Phase Coherence as a Measure of Acoustic Quality, part one: the Neural Mechanism

David Griesinger

Consultant, 221 Mt Auburn St #107, Cambridge, MA 02138, USA

PACS: 43.55.Fw, 43.55.Mc, 43.66.Ba, 43.66.Hg, 43.66.Jh, 43.66.Qp

ABSTRACT

The three papers in this series focus on *engagement*, which results when sounds demand, and hold, our attention. Engagement is related to the perception of distance, and is distinct from intelligibility. Sounds perceived as close demand attention and convey drama. Sounds perceived as further away can be perfectly intelligible, but can be easily ignored. The properties of sound that lead to engagement also covey musical clarity – one is able, albeit with some practice, to hear all the notes in a piece, and not just the harmonies. Historically halls for both music and drama were designed to maximize engagement through their size, shape, and liberal use of fabric, tapestries, and ornamentation. Most classical music was composed to be heard in such venues. Modern drama theatres and cinemas still maximize engagement, as do the concert halls at the top of Beranek's ratings. But there is little recognition of engagement in acoustical science, and too few modern music venues provide it. The first of these papers describes the physics and physiology that allow humans to perceive music and speech with extraordinary clarity, and how this ability is impaired by inappropriate acoustics. It also shows how engagement can be measured – both from an impulse response and from recordings of live music. The second paper describes the psychology behind engagement, and makes a plea for concert halls and opera designs that maximize engagement and musical clarity over a wide range of seats. The third paper presents some of the architectural means by which this can be achieved. The conclusions are often radical. For example, excess reflections in the time range of 10ms to 100ms reduce engagement, whether they are lateral or not.

INTRODUCTION

These three talks are centered on the properties of sound that promote *engagement* – the focused attention of a listener. Engagement is usually subconscious, and how it varies with acoustics has been insufficiently studied. In some art forms the phenomenon is well known: drama and film directors insist that performance venues be acoustically dry, with excellent speech clarity and intelligibility. Producers and listeners of popular music, and customers of electronically reproduced music of all genres, also expect – and get – recordings and sound systems that demand our attention.

The author strongly believes that the acoustic properties that convey the excitement of a play, pop song, or film also increase the impact of live classical music, and can co-exist with reverberation. But many seats in current halls and opera houses are not acoustically engaging. They encourage sleep, not attention. Most classical music – and nearly all operas – were not written to be performed in such halls.

Engagement is associated with sonic clarity – but currently there is no standard method to quantify the acoustic properties that promote it. Acoustic measurements such as "Clarity 80" or C80, were developed to quantify intelligibility, not engagement. C80 considers all reflections that arrive within 80ms of the direct sound to be beneficial. As we will see, this is not what engagement requires. Venues often have adequate intelligibility – particularly for music – but poor engagement.

But since engagement is subconscious, and reverberation is not, acoustic science has concentrated on sound decay – and not on what makes sound exciting. Acoustic engineers and architects cannot design better halls and opera houses without being able to specify and verify the properties they are looking for. We desperately need measures for the kind of clarity that leads to engagement. The work in this talk attempts to fill this gap.

The first part of this talk is concerned with the physics of detecting and decoding information contained in sound waves. Specifically we seek to understand how our ears and brain can extract such precise information on the pitch, timbre, horizontal localization (azimuth), and distance of multiple sound sources at the same time. Such abilities would seem to be precluded by the structure of the ear, and the computational limits of human neurology.

We present the discovery that this information is encoded in the phases of upper harmonics of sounds with distinct pitches, and that this information is scrambled by reflections. The reflections at the onsets of sounds are critically important. Human neurology is acutely tuned to novelty. The onset of any perceptual event engages the mind. If the brain can detect and decode the phase information in the onset of a sound – before reflections obscure it – pitch, azimuth and timbre can be determined. The sound, although physically distant, is perceived as psychologically close. The work in part one shows the mechanisms by which the ear and brain can detect pitch, timbre, azimuth and distance by analyzing the information that arrives in a 100ms window after the onset of a particular sound event.

But the significance of this discovery for these papers is that phase information is scrambled predictably and quantifiably by early reflections. Given a binaural impulse response or a recording of a live performance the degree of harmonic phase

coherence in a 100ms window can be used to measure the degree of engagement at a particular seat.

Because engagement is mostly subconscious and largely unknown in acoustic literature, part two of this talk presents some of the experiences and people that taught me to perceive and value engaging sound. Together these experiences become a plea for hall designs that deliver excitement and clarity along with reverberation. Part three of this talk presents some of the reasons a few well known halls are highly valued, and how new halls can mimic them.

"NEAR", "FAR" AND LOCALIZATION

The perception of engagement and its opposite, muddiness, are related to the perception of "near" and "far". For obvious reasons sounds perceived as close to us demand our attention. Sounds perceived as far can be ignored. Humans perceive near and far almost instantly on hearing a sound of any loudness, even if they hear it with only one ear – or in a single microphone channel. An extended process of elimination led the author to propose that a major cue for distance- or near and far – was the phase coherence of upper harmonics of pitched sounds. [1], [2]. More recent work on engagement – as distinct from distance – led to the realization that engagement was linked to the ability to reliably perceive azimuth, the horizontal localization of a sound source. For example, if the inner instruments in a string quartet could be reliably localized the sound was engaging. When (as is usually the case) the viola and second violin could not be localized the sound was perceived as muddy and not engaging. Engagement is usually a subconscious perception, and is difficult for subjects to identify. But localization experiments are easy to perform reliably. I decided to study localization as a proxy for engagement.

Direct sound, Reflections, and Localization

Accurate localization of a sound source can only occur when the brain is able to perceive the direct sound – the sound that travels directly from a source to a listener – as distinct from later reflections. Experiments by the author and with students from several universities discovered that the ability to localize sound in the presence of reverberation increased dramatically at frequencies above 700Hz. Localization in a hall is almost exclusively perceived through harmonics of tones, not through the fundamentals. Further experiments led to an impulse response based measure that predicts the threshold for horizontal localization [3][4]. The measure simply counts the nerve firings that result from the onset of direct sound above 700Hz in a 100ms window, and compares that count with the number of nerve firings that arise from the reflections in the same 100ms window.

$$S = 20 - 10 * \log \int_{.005}^{\infty} p(t)^2 dt$$

LOC in dB =

$$S - 1.5 + 10 * \log \int_0^{.005} p(t)^2 dt$$

$$-(1/D) * \int_0^{D-.005} POS(S + 10 * \log \int_{.005}^{\tau} p(t)^2 dt) d\tau$$

pressure at which nerve firings cease, assumed to be 20dB below the peak level of the sum of the direct and reverberant energy. ***p(t)*** is an impulse response measured in the near-side ear of a binaural head. ***p(t)*** is band limited to include only frequencies between 700Hz and 4000Hz. ***LOC*** is a measure of the ease of localization, where ***LOC*** = 0 is assumed to be the threshold, and ***LOC*** = +3dB represents adequate perception for engagement and localization. ***POS*** means positive values only. ***D*** is the 100ms width of the window.

The first integral in ***LOC*** is the log of the sum of nerve firings from the direct sound, and second integral is the log of the sum of nerve firings from the reflections. The parameters in the equation (the choice of 20dB as the dynamic range of nerve firings, the window size *D*, and the fudge factor -1.5) were chosen to match the available localization data. The derivation and use of this equation is discussed in [3]. The author has tested it in a small hall and with models, and found it to accurately predict his own perception. Similar results have been obtained by professor Omoto at the University of Kyushu.

MEASURING ENGAGEMENT AND LOCALIZATION WITH LIVE MUSIC - THE IMPORTANCE OF PHASE.

The equation for LOC presented above requires binaural impulse responses from fully occupied halls and stages to be useful. These are extremely difficult to obtain. The author has struggled for some time to find a way to measure both localization and engagement from binaural recordings of live music. It ought to be easy to do – if you can reliably hear something, you can measure it. You just need to know how!

In the process of trying to answer this question, the author came to realize that the reason distance, engagement, and localization are related is that they all arise from the same stream of information: *the phase relationships of harmonics at the frequencies of speech formants.*

Perplexing Phenomena of Hearing

Human hearing uses several ways of processing sound. The basilar membrane is known to be frequency selective, and respond more or less logarithmically to sound pressure. With the help of sonograms much has been learned about speech perception. But these two properties of hearing are inadequate to explain our extraordinary ability to perceive the complexities of music – and our ability to separate sounds from several simultaneous sources.

For example, the frequency selectivity of the basilar membrane is approximately 1/3 octave (~25% or 4 semitones), but musicians routinely hear pitch differences of a quarter of a semitone (~1.5%). Clearly there are additional frequency selective mechanisms in the human ear.

The fundamentals of musical instruments common in Western music lie between 60Hz and 800Hz, as do the fundamentals of human voices. But the sensitivity of human hearing is greatest between 500Hz and 4000Hz, as can be seen from the IEC equal loudness curves. In addition, analysis of frequencies above1kHz would seem to be hindered by the maximum nerve firing rate of about 1kHz. Even more perplexing, a typical basilar membrane filter above 2kHz has three or more harmonics from each voice or instrument within its bandwidth. How can we possibly separate them? Why has evolution placed such emphasis on a frequency range that is difficult to analyze directly, and where several sources seem to irretrievably mix?

But in a good hall I can detect the azimuth, pitch, and timbre of three or more musicians at the same time, even in a concert where musicians such as a string quartet subtend an angle of +-5 degrees or less! (The ITDs and ILDs at low fre-

quencies are miniscule.) Why do some concert halls prevent me from hearing the inner voices of a quartet?

As a further example, the hair cells in the basilar membrane respond mainly to negative pressure – they approximate half-wave rectifiers, which are strongly non-linear devices. How can we claim to hear distortion at levels below 0.1% ?

Why do so many creatures – certainly all mammals – communicate with sounds that have a defined pitch? Is it possible that pitched sounds have special importance to the separation and analysis of sound?

Answer – it's the phases of the harmonics!

Answers to these perplexing properties of hearing become clear with two basic realizations:

1. The phase relationships of harmonics from a complex tone contain more information about the sound source than the fundamentals.

2. And these phase relationships are scrambled by early reflections.

For example: my speaking voice has a fundamental of 125Hz. The sound is created by pulses of air when the vocal chords open. All the harmonics arise from this pulse of air, which means that exactly once in a fundamental period all the harmonics are in phase.

A typical basilar membrane filter at 2000Hz contains at least four of these harmonics. The pressure on the membrane is a maximum when these harmonics are in phase, and reduces as they drift out of phase. The result is a strong amplitude modulation in that band at the fundamental frequency of the source. When this modulation is below a critical level, or noise-like, the sound is perceived as distant and not engaging.

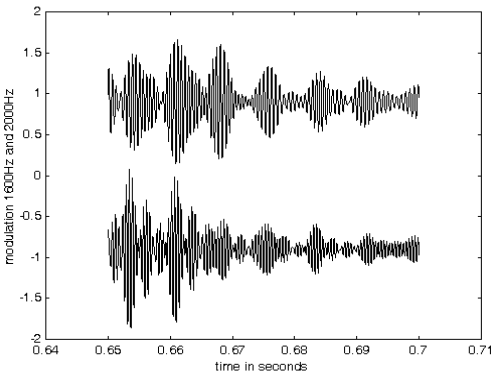

Figure 1: Top trace: The motion of the basilar membrane at a region tuned to 1600Hz when excited by a segment of the word "two". Bottom trace: The motion of a 2000Hz portion of the membrane with the same excitation. The modulation is different because there are more harmonics in the higher frequency band. In both bands there is a strong (20dB) amplitude modulation of the carrier, and the modulation is largely synchronous between the two bands.

Amplitude Modulation

The motion of the basilar membrane above 1000Hz as shown in figure 1 appears to be that of an amplitude modulated carrier. Demodulation of an AM radio carrier is achieved with a diode – a half-wave rectifier – followed by a low pass filter. Although the diode is non-linear, radio demodulation recovers linear signals, meaning that sounds in the radio from several speakers or instruments are not distorted or mixed together. A similar process occurs when the basilar membrane decodes the modulation induced by the phase relationships of harmonics. Harmonics from several instruments can occupy the same basilar region, and yet the modulations due to each instrument can be separately detected.

Both in an AM radio and in the basilar membrane the demodulation acts as a type of sampling, and alias frequencies are detected along with the frequencies of interest. In AM radio the aliases are at high frequencies, and can be easily filtered away. The situation in the basilar membrane is more complicated – but can still work successfully. This issue is discussed in [3].

Figure 2 shows a model of the basilar membrane which includes a quasi-linear automatic gain control circuit (AGC), rather than a more conventional logarithmic detector. The need for an AGC is discussed in [3], but in other ways the model is fairly standard. The major difference between the model in figure 2 and a standard model is that the modulations in the detected signal are not filtered away. They hold the information we are seeking.

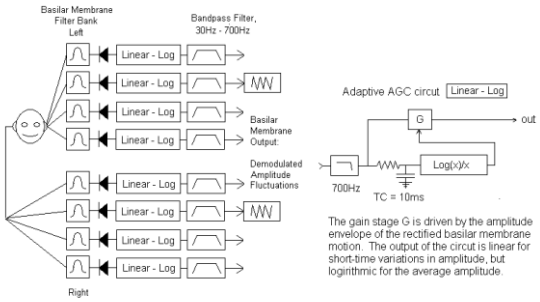

Figure 2: A basilar membrane model based on the detection of amplitude modulation. This model is commonly used in hearing research – but the modulation detected in each band is normally not considered important.

There is one output from figure 2 for each (overlapping) frequency region of the membrane. We have converted a single signal – the sound pressure at the eardrum - into a large number of neural streams, each containing the modulations present in the motion of basilar membrane in a particular critical band.

How can we analyze these modulations? If we were using a numeric computer some form of autocorrelation might give us an answer. But autocorrelation is complex – you multiply two signals together – and the number of multiplications is the square of the number of delays. If you wish to analyze modulation frequencies up to 1000Hz in a 100ms window more than 40,000 multiplies and adds are needed

I propose that an analyzer based on neural delay lines and comb filters is adequate to accomplish what we need. Comb filters are capable of separating different sound sources into independent neural streams based on the fundamental pitch of the source, and they have high pitch acuity. Comb filters have interesting artifacts – but the artifacts have properties that are commonly perceived in music. A comb filter with 100 sum frequencies in a 100ms window requires no multiplies, and only 2000 additions. The number of or taps (dendrites) needed is independent of the delay of each neuron – which means in this model that the number of arithmetic operations is independent of the sample rate.

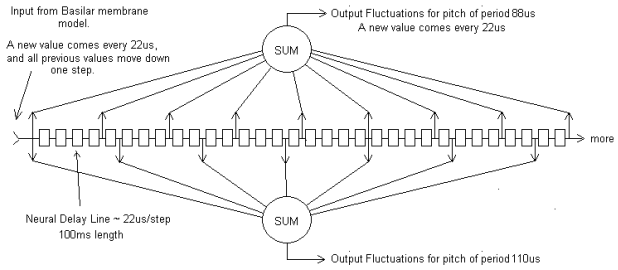

Figure 3: A comb filter analyzer showing two tap periods, one a period of four neural delay units, and one of five neural delay units. In human hearing such a delay line would be 100ms long, and be equipped with perhaps as many as 100 tap sums, one for each frequency of interest. There is one analysis circuit for each overlapping critical band. I have chosen a sample rate of 44.1kHz for convenience, which gives a neural delay of 22us.

Figure 3 shows the analyzer that follows the basilar membrane circuit in the author's model. The analyzer is driven by the amplitude modulations created by the phase coherence of harmonics in a particular critical band. When the fundamental frequency of a modulation corresponds to the period of one of the tap sums, the modulations from that source are transferred to the tap sum output, which becomes a neural data stream specific to that fundamental. The analysis circuit separates the modulations created by different sound sources into independent neural streams, each identified by the fundamental frequency of the source.

If we use a 100ms delay window and plot the outputs of the tap sums as a function of their frequency, we see that the analyzer has a frequency selectivity similar to that of a trained musician – about 1%, or 1/6th of a semitone.

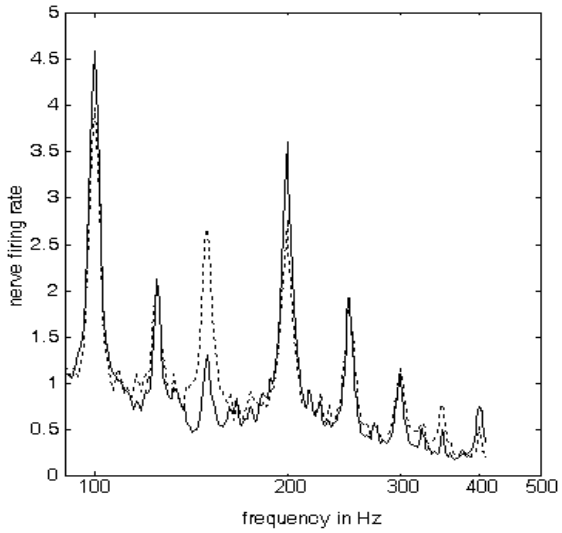

Figure 4: The output of the analysis circuit of figure 3 after averaging the tap sums of six 1/3 octave bands from 700Hz to 2500Hz. Solid line: The modulations created by the harmonics of pitches in a major triad – 200Hz, 250Hz, and 300Hz. Dotted line: The modulations created by harmonics of the pitches from the first inversion of this triad – 1500Hz, 200Hz, and 250Hz. Note the patterns are almost identical, and in both cases there is a strong output at the root frequency (200Hz) and its subharmonic at 100Hz.

Figure 4 shows one of the principle artifacts – and musical advantages – of the comb filter used as an analyzer. The advantage is that the comb filter inherently repeats triadic patterns regardless of inversions or octave, and produces similar output patterns for melodies or harmonies in any key.

The reason for this advantage – and a possible disadvantage – is that the tap sums are equally sensitive to the frequency corresponding to their period and to harmonics of that frequency. In practice this means that there is an output on a tap sum which is one octave below the input frequency. The subharmoic is not perceived, which suggests that the perception is inhibited because of the lack of output from a region of the basilar membrane sensitive to this fundamental frequency (in this case 100Hz).

The comb filter analyser is composed of simple neural elements: nerve cells that delay their output slightly when excited by an input signal, and nerve cells that sum the pulses present at their many inputs. The result is strong rate modulations at one or more of the summing neurons, effectively separating each fundamental pitch into an independent neural stream.

Not only is the fundamental frequency of each pitch at the input determined to high accuracy, once the pitches are separated the amplitude of the modulations at each pitch can be compared across critical bands to determine the timbre of each source independently.

The modulations can be further compared between the two ears to determine the interaural level difference (ILD) and the interaural time delay (ITD). The ILD of the modulations is a strong function of head shadowing, because the harmonics which create the modulations are at high frequencies, where head shadowing is large. This explains our abilities to localize to high accuracy, even when several sources subtend small angles.

Simple experiments by the author have shown that humans can easily localize sounds that have identical ITD at the onset of the sound, and identical ILDs, but differ in the ITD of the modulations in the body of the sound, even if the bandwidth of the signal is limited to frequencies above 2000Hz. A demonstration of this ability using pink noise is on the author's web-site.

WHY THE HEARING MODEL IS USEFUL

The hearing model presented here need not be entirely accurate to be useful to the study of acoustics. The most important aspect of the model is that is demonstrates that many of the perplexing properties of human hearing can be explained by the presence of information in harmonics above 700Hz, that this information can be extracted with simple neural circuits, and that this information is lost when there are too many reflections.

Our model detects and analyses modulations present in the motion of many overlapping regions (critical bands) on the basilar membrane. Although the detection process is non-linear, as in AM radio the modulations themselves are (or can be) detected linearly. The analysis process creates perhaps as many as one hundred separate neural streams from each critical band. But most of these streams consist of low amplitude noise. A few of the outputs will have high amplitude coherent modulations, each corresponding to a particular source fundamental. The frequency selectivity is very high – enabling the pitch to be determined with accuracy. The brain can analyse the outputs from a single pitch across critical bands to determine timbre, and between ears to determine azimuth.

The length of delay line in the analyser (~100ms) was chosen to match our data on source localization. As the length of the delay line increases the pitch acuity increases – at the cost of reduced sensitivity and acuity to sounds (like speech) that vary rapidly in pitch. Tests of the model have shown 100ms to be a good compromise. As we will see, the model easily detects the pitch-glides in speech, and musical pitches are determined with the accuracy of a trained musician. The comb filter analyser is fast. Useful pitch and azimuth discrimination is available within 20ms of the onset of a sound, enabling a rapid response to threat.

But the most important point for these papers is that the fine perception of pitch, timbre, and azimuth all depend on phase coherence of upper harmonics, and that the acuity of all these perceptions is reduced when coherence is lost. When coherence is lost the brain must revert to other means of detecting pitch, timbre, and azimuth. When the coherence falls below a critical level a sound source is perceived as distant – and not engaging.

The degree of coherence in harmonics is a physical property. The model presented above can be used to measure coherence, and this measure can be useful in designing halls and opera houses.

THE EFFECTS OF REFLECTIONS ON HARMONIC COHERENCE

The discrimination of pitch

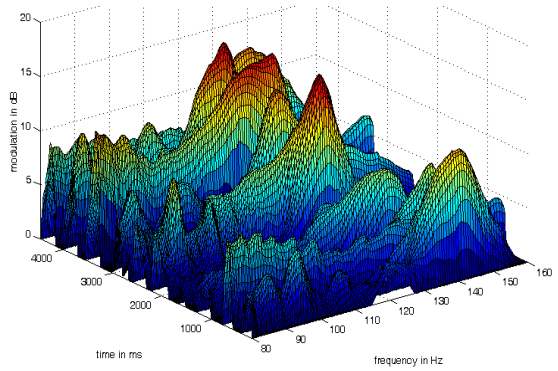

Figure 5: The syllables “one” to “ten” in the 1.6kHz to 5kHz bands. Note that the voiced pitches of each syllable are clearly seen. Since the frequencies are not constant the peaks are broadened – but the frequency grid is 0.5%, so you can see that the discrimination is not shabby.

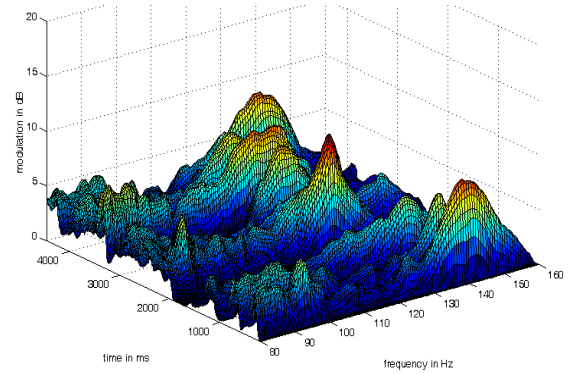

Figure 6: The same syllables in the presence of reverberation. The reverberation used was composed of an exponentially decaying, spatially diffuse, binaural white noise. The noise had a reverberation time (RT) of 2 seconds, and a direct to reverberant ratio (D/R) of -10dB. Although the peak amplitude of the modulations is reduced, most of the pitch-glides are still visible. The sound is clear, close, and reverberant.

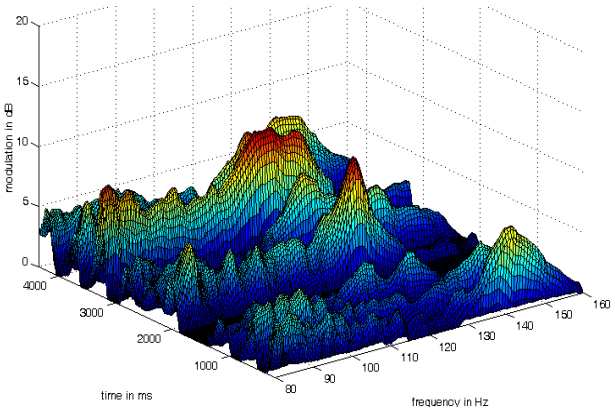

Figure 7: The same as figure 6, but with a reverberation time of 1 second, and a D/R of -10dB. The shorter reverberation time puts more energy into the 100ms window, reducing the phase coherence at the beginning of each sound. Notice that many of the pitch-glides and some of the syllables are no longer visible. The sound is intelligible, but muddy and distant.

The discrimination of horizontal direction (ILD)

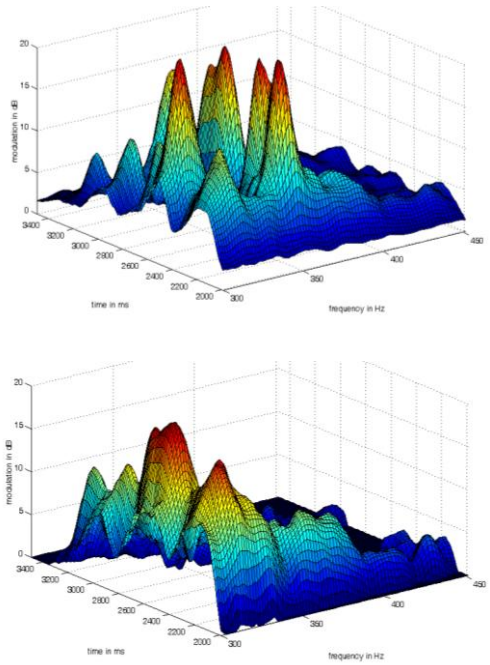

Figure 8: The modulations from two violins playing a semitone apart in pitch, binaurally recorded at +-15 degrees azimuth. The top picture is the left ear, the bottom picture is the right ear. Note the higher pitched violin (which was on the left) is hardly visible in the right ear. There is a large difference in the ILD of the modulations.

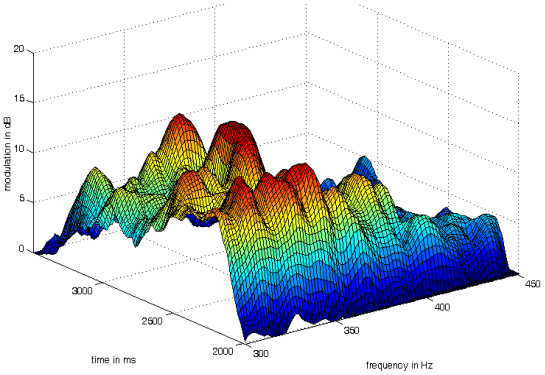

Figure 9: The same picture as the top of figure 8, but with the 1 second RT of figure 7. Note the difference in ILD is far less. The pitch of the higher frequency violin can still be determined, but the two violins are perceived as both coming from the centre. The azimuth information is lost.

Timbre – comparing modulations across critical bands

Once sources have been separated by pitch, we can compare the modulation amplitudes at a particular frequency across each 1/3 octave band, from (perhaps) 500Hz to 5000Hz. The result is a map of the timbre of that particular note – that is, which groups of harmonics or formant bands are most prominent. This allows us to distinguish a violin from a viola, or an oboe from a clarinet.

I modified my model to select the most prominent frequency in each 10ms time-slice, and map the amplitude in each 1/3 octave band for that frequency. The result is a timbre map as a function of time.

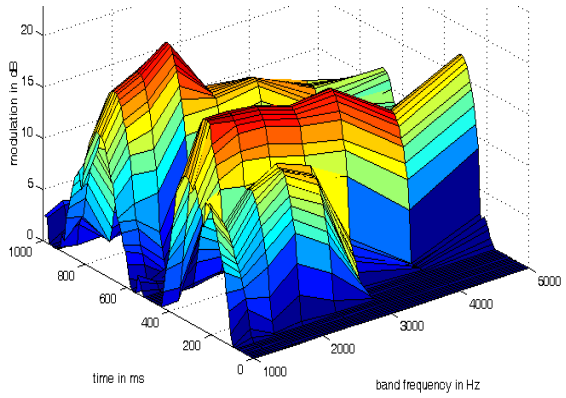

Figure 10: Timbre map of the syllables “one” and “two”. All bands show moderate to high modulation, and the differences in the modulation as a function of frequency identify the vowel. Note the difference between the “o” sound and the “u” sound.

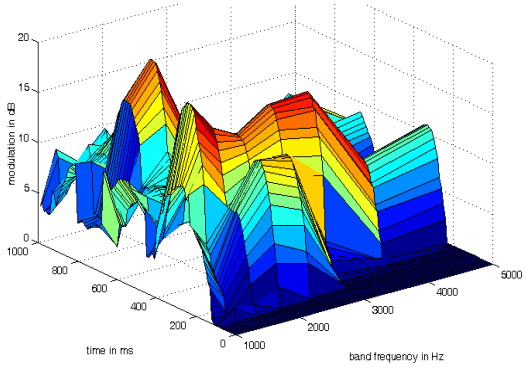

Figure 11: Timbre map of the signal in figure 11, but with a 2 second RT at a D/R of -10dB. Although there is less modulation the timbre pattern of both syllables is almost identical to Figure 10, where no reverberation is present.

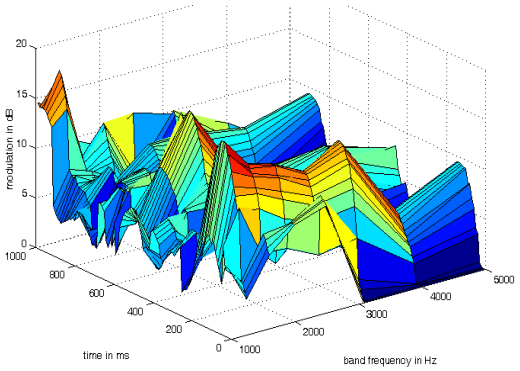

Figure 12: The same as figure 11, but with a 1 second RT. Note that the timbre information is mostly lost. The speech is intelligible – but the primary perception is that the timbre is different – and that the sound is muddy.

SUMMARY OF PART ONE

We postulate that the human ear has evolved not only to analyze the average amplitude of the motion of the basilar membrane, but also fluctuations or modulations in the amplitude of the basilar membrane motion when the membrane is excited by harmonics above 1000Hz. These modulations are at low frequencies, and easily analyzed by neural circuits. As long as the phases of the harmonics that create the modulations are not altered by reflections, the modulations from several sources can be separated by frequency and separately analyzed for pitch, timbre, azimuth, and distance.

The modulations – especially when separated – carry more information about the sound sources than the fundamental frequencies, and allow precise determination of pitch, timbre, and azimuth.

The phases of the harmonics that carry this information are scrambled when the direct sound from the source is combined with reflections from any direction. However if the amplitude of the sum of all reflections in a 100ms window starting at the onset of a sound is at least 3dB less than the amplitude of the direct sound in that same window the brain is able to perceive the direct sound separately from the reverberation, and timbre and azimuth can be perceived. The

sound is likely to be perceived as psychologically close, and engaging.

Reflections from any direction – particularly early reflections – scramble these modulations and create a sense of distance and disengagement. But they are only detrimental to music if they are too early, and too strong. The model presented above makes it possible to visualize the degree to which timbre and pitch can be discerned from a binaural recording of live music in occupied venues.

At present the pictures which result from the model with live music sources need to be subjectively evaluated to determine if the sound is engaging, but with and further calibration a single-number measure for engagement should be possible.

REFERENCES

1 D.H. Griesinger, "Pitch Coherence as a Measure of Apparent Distance and Sound Quality in Performance Spaces"Preprint for the conference of the British Institute of Acoustics in May, 2006. Available on the author's web site: www.davidgriesinger.com

2 D.H. Griesinger, "Pitch Coherence as a Measure of Apparent Distance and Sound Quality in Performance Spaces" A powerpoint presentation given as the Peter Barnett memorial lecture to the Institute Acoustics conference in Brighton, November 2008. Available on the author's web-page.

3 D.H. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment" A power point presentation with audio examples given at the Portland meeting of the Acoustical Society of America, May 2009.

4 D.H. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment" A preprint for a presentation at the 126th convention of the Audio Engineering Society, May 7-10 2009. Available from the Audio Engineering Society.