on) the information carrying symbols. An analogous interference reduction process seems to occur in speech, as a result of the frequency diversity processing.

The periodic transmission of vowels, rich in harmonics, enables the frequency diversity processing to identify and deweight any frequency bands that contain power, but does not match the correct harmonic frequency. Such bands probably contain noise or interference. Harmonic vowel detection thus allows the receiver to act as a graphic equalizer, constantly readjusting the gain of the summed filter outputs, to reduce the interference on the next information carrying consonant. The speech equalizer is attempting to alleviate multisource interference, whereas the modem equalizer alleviates multipath interference.

It is important to note that speech is making use of several distinct techniques to combat interference. First, frequency diversity processing enables the system to deal with multipath induced, frequency selective fading on vowel harmonics. The use of long symbol durations reduces multipath induced intersymbol interference. These techniques, in turn, enable the reliable detection of vowels. That, in turn, enables the system to use the detected vowels in a manner similar to the use of a priori known modem symbols (channel probes); it drives an adaptive equalization process designed to improve the reception of the harder to detect symbols; consonants, in the case of speech, that convey most of the information content.

These various interference reduction mechanisms contribute substantially to the "cocktail party effect", in which a listener can pick out a voice in a crowd. But like the modem case, their effect is negligible when there is a clean signal. Unvoiced speech, including whispering, which has no harmonics to "lock onto", can be readily understood in clean environments, just as modem signals do not need interleaved channel probes when there is little channel distortion or interference. The information contained within a whisper (or any other sound) is being conveyed predominately by the amplitude modulations on the filterbank outputs. All the FM and frequency diversity processing is there, primarily to determine how many acoustic sources exist, and which filterbank outputs originated with which source; it is largely superfluous when there is only one source.

Still other techniques, all of which have analogies in digital data transmission, are used to further improve the reliability of speech detection. These are summarized in table 1.

Gray coding is a technique in which similar symbols are used to symbolize similar information. Hence, if a small amount of distortion causes the detector to misidentify a transmitted symbol, it is likely to identify it, incorrectly, as a symbol that appears similar to the correct one. As long as the misidentified symbol encodes information that is similar to that encoded by the correct symbol, the intended message may not be totally lost. In speech, for example, misidentifying the symbol "hits" as "hit", will probably result in a correctable error. It is significant that spoken languages commonly modify word endings, in order to construct "Gray coded" vocabularies. Multipath more frequently garbles word endings than beginnings. Placing the root word at the beginning, where it is least likely to be garbled, is probably the result of natural selection, between competing vocabularies, favoring the information coding that is most likely to be understood in multipath environments. It is interesting to note, in this context, that languages that do not have a long evolutionary history, such as Creoles, also do not use Gray coding as extensively as those that do.

Beyond employing techniques designed to improve the probability of correctly identifying individual symbols, it is possible to exploit additional techniques, which improve the probability of correctly identifying sequences of symbols. In this regard, grammar and punctuation seem to be analogous to techniques such as Trellis Coded Modulation and Framing employed in data transmission. Such techniques enable the receiver to detect, and frequently correct, additional garbling in a message, by noting that received sequences of symbols violate a priori known restrictions on the allowable sequences of transmitted symbols. Also, by imbedding special "control symbols" into a transmitted sequence, such as a voice inflection used to indicate a question, the receiver may deduce that special message handling (giving an answer) may be required, even before it decodes the underlying message. Requesting retransmission of garbled messages is also common to both speech and data communications.

CHARACTERISTIC	SPEECH	HF MODEM
Low symbol rate used to mitigate multipath and intersymbol interference	Syllable duration (100 ms) is much greater than typical multipath delays (>10 ms)	Morse code, low baudrate FSK and multitone modems use symbol durations much greater than the 1-10 ms multipath delay
Use of frequency diversity signaling, which transmits the same information on two or more different carrier frequencies, to alleviate frequency selective fading	Vowel harmonics each encode the same modulation of the logarithm of the instantaneous fundamental frequency, F(t): Log[nF(t)] = log[n] + log[F(t)]	Frequency shift keyed (FSK) and multitone modems transmit two or more copies of the same information, at different frequencies; FSK employs anti-correlated on off keyed signals
Use of adaptive equalization based on alternating transmission of easily detected "probes" with harder to detect information	Alternating transmission of vowels (with narrow bandwidth harmonics) and consonants with wide bandwidths	Alternating transmission of a priori known channel probes and unknown information carrying symbols, at high baud rates
Similar symbols used to convey similar information	Modification of a root word to form related words; Singular/plural (dog/dogs) Verb tense (hit, hits, hitting)	Gray codes used to map symbols to bits to reduce bit errors
Restrictions on allowable sequences of valid symbols	Grammar: The sky is blue. (correct) The sky blue is. (incorrect)	Trellis Coded Modulation and types of forward error correction codes, decoded via sequence estimation
Special encoding of signal routing and handling information. Delineation of separate symbols, or groups of symbols	Punctuation: route questions to the mental process needed to give an answer Emphasis: shout to get attention	Framing and synchronization Call setup
Request to repeat garbled messages	I heard you say something, but I did not catch what it was, would you repeat it please?	ARQ protocols based on parity error detection
Power and mode control in difficult environments	Shouting and dragging out easily detected vowels: Biiiillyyy! I told you to stop that nooooow! Mother speaking to infant in slow "baby" talk	Power management and fallback modes to slower bit rates Insertion of longer probe/training sequences

Table 1: Analogies between speech and high frequency (short wave) radio modems. Both types of signals have many characteristics that help the receiver to do a better job of recovering the transmitted information, by reducing the number of errors in the received symbols.

Besides providing a variety of interference reduction and error correction techniques, the AM and FM measurement processes provide a simple parametric representation of speech, that can be exploited to easily modify the characteristics of synthesized speech, such as a speaker's pitch, effective vocal tract length and speaking rate. Obviously, the brain does not resynthesize speech. But it may nevertheless perform some type of speaker normalization, by simply modifying the parametric representation itself, so that all speech appears, to subsequent processing, to be more alike than it actually is. This may simplify the task of speech recognition and understanding. In this regard, it is interesting to note that the speech rate might be continually adjusted, in a manner analogous to the decision-directed tracking techniques employed in modems. Once a decision is made, identifying the last symbol (syllable) to be received, the timing between that and the previous syllables, can be exploited to direct the continuous adjustment of the output speech rate parametric representation, partially compensating for differences in input speech rate. Making such an adjustment is trivial. The real problem is determining what adjustment should be made in the first place. Decision-directed techniques solve that problem in modems. A similar technique might be employed for vocal tract normalization.

6. Conclusions

The considerations presented, lead us to believe that speech communication is not the ad hoc process that it is commonly believed to be. The auditory system seems to employ a number of sophisticated techniques, whose performance enhancing capabilities are well understood in the context of communication theory. Furthermore, there are simple and plausible explanations for how these techniques evolved, and for how they might be implemented. The proposed auditory model can account for many of the performance characteristics of the auditory system. But it is far from being complete. There are still many unknowns, like precise filter shapes, whose impact has been discussed. Others, such as the control law used to adapt the weights used in frequency diversity processing, have only been hinted at, but have a significant impact on performance, including the system's ability to separate multiple acoustic sources. These areas will provide fertile grounds for future research.

Appendix 1: A Channelized Receiver Algorithm

There are numerous techniques for constructing channelized receivers. Many are specialized to the case in which all the channels are equally spaced in frequency, and have the same bandwidth. The technique presented here is more general. It is well known that Fast Fourier Transform (FFT) algorithms, can be used to efficiently implement frequency domain filtering, with highly-selective (sharp cut-off frequency), linear phase, finite impulse response (FIR) filters. It is not so well known that frequency tuning (such as down conversion to baseband) and sample rate decimation, can be accomplished at the same time. This enables the construction of inexpensive, multichannel receivers, which simultaneously provide continuous FFT spectral data, that can be used for signal activity detection. For the latter, frequency domain windowing can be performed after the FFT has been computed, rather than applying a time domain window to the data prior to computing the FFT.

The technique works as follows. A wideband signal is digitized, then overlapping blocks of samples are Fourier transformed (overlap-save or overlap-add algorithms). The FFT data is then multiplied by an FIR filter's FFT, and the filtered signal is converted back to the time domain via an inverse FFT. Frequency tuning is accomplished by simply shifting the FFT "bins" of interest; take the data values from one region of the FFT output array, and reposition them at a different region, such as centered at zero frequency, prior to the inverse FFT.

If the filter response highly attenuates all the spectrum outside of the region of interest near baseband, then only the baseband frequencies contribute anything to the inverse FFT output. Consequently, a decimated output signal can be constructed by merely performing an inverse FFT that is much smaller than the forward FFT; one need not transform the near-zero values in the filter stopband. The overall cost of the receiver per channel is low, because the cost of the comparatively expensive forward FFT is shared by many channels, including the spectral analysis subsystem, while the inverse FFTs are inherently low cost due to their much smaller size.

Depending on how the overlap is performed in the overlap save, a phase correction factor may or may not have to be applied from one block of data to the next, and one may or may not be able to "tune" to every bin within the forward FFT. These considerations result from the fact that how the overlap and tuning is done, impacts the continuity of the FFT's underlying basis functions, from one block to the next.

Unfortunately, the technique is not appropriate, when simulating the halfwave rectification process, that occurs within the auditory system. The efficiency of the technique derives from the ability to decimate the outputs. But rectification would increase the bandwidth of the outputs. That would result in aliasing, if the output sampling rate had been significantly reduced by the filtering prior to rectification.

Appendix 2: Signal Files and Sonograms

For each signal, there is a pair of files, an audio file and its sonogram. Each audio file has approximately five seconds of speech, sampled at 16 kHz, with 16-bit, 2's complement samples. Sonograms are in JPEG format. For signals 10-14, the analysis filterbank was identical, but the synthesis parameters differ.

Signal 1: Original: "The quick brown fox jumps over the lazy dog and the cow jumps over the moon".

Signal 2: Lowpass filtered copy of signal 1, with cutoff frequency of 4 kHz.

Signal 3: Phase vocoding with one detector per octave, 5 detectors (paired a and b filterbanks).

Signal 4: Phase vocoding with three detectors per octave, 15 detectors (paired a and b filterbanks).

Signal 5: Phase vocoding with four detectors per octave, 22 detectors (paired a and b filterbanks).

Signal 6: Phase vocoding with four detectors per octave, 22 detectors (paired a and b filterbanks), adjacent channel suppression; On a sample by sample basis, a test is performed to determine if a stronger channel is adjacent to a weaker one. If there is, the weaker channel's amplitude is set to zero. This reduces the beating caused when adjacent channels respond to the same input tone, and similar, but not identical frequency tones are synthesized.

Signal 7: Phase vocoding with twelve detectors per octave, 65 detectors (paired a and b filterbanks).

Signal 8: Phase vocoding with twelve detectors per octave, 65 detectors, (paired a and b filterbanks), adjacent channel suppression

Signal 9: Phase vocoding with mel scale detectors, 41 detectors (one filterbank, paired a_n and a_{n+1})

Signal 10: Harmonic coding, 42 harmonics, phase jittered

Signal 11: Same as Signal 10, but no even harmonics (n>10)

Signal 12: Instantaneous frequency coding, 42 harmonics, with the n'th harmonic's FM replaced, on a sample by sample basis, by n*pitch, if the FM differs from n*pitch by more than one percent. No even harmonics (n>10).

Signal 13: Same as Signal 10, with each harmonic's AM replaced by AM raised to the power 1.3. This acts somewhat like a matched filter; it amplifies stronger signal components more than weaker ones.

Signal 14: Harmonic coding, 20 harmonics, phase jittered, pitch doubled, speed doubled. AM as a function of frequency is determined by linear interpolation between the measured harmonics.

References:

Dai, H., Nguyen, Q.T., Green, D.M., 1995. A two-filter model for frequency discrimination. Hearing Research 85, 109-114.

Kohata, M., 1999. 1.2 Kbit/sec Harmonic Coder Using Auditory Filters. In: Proc. Internat. Conf. Acoust. Speech Signal Process. Phoenix, AZ, March 15-19, 1999. pp 469-472.

Macon, M.W., Clements, M.A., 1997. Sinusoidal Modeling and Modification of Unvoiced Speech. IEEE Trans. Speech and Audio Processing 5, pp. 557-560.

McAulay, R.J., Quatieri, T.F., 1992. Low-Rate Speech Coding Based on the Sinusoidal Model. In: Furui, S. and Sondhi, M.M, (Eds.) Advances in Speech Signal Processing, Marcel Dekker, Inc., New York, pp.165-207.

McEachern, R., 1992. How the ear really works. In: Proceedings IEEE International Symposium on Time-Frequency and Time-Scale Analysis. Victoria, BC, Canada, pp. 437-440.

McEachern, R., 1994a. Ratio detection precisely characterizes signal's amplitude and frequency. In: EDN, March 3, 1994, pp. 107-112.

McEachern, R., 1994b. Hearing it like it is – audio signal processing the way the ear does it. In: DSP Applications, Vol. 3, No. 2, pp. 35-47.

McEachern, R., 1994c. Hearing it like it is Part 2 – sound demodulation via parallel filter banks. In: DSP & Multimedia Technology, Vol. 3, No. 5, pp. 23-37.

Plomp, R., 1967. Beats of Mistuned Consonances, Jour. Acoust. Soc., Amer., 42, pp. 462-474

Quatieri, T.F., Hanna, T.E., O'Leary, G.C., 1996. AM-FM separation using auditory-motivated filters. In: Proc. IEEE Int. Conf.. Acoust., Speech, Signal Process., Atlanta, GA, May 1996, pp. 977-980.

Quatieri, T.F., Hanna, T.E., O'Leary, G.C., 1997. AM-FM separation using auditory-motivated filters. IEEE Trans. Speech and Audio Processing 5, 465-480.

Rose, J.E., Hind, J.E., Anderson, D.J., Brugge, J.F., 1971. Some Effects of Stimulus Intensity on Response of Auditory Nerve Fibers in the Squirrel Monkey. Jour. Neurophysiol. 34, pp. 685-699

Teich, M.C., Khanna, S.M., Guiney, P.C., 1993. Spectral Characteristics and Synchrony in Primary Auditory-Nerve Fibers in Response to Pure-Tone Acoustic Stimuli. Jour. Stat. Physics 70, pp. 257-279

Warren, Richard M., 1999a. Auditory Perception – A New Analysis and Synthesis, Cambridge Univ. Press, pp. 65-69
Warren, Richard M., 1999b. op cit pp. 172-173

Zwicker, E., Fastl, H., 1999. Psychoacoustics – facts and models, Springer. pp. 113-115.