1aAA14. The audibility of direct sound as a key to measuring the clarity of speech and music

Session: Monday Morning, Oct 31

Time: 11:30

Author: David H. Griesinger
Location: David Griesinger Acoustics, 221 Mt. Auburn St., Cambridge, MA 02138
dgriesinger@verizon.net
www.davidgriesinger.com

Abstract:

Human ear/brain systems evolved to decode the direction, timbre, and distance of multiple sounds in a complex and noisy environment. In a reverberant space this information is only available at the onset of a sound, before reflections overwhelm it. We propose that the perception of clarity depends on our ability to detect the information that this brief onset contains, and to be able to use it to separately analyze multiple sounds for direction, timbre and distance. In a reverberant environment source separation and the perception of direction, timbre and distance are closely related, as they all degrade at the same time as reverberant level increases. We find that all these abilities depend on phase relationships between harmonics of complex tones, and that these phase relationships are scrambled in predicable ways by reflections and reverberation. Well known properties of human hearing are used to develop both a physical model for the neurology of onset detection and an impulse response measure for the ability to clearly localize sounds in a reverberant field. A C language implementation of the physical model is capable of predicting and perhaps measuring the localizability of individual musicians in a binaural recording of live music, possibly leading to a real-time measure of clarity.

1 INTRODUCTION

Sabine measured the reverberation time of spaces by blowing a continuous tone on an organ pipe, stopping the flow of air, and then measuring the time it took for the sound to become inaudible with a stopwatch. He measured reverberation time this way because the equipment was simple and the data was repeatable. His method, with some refinement, is still in use. The data correlates to some degree with the subjective impression of rooms. But it is not by itself predictive of how successful the space will be for either speech or music. Current standardized measures of room acoustics were developed the same way. We find a technology that might be used to measure a physical property of sound, hoping the data correlates with some subjective property. Sometimes it does correlate, but only if we average many rooms. Our ability to predict the success of a particular space remains limited.

The problem is compounded by the difficulty of defining the properties of sound we would ideally like to hear. It is hard to accurately define something you cannot measure, and it is hard to design a measure for something you cannot define. But if we want to have the tools we need to

reliably design spaces with appropriate acoustics for their use, we have to break out of this dilemma.

A possible path out of the dilemma may be to examine how the ear and brain extract such an extraordinary amount of information from a noisy, complex, and confusing sound field. Along with nearly all animals we can perceive and localize tiny sounds in the presence of enormous background noise and other possibly simultaneous noises, evaluate these sounds for threat, and respond appropriately. As social animals we have evolved to be able to choose to pay attention to one of three or more simultaneous conversations. If someone we are not paying attention to speaks our name we instantly shift our attention to that voice. This is the cocktail party effect, and it implies that we can detect the vocal formants of three or more speakers independently, form their speech into independent neural streams, and at a subconscious level scan these streams for content.

But when reflections and reverberation become too strong the sonic image becomes blurred. We can no longer form independent neural streams and separately localize simultaneous sounds. For speech the result is babble – although we may be able with difficulty to comprehend the loudest voice. All sounds blend together to form a sonic stew. With music such a stew can be pleasing, even if the detail of performance or composition is lost. But the brain is operating in a backup mode, and our minds can easily wander.

Additional insight into this phenomenon can be found in the work on classroom acoustics by SanSoucie. [1] Research has shown that it is not sufficient that the teacher's words be intelligible in the rear of the classroom. They must be sufficiently clear that the brain can recognize each vowel and consonant without guesswork or context. When conditions are poor working memory is insufficient to hold the incoming speech long enough to both decode it and then to process and remember it. In average classroom acoustics students can hear the teacher but they cannot remember what was said.

Another example might come from the arcane field of stage acoustics. A physicist/musician friend was complaining to me about the difficulty of hearing other instruments in a small concert stage with a low ceiling. He suggested adding reflectors overhead to increase the loudness of his colleagues. But experiments showed this only made the problem worse. The problem was not the lack of level from the other musicians, it was the inability of the players to perform the cocktail party effect. They could hear their own instruments, but could not separate other instruments from the sonic muddle. The solution was to increase the clarity on stage by *reducing* the strength of early reflections.

This paper is primarily concerned with clarity. Not the kind of clarity that is measured with C80 or C50, but the kind of clarity that enables us to easily form independent neural streams for simultaneous sounds, and then find their direction, timbre, and distance. This is what our brains evolved to do, and when we can do it what we hear becomes more understandable, beautiful, and interesting. We find that the mechanisms behind the cocktail party effect also predict the ease with which we identify vowels, and hear the direction and distance of multiple sources. Once we understand how the brain performs this miracle, making a measure for it becomes possible. We will show the physics of the sonic data that enables the cocktail party effect, and how the brain

has evolved to decode it. We will present a relatively simple formula for measuring from a binaural impulse response the ease with which we can perceive the details of sound.

2 THE PHYSICS OF HEARING

2.1 What Do We Already Know?

1. The sounds we want to hear in a performance space are speech and music, both of which consist of segments of richly harmonic tones 25ms to 500ms long, interspersed with bursts of broadband high frequency energy. It is likely we will not understand hearing or acoustics without understanding the necessity of harmonic tones.

2. There is a tremendous improvement in signal to noise ratio (S/N) if an organism possesses the ability to analyze the frequency of incoming sound with high precision, as then most of the background noise can be filtered out. Tones with a large number of high harmonics all related by a single pitch contain the information we need to identify potential threats, the vowels in speech, the source direction and how quickly we must act, all while retaining the S/N advantage of a single pitch. Speech and the complexities of music depend on such tones.

3. We can perceive pitch, timbre, direction and distance of multiple sources at the same time, and in the presence of background noise. This is the well-known cocktail party effect, essential to our successful navigation of difficult and dangerous social situations.

4. Perhaps as a consequence human hearing is extraordinarily sensitive to pitch. A musician can tune an instrument to one part in one thousand, and the average music lover can perceive tuning to at least an accuracy of one percent. This is amazing given the frequency selectivity of critical bands, which are about one part in five. Such pitch acuity did not evolve by accident. It must play a fundamental role in our ability to hear – and might help us understand how to measure acoustics.

5. The fact that the pitch of low frequency sine tones varies with the loudness of the tone would seem to make playing music difficult. But the acuity to the pitch of sine-tones is a maximum at about 1000Hz. When we perceive the pitch of low tones primarily from the frequencies of their upper harmonics the perceived pitch is stable with level. We need to explain how we can perceive the pitches from the upper harmonics of several instruments at the same time when such harmonics are typically unresolved by critical bands.

6. Physics tells us that the accuracy with which we can measure the frequency of a periodic waveform depends roughly on the product of the signal to noise ratio (S/N) of the signal and the length of time we measure it. If we assume the S/N of the auditory nerve is about 20dB, we can predict that the brain needs about 100ms to achieve the pitch acuity of a musician at 1000Hz. So we know there is a neural structure that can analyze sound over this time period.

7. Physics also tells us that the amount of information that any channel can carry is roughly the product of the S/N and the bandwidth. The basilar membrane divides sound pressure into more than 40 overlapping channels, each with a bandwidth proportional to its frequency. So a critical band at 1000Hz is inherently capable of carrying ten times as much information as a critical band at 100Hz. Indeed, we know that most of the intelligibility of speech lies in frequencies between

700 and 4000Hz. We need to know the physics of how information is encoded into sound waves at these frequencies, and how this information is decoded.

8. The cocktail party effect implies that we can detect the vocal formants of three or more speakers independently, even when the sounds arrive at our ears at the same time. Pitch is known to play a critical role in this ability. Two speakers speaking in monotones can be heard independently if their pitch is different by half a semitone, or three percent.[2] If they whisper, or speak at the same pitch, they cannot be separated. The vocal formants of male speakers are composed of numerous harmonics of low frequency fundamentals. When two people are speaking at once the formant harmonics will mix together on the basilar membrane, which is incapable of separating them. We should hear a mixture of formants, and be unable to understand either speaker. But it is clear that the brain can separate the harmonics from two or more speakers, and that this separation takes place before the timbre – and thus the identity of the vowel – is detected. We believe that our acuity to pitch evolved to enable this separation. A few audio examples of sound separation by pitch can be found in [9].

9. Onsets of the sound segments that make up speech and music are far more important to comprehension than the ends of such segments. Convolving a sentence with time-reversed reverberation smoothes over the onset of each syllable while leaving the end clear. The modulation transfer function – the basis of STI and other speech measures – is unchanged. But the damage wrought to comprehension is immensely greater when reverberation is reversed.

10. When there are too many reflections we can sometimes understand speech from a single source, but in the presence of multiple sources our ability to perform the cocktail party effect is nullified and the result is babble. In the presence of too many reflections our ability to detect the timbre, distance, and direction of single sources is reduced, and the ability to separately detect these properties from multiple sources is greatly reduced.

11. We have found that accurate horizontal localization of sound sources in the presence of reverberation depends on frequencies above 1000Hz, and accuracy drops dramatically when the direct to reverberant ratio (D/R) decreases only one or two dB below a certain value.

The threshold for accurate horizontal localization as a function of the D/R and the time delay of reflections can be predicted from a binaural impulse response using relatively simple formula, which will be discussed later in this paper.

2.2 Amplitude Modulation - The key to this paper

A final bit of physics makes these observations understandable. Harmonics of complex tones retain in their phase vital information about the process that created them. Almost invariably these harmonics arise from a pulsed excitation – the opening of the vocal cords, the release of rosin on a string, the closing of a reed, etc. Thus at the moment of creation all the harmonics are in phase, and the amplitude of the sound pressure is a maximum. Since the harmonics are all at different frequencies they drift apart in phase, only to be forced back together once in every fundamental period. In the absence of reflections this phase alignment is preserved as sound travels to a listener. Once in every fundamental period the harmonics align in phase and produce a maximum of sound pressure. As they drift apart they destructively interfere with each other, and the sound pressure decreases. In the absence of reflections the modulation of the pressure is

large – approaching a 20dB difference between pressure maxima and minima. These modulations can be seen in Figure 1.

A typical male speaking voice has a fundamental at about 125Hz, and generates harmonics at multiples of this frequency. Thus in a critical band centered at 1000Hz we expect to find harmonics at 750Hz, 875Hz, 1000Hz, 1125Hz, and 1250Hz. These harmonics interfere with each other to produce a modulation in the motion of the membrane that resembles the signal of an AM radio. As can be seen in figure 1 there is a carrier at the frequency of the basilar filter, and this carrier is strongly amplitude modulated at the frequency of the fundamental and some of its harmonics. Not coincidentally the basilar membrane detects this motion exactly as an AM radio would. It rectifies the signal, detects the modulation, and passes the modulation to the brain without the carrier.

Thus the ear detects not only the average amplitude in a critical band, but also modulations in that amplitude at the frequencies of the fundamentals of complex tones. Moreover, the modulation and detection process is linear. If there are harmonics from two or more tones present at the same time they are all detected and passed to the brain without intermodulation. Evolution has found a method of utilizing the inherent information carrying ability of higher frequencies without requiring that the carrier frequencies be detected directly. And it has found a way of linearizing an inherently non-linear detector.

2.3 Summary of the known Physics and Psychophysics of Sound

1. Vital information in speech and music is carried primarily in frequencies above 700Hz.

2. Onsets of speech and musical sounds are far more important to comprehension than the way sound decays. The small segment of direct sound that carries with it accurate information about the timbre and localization of the source is often quickly overwhelmed by reflections. To predict acoustic quality we need to know under what conditions precise data on timbre and localization are lost.

3. Separately determining timbre, direction, and distance of sound from several simultaneous sources in a complex sound field depends on the presence of tones with many high harmonics, and on the likelihood that the pitches of the tones from separate sources are slightly different. This dependency has driven the evolution of our acute sensitivity to pitch. We also know that human pitch perception is circular in octaves. Do Re Mi is the same in any octave.

4. Our ability to separate the harmonics in the vocal formant range from two or more sources at the same time depends on the phase alignment of the harmonics from each source. The phase alignment of the harmonics from each source creates amplitude modulation of the basilar membrane at the frequency of each fundamental, and these modulations combine linearly. The brain stem can separate them from each other and from background noise by their pitch.

5. Reflections from any direction alter the phase relationships between harmonics of complex tones, reducing and randomizing the amplitude modulation of the basilar membrane. The result is intermodulation between sources, distortion, and noise. Separation of sources by pitch becomes difficult. The brain stem must revert to a simpler method of decoding sound. The sources blend together, and only the strongest of them can be accurately perceived and localized.

Our current acoustic measurements do not take these facts of human perception into account. The reverberation time (RT) has been standardized to follow Sabine's method. The standard is equivalent to exciting the room with an infinitely long continuous signal, and measuring the rate of decay when the signal stops. Measures such as clarity, (C80 or C50), measure aspects of the response of a room to an impulse – an infinitely short signal. C80, C50, and IACC measure aspects of the onset of sounds, but only for the sounds pistols – fortunately rare in speech and music. Neither the infinitely long or the infinitely short excitation resemble the properties of music, either in the duration of the excitation or in the essential presence of richly harmonic tones.

There are also a number of myths that dominate acoustic thought. One of the most misleading of these myths is the "law of the first wave-front" which is widely interpreted to mean that the direct sound – the sound that travels to the listener before the reflections arrive – is always distinctly audible. The definitions of C80, C50, IACC and others rely on this so-called law. They start their measurement time with the arrival of the direct sound, whether it is audible or not. Indeed, the direct sound in an impulse response always looks like it should be audible. But this is a consequence of using an infinitely short signal as an excitation. Real signals nearly always have a significant rise time and a finite duration. Will the direct sound still be audible – or even visible in a graph? What if the sum of early reflection energy is greater than the direct sound? Will the direct sound be audible?

To complicate matters further, both RT and the early decay time (EDT) measure the way sound decays in rooms. (The current standardized measurement for EDT is flawed both in its mathematical definition and its intended meaning.) But it is clear that the human ear and brain are uninterested in how sound decays. Sound decay is essentially noise. It can be beautiful, but much of the information the sound might contain – such as its unique timbre and the direction of the source – is lost in the decay. It is the onsets of sounds that convey their meaning, and our ears and brains have evolved to extract as much of this information as possible before reflections and reverberation overwhelm it.

3 A PHYSICAL MODEL OF SOUND DETECTION

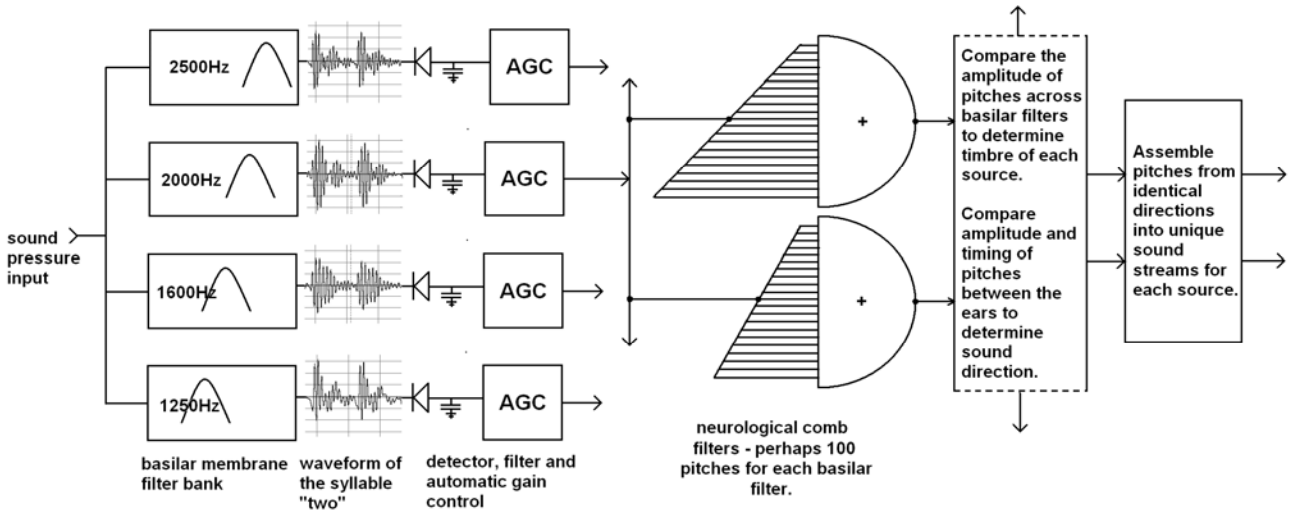

Figure 1: Sounds entering the ear are separated into frequency bands by a bank of overlapping mechanical filters with relatively low selectivity. At the vocal formant frequencies each filter typically contains three or more harmonics of speech or musical fundamentals. These harmonics interfere with each other to create a strongly amplitude modulated signal. Actual waveforms of these amplitude modulations are shown in figure one. Note that the modulation depth is large, and the peak amplitudes align in time. The modulations in the signal are detected linearly by the hair cells, but like an AM radio with automatic gain control the nerve firing rate for time variations longer than about 20 milliseconds is approximately logarithmically proportional to the sound pressure. The brain stem separates these modulations by pitch using a number of comb filters each ~100ms long. Two filters out of about one hundred for each critical band are shown in the figure. The filters detect pitches using the travel speed of nerve pulses in tiny fibers.

Once separated by pitch the brain stem compares the amplitude of the modulations for each pitch across the critical bands to determine the timbre of the source, and compares the amplitude and timing of the modulations at each pitch between the two ears to determine sound direction. Using these cues the brain stem assembles events into separate foreground sound streams, one for each source. Sound left over after the foreground is extracted is assigned to a background sound stream.

Reflections and reverberation randomize the phases of the harmonics. When the reflections are too strong the modulations in each frequency band become noise-like, and although pitch is still detectable, timbre and direction are not. The mechanism in figure one is similar to current models by other researchers, except in our model complex tones are separated by pitch before analysis for timbre and localization. Distance (near/far) is inferred by the ease with which pitch separation takes place.

4 A SIMPLIFICATION BASED ON AN IMPULSE RESPONSE

The above model can be used to analyze the localizability of sound sources in a binaural recording of live music. But it would be very useful to predict localizability – and thus a measure of sound quality – from a measured impulse response. There is a simple graphic that explains a method for developing such a measure. It first mathematically manipulates an impulse response to resemble the sound pressure from a sound of finite length, and then graphs way the energy of reflections between 700Hz and 4000Hz build up with time. The graphic enables to us to visualize the process by which the brain extracts information from the onset of a sound.

Let's assume we have a sound source that suddenly turns on and then holds a constant level for more than 100ms. Initially only the direct sound stimulates the basilar membrane. Soon the first reflection joins it, and then the next, etc. The nerve firing rate from the combination of sounds is approximately proportional to the logarithm of the total sound pressure, but we can plot the rate of nerve firings from the direct sound and the reflections separately. In the following graphs the vertical axis is labeled "rate of nerve firings", normalized such that the rate is 20 units for the sum of both rates once the reverberation is fully built-up. The scale is chosen so that the value of the rate is proportional to the sound pressure in dB. (To simplify the graph we assume the nerve firings cease 20dB below the final maximum sound pressure, implying a S/N of 20dB.) Thus in figure two the rate for the direct sound is about 13, implying that the total sound pressure will eventually be 7dB stronger than the direct sound. The data shown in these graphs were measured by the author in the unoccupied Boston Symphony Hall (BSH). They use the ipeselateral (source

side) signal from the author's binaural microphone. The omnidirectional source was at the conductor's position. The binaural microphone is equalized to have essentially flat frequency response from 30Hz to 5000Hz for sounds from the front. (Ideally we should equalize to match an inverse equal loudness curve.)

We postulate that if the total number of nerve firings from the direct sound exceeds the total number of nerve firings from the reflections in the first 100ms, then a sound source will be localizable. If the total number of nerve firings from the reflections exceeds the total number from the direct sound, the sound will not be localizable.

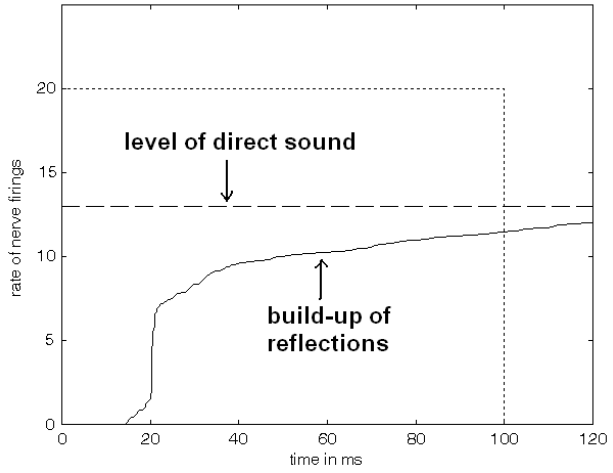

Figure 2: The relative rate of nerve firings from the direct sound and the build-up of reverberation in the frequency range of 1000Hz to 4000Hz in unoccupied Boston Symphony Hall (BSH) row R, seat 11, with a source at the podium. The dashed line shows the rate of nerve firings for a sound of constant level that begins at time zero. The solid line shows the firing rate due to the reverberation as it builds up with time. The dotted line marks the combined final firing rate for a continuous excitation, and the 100ms length of the time window the brain stem uses to detect the direct sound.

In this seat the direct sound is strong enough that the ratio of the area in the window under the direct sound (the total number of nerve firings from the direct sound in this window) to the area in the window under the build-up of the reflections is 5.5dB. This is the value for ***LOC*** – the measure that will be discussed in the next section. This value implies excellent localization and clarity.

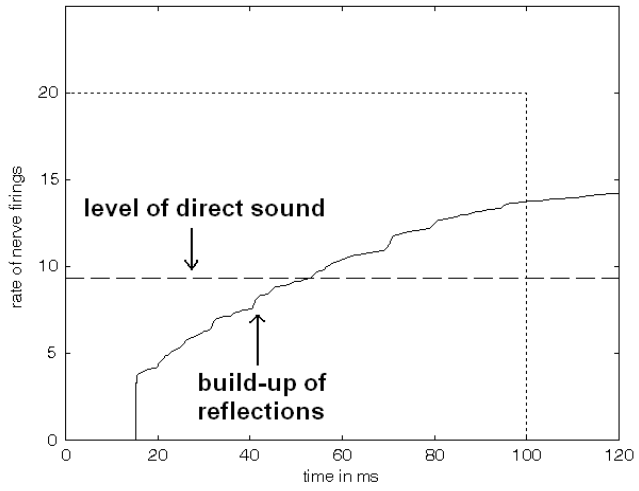

Figure 3: Nerve firing rates for the direct sound and the build-up of reflections in unoccupied BSH, row DD, seat 11. ~90ft from the stage. Notice the direct sound is weaker than in row R, and there is a strong high-level reflection at about 17ms that causes the reflected energy to build up quickly. The ratio of the areas (the total number of nerve firings) for the direct sound in the first 100ms to the area under the line showing the build-up of the reflections is 1.5dB. Localization in the occupied hall is poor in this seat. Subjectively the ratio of areas would be below zero. It is likely that in the occupied hall audience absorption in front of this seat reduces the direct sound.

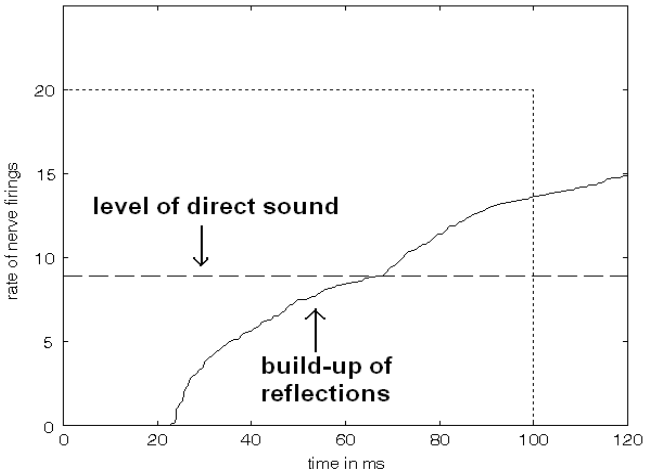

Figure 4: Rates of nerve firings for the direct sound and build-up of reflections in BSH, front of first balcony, row A, seat 23 ~110ft from the stage. The direct sound is weaker here – but there are no strong early reflections. The ratio of areas is +2.2dB, and localization is better than in row DD on the floor. (Subjectively this seat is superb. The clarity is better than this graphic predicts, and the envelopment is amazing. An occupied measure would likely show a higher value for LOC.)

The postulate used to define LOC holds up well in the author's experience. The graphic and the formula for LOC came from a series of experiments on the threshold of localization in the presence of reflections of various amplitude and time delay [3][4].The parameters in the model – the choice of -20dB for the zero of nerve firings and the 100ms length of the time window can be adjusted slightly to fit the localization data. But in experiments in a small 300 seat concert hall and in the BSH data shown above the model predicts the seats where localization is difficult. Given the sharpness of the threshold for localization, the accuracy of prediction is remarkable.

4.1 A Mathematical Equation for Predicting Localizability from an Impulse Response

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, implying that 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 for male speech [3][4]. The measure simply counts the nerve firings above 700Hz in a 100ms window that result from the onset of a continuous direct sound, and compares that count with the number of nerve firings that arise from the reflections in the same 100ms window.

(1)
$$S = 20 - 10 * \log \int_0^{\infty} p(t)^2 dt$$

(2) ***LOC*** in dB =

$$S + 1.5 + 10 * \log \int_0^{.005} p(t)^2 dt - (1/D) * \int_{.005}^{D} POS\,(S + 10 * \log \int_{.005}^{\tau} p(t)^2 dt) d\tau$$

In equation 1 above ***S*** is a constant that establishes a sound 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. Equation 2 calculates the value of ***LOC.*** It 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 section in ***LOC*** is the log of the sum of nerve firings from the continuous direct sound. The second section contains a double integral. The right-hand integral calculates the build up of reflected energy from a continuous tone as a function of time, and the left-hand integral finds the sum of nerve firings that result from that build-up of energy. Note that the left-hand section integrates the LOG of the build up of pressure, and not the build up directly. This distinction is extremely important. Nerve firings are roughly proportional to the LOG of pressure, not pressure itself. If we attempt to integrate the pressure and not the log of pressure the variation of ***LOC*** with both time delay and level of reflections does not match our data at all. Because the effect of the reflections is logarithmic with pressure, the earlier a reflection comes to the direct sound the larger the effect will be on the value of ***LOC***. This effect can be easily seen by comparing figures 3 and 4 above.

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][4]. The author has tested it in a small hall and with models, and found it to accurately predict his own perception. The latest Matlab code for calculating ***LOC*** and producing the graphs shown in figures 2,3, and 4 is in [4].

In practice using the measure needs care. Orchestral instruments are not omnidirectional, and localization and engagement are often better than the LOC measure would suggest. Ideally the source directivity for the impulse response must match the directivity of a particular instrument. Using an omnidirectional microphone rather than a binaural microphone will also underestimate the localizability, as there is substantial head shadowing in a binaural microphone, which reduces the strength of lateral reflections in the ipselateral ear. So LOC is useful, but not yet predictive of the localizability or engagement of every instrument in every concert.

5 THE SUBJECTIVE IMPORTANCE OF *LOC*

5.1 Timbre, Localization, and Distance

LOC does not depend on the hearing model shown in figure 1. It was developed to predict (as precisely as possible) our data on the threshold for localization of speech in the presence of reflections and reverberation. But its design is based on the known facts of hearing outlined above. First, it manipulates the impulse response to represent the room's response to a sound of finite duration. Second, it analyzes the onset of such a sound, not the decay. Third, it includes a window, or region of interest, of ~100ms, a time interval that crops up in loudness detection and many other aspects of hearing. Fourth, the threshold is predicted by a simple signal-to-noise argument – if the number of nerve firings from the direct sound exceed the number from reflections in the first 100ms, then the sound will be localizable. So far as I have been able to test it, ***LOC*** is predictive of localization. It does not simply correlate with it. If ***LOC*** is above +3dB, a sound will be sharply localized even in the presence of other sounds.

The hearing model in figure one may not be accurate in detail. Biological systems may not contain comb filters – although I know of no other filter design that is circular in octaves, can achieve the needed pitch resolution, and uses so little hardware. But the physical properties of band filtered sound on which figure one is based – namely the amplitude modulation induced by the phase relationships of upper harmonics – is real, observable, and can be modeled. The effects of reflections and reverberation on this stored information can be measured and calculated. This is not guesswork, it is straightforward science. We have modeled the mechanism in figure one with a mixture of C language and Matlab. The model appears to be able to predict from a live recording the localizability of a string quartet in two closely adjacent rows in a concert hall. [5]

The physics of sound on which figure one is based predicts that the same mechanism – amplitude modulation induced by coherent phases – also powers the ability to perceive the timbre of multiple sources. There is no other adequate explanation for our ability to perform the cocktail party effect. The ease of timbre perception is the key element in recognizing vowels – and a major component of the ease with which we perceive, recognize, and remember speech. [6] So ***LOC*** may be useful in quantifying speech quality. The research described in this paper started with an attempt to understand the sonic perception of distance [7], where the connection between harmonic tones and amplitude modulation was first made. So our perception of distance – and thus the psychological necessity of paying attention – depends on the same physical mechanism as localization. The importance of low perceived sonic distance to drama and engagement is explored in [3][4] and [8].

Cochlear implants show that a standard hearing model – one based only on the amplitudes of sound pressure in critical bands – is adequate to comprehend speech. But users of these implants find music incomprehensible, and the cocktail party effect out of the question. Acoustic measures based on standard hearing models may be similarly flawed.

5.2 Stream Formation and Envelopment

The caption of figure 1 shows a proposed mechanism by which brain stem assembles independent neural streams from each source in a sound field. But there is another interesting aspect of stream formation. When it is possible to detect the direct sound – and thus the timbre and localization of sound sources – it is possible for the brain to separate this perception from the perception of reflections and reverberation. The timbre and location of the direct sound – detected at the onsets of sounds – is perceived as extending through the note, even though the information has been overwhelmed by reflections. This is the foreground data stream. But since the separation has already been made, the brain can assign the loudness and direction of the reverb to a different type of stream – the background stream. It is usually impossible to identify independent sources in the background stream. Reverberation is heard as harmony, and can be very beautiful. In our experiments with localization we find that in a typical hall when the direct sound is not detectable, not only is timbre and direction difficult to perceive, the reverberation and the notes become one sonic object, and this object – although broad and fuzzy – is located in front of the listener. When the D/R increases just a little bit, suddenly the sound image can become clear, and the reverberation is perceived as both louder and more surrounding the listener. In demonstrating this effect to audiences of 50 to 100 people I have found that many – but by no means all – listeners can easily perceive the change from frontal to enveloping. It may take a bit of learning to perceive this effect, but it is quite real. The enveloping reverberation is

more attractive than the muddled together front image and reverberation combined. This is the envelopment we are looking for in hall design – and it too appears to depend on ***LOC***. This effect is seen in the data on Boston Symphony Hall presented in figures 2, 3, and 4. The seat in figure 3, with the lowest value of LOC, has not only poor localization, it has the least enveloping sound.

6 CONCLUSIONS

We have proposed that an under-researched aspect of human hearing – the amplitude modulations of the basilar membrane motion at vocal formant frequencies – is responsible for much of what makes speech easily heard and remembered, makes it possible to attend to several conversations at the same time, and makes it possible to hear the individual voices that make up much of the complexity and delight of music performance. A model based on these modulations predicts a great many of the seemingly magical properties of human hearing.

The power of this proposal lies in the relatively simple physics behind these hearing mechanisms. Understanding the relationships between acoustics and the perception of timbre, direction and distance of multiple sound sources becomes a physics problem – namely how much do reflections and reverberation randomize the phase relationships and thus the information carried by upper harmonics. The advantage of putting acoustics into the realm of physics is that the loss of information can be directly quantified. It becomes independent of the training and judgment of a particular listener.

A measure, ***LOC,*** is proposed that is based on known properties of speech and music. In our limited experience LOC predicts – and does not just correlate with – the ability to localize sound sources simultaneously in a reverberant field. It may (hopefully) be found to predict the ease of understanding and remembering speech in classrooms, the ease with which we can hear other instruments on stages, and the degree of envelopment we hear in the best concert halls.

A computer model exists of the hearing apparatus shown in figure one. The amount of computation involved is something millions of neurons can accomplish in a fraction of a second. But the typical laptop finds it challenging. Preliminary results indicate that a measure such as LOC can be derived from live binaural recording of music performances.

7 REFERENCES

1. S. SanSoucie ‘Speech comprehension while learning in classrooms’ Dot Acoustics www.dotacoustics.com (June 2010)
2. A.S. Bregman ‘Auditory Scene Analysis’ page 560. MIT Press 1994
3. D. Griesinger ‘The Relationship between Audience Engagement and the ability to Perceive Pitch, Timbre, Azimuth and Envelopment of Multiple Sources’ Tonmeister Tagung 2010. (this paper is on the author’s web-page – www.davidgriesinger.com)
4. D. Griesinger – ‘Listening to Acoustics’ slides from a lecture at BBM acoustics Munich containing Matlab code and much else - http://www.davidgriesinger.com/engagement11.ppt
5. D. Griesinger - With the permission of the Pacifica String Quartet we can hear two examples from a concert in at 1300 seat shoebox hall. The sound in row F is quite

different from the sound in row K. The recordings are from the author's eardrums, and are equalized for playback over loudspeakers or headphones equalized to sound identical to loudspeakers. (Most headphones have too bright a sound to reproduce them correctly. Pink noise played though the headphones should sound identical in timbre to the same noise played through a frontal loudspeaker.)
"Binaural Recording of the Pacifica String Quartet in Concert row F";
(http://www.davidgriesinger.com/Acoustics_Today/row_f_excerpt.mp3)
"Binaural Recording of the Pacifica String Quartet in Concert row K";
(http://www.davidgriesinger.com/Acoustics_Today/row_k_excerpt.mp3)

6. H. Sato 'Evaluating the effects of room acoustics and sound systems using word intelligibility and subjective ratings of speech transmission quality' ISRA Awaji Japan April (2004)
7. D. Griesinger 'Subjective aspects of room acoustics' ISRA Awaji Japan April (2004)
8. D. Griesinger 'Clarity, Cocktail, and Concerts: Listening in Concert Halls' Acou. Today Vol. 7 Issue 1 pp 15-23 January (2011)
9. Audio examples of clarity can be found at the following link: "Audio Examples of Clarity and Pitch Separation" www.davidgriesinger.com/Acoustics_Today/examples.mp3