INTERNATIONAL SYMPOSIUM on MUSICAL and ROOM ACOUSTICS
ISMRA 2016
September 11-13, 2016
La Plata, Buenos Aires, Argentina

Evaluation of Concert Halls / Opera Houses: Paper ISMRA2016-28

The effects of room acoustics on the physics and neurology that enable us to separate information in sonically complex environments

David Griesinger[(a)]

[(a)] David Griesinger Acoustics, USA, dgriesinger@verizon.net

Abstract

Humans can tell instantly, independent of timbre or loudness, if a sound is close to us. We are also able in a crowded room to switch attention at will between at least three simultaneous conversations, and involuntarily switch to one of them if our name is spoken. These are the abilities that allow musicians to hear on stage, and the audience to hear music as composers intended. But these feats are only possible if individual voices can be separated into independent neural streams. We will present data showing that the ability to do this relies on the phase relationships between the harmonics above 1000 Hz that encode speech and music information, and the neurology of the inner ear that has evolved to detect them. Once in each fundamental period harmonic phases align to create massive peaks in the sound pressure at the fundamental frequency. Pitch-sensitive filters can detect and separate these peaks from each other and from noise with amazing acuity. But reflections and sound systems randomize phases, with serious effects on attention, source separation, and intelligibility. We will show how ears and speech co-evolved, and present recent work on the importance of phase in acoustics and sound design.

Keywords: Presence, Proximity, Phase, Stream separation

INTERNATIONAL SYMPOSIUM on MUSICAL and ROOM ACOUSTICS
ISMRA 2016
September 11-13, 2016
La Plata, Buenos Aires, Argentina

The effects of room acoustics on the physics and neurology that enable us to separate information in sonically complex environments

1 Introduction: What is “proximity”

Humans can tell in a fraction of a second if a sound is close to us. Recent work by Lokki [1] has identified this perception as an important – perhaps the most important – sonic perception that predicts preference in concert hall sound. He has named the perception “proximity”. The author of this preprint has been focused on this perception for many years, and has used other names, such as “sonic distance”, “engagement”, and “presence”. We find “proximity” to be a better description for this perception, and we will use it in this paper.

Rapidly detecting proximity has survival value, as close sounds demand attention. But detecting proximity is just the tip of the iceberg for the neural circuits that cutthroat evolution has provided for us. The neural mechanisms that detect proximity enable our ears to separate simultaneous speech sounds from each other and from noise of many types. We have found – largely by a process of elimination – that both proximity and separation of simultaneous sources depends on the presence of signals where the essential information is encoded in the harmonics of complex tones. For example, humans are able to separately understand two monotone sentences if their pitches are different by only half a semitone. If they are at the same pitch the task is impossible. So pitch – a distinct fundamental creating multiple harmonics – is essential. It is no surprise that we use such an encoding for speech. Musical instruments are similar. Their distinct timbre depends on the spectrum of their harmonics. And if they differ in pitch, we can choose to follow the lines of one or more at the same time.

2 Proximity and attention

We care about proximity because it influences our behavior. When a sound is perceived as close it involuntarily evokes attention. We can choose to ignore it, but that takes effort. A sound perceived as far away can be easily ignored. Drama and cinema directors know this effect very well, and demand clear direct sound to their audiences. 18th and 19th opera houses were also direct-sound dominated, with audience sitting as close as possible to the performers, and with lots of absorption all around.

Older concert venues were much less reverberant than is now popular. The Alte Gewandhaus in Leipzig had a reverberation time of 1.2 seconds when Haydn and Mozart performed there. The Thomas Kirche in Leipzig was festooned with banners during Bach’s time, with a reverberation time of only about 1.6 seconds.

But proximity seems to be largely ignored in modern concert halls, opera houses, and even classrooms. To make matters worse, all the standard acoustic quality measures are blind to proximity. But once the physics of the mechanism for detecting presence – the same

mechanism we use for source separation and fine localization – is known we can predict how it will vary in a space, and optimize the space for both proximity and reverberance.

3 Physics of proximity

We believe these abilities can all be explained by the physics of harmonic tones. Speech and musical tones are created by bursts of energy: the opening of the vocal cords, the release of rosin on a string, or the opening of a reed. All these mechanisms create a short burst of energy at the fundamental frequency. The harmonics we hear are created in that instant. At that instant they are all in phase, and the air pressure is a maximum. Once in every fundamental period they are forced to be in phase again, producing a maximum air pressure.

For more than a hundred years speech and other sounds have been analyzed using spectrograms, which plot of the frequency content in the sounds as a function of time. Spectrograms can be created by FFT analysis, which are creatures of block frequencies and time. But the ear does not use FFTs. It analyzes sound in the time domain. The ear uses continuous, relatively broad filters in the basilar membrane. They need to be broad enough that each critical band contains three or more harmonics. You need at least three to recreate the pulses that are so obvious in the rapid amplitude structure of speech and music.

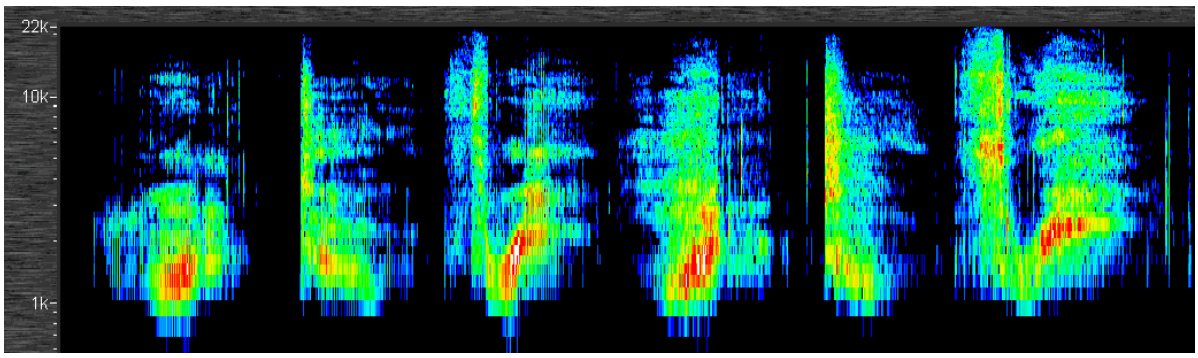

Figure 1: A spectrogram of the author speaking “one two three” first voiced, then whispered. Note the vocal formant bands above 1000Hz are clearly seen, and are nearly identical in the two spectra. But the time structure of the waveform is not visible. Spectrograms mask this detail.

These sharp spikes to the ears of a listener, where they can be easily observed with a microphone, and easily heard by the ear. In textbooks sound is almost always represented as a continuous wave. This is profoundly misleading. The sounds we rely on for communication and for pleasure consist of a series of pulses generated by a fundamental pitch.

The regular spikes shown in figure two are created by the vocal cords. Once in every fundamental period they open with a burst of pressure, creating a spike, a mini delta function, that creates the upper harmonics we hear. The harmonics are locked by their phases to re-align once in each fundamental period to re-create the pressure spike that made them.

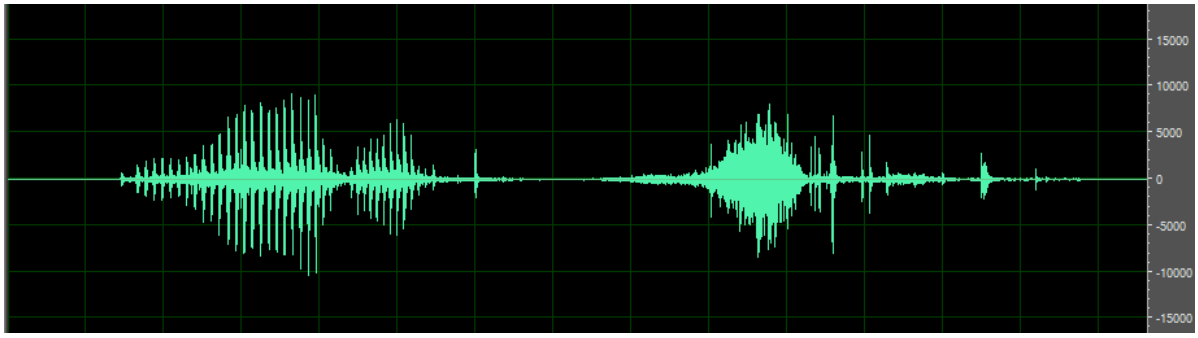

Figure 2: the time waveform of the syllable "one" high pass filtered at 1000Hz. The first waveform is voiced, the second whispered. They have identical sound power. Note the sharp regularly spaced spikes in the time waveform of the voiced syllable. The spectrograms are practically identical but the sound is very different. Although the fundamental has been filtered away, it can be clearly heard in the voiced signal.

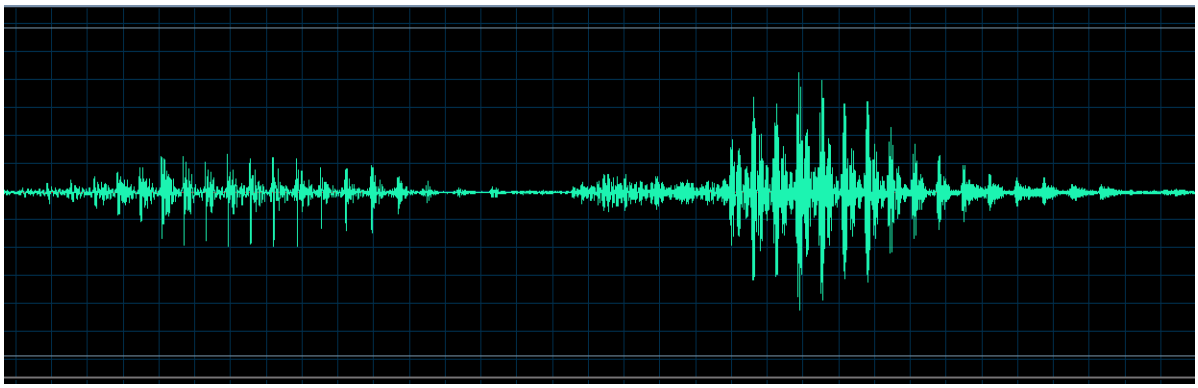

Figure 3: The time waveform of the syllables "one" and "two" with no reflections. Note that there are spikes, and the spikes have a regular period. If we high pass filter these signals at 1000Hz the fundamental is clearly heard.

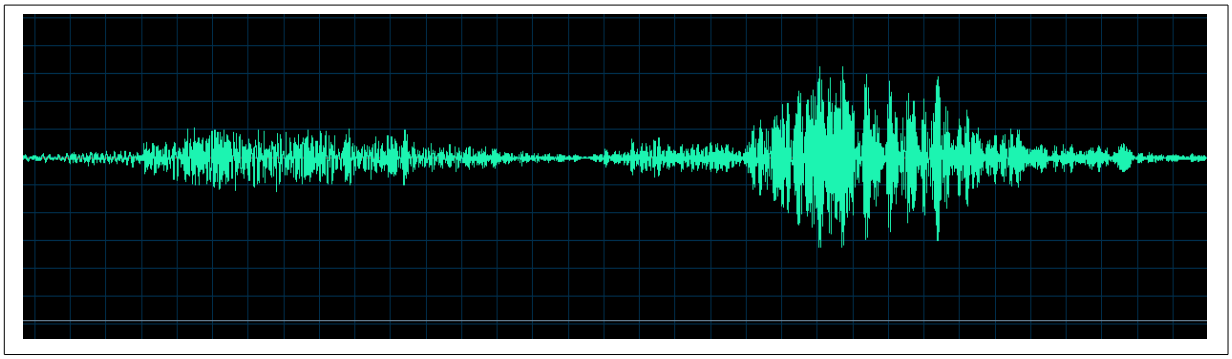

Figure 4: The same two syllables with too many early reflections. There are fewer spikes, and those that exist are at random times. When high pass filtered at 1000Hz the fundamental cannot be heard. These syllables sound distant, and they cannot be separated from other sounds by pitch.

But the phases can be fragile. Reflections that arrive from any direction can alter the phases of harmonics. They need only be delayed by a half a period to have this effect. We can calculate the delay needed: about 5ms for male speech and 3ms for female speech. Any reflection greater than these delays will alter harmonic phase. Too many reflections coming too soon and the spikes will no longer be detectable by the ear. The sound source will sound distant, and no longer be sharply localized.

4 The necessity of a fundamental period

Human speech is voiced. We identify vowels by formant profiles at frequencies mostly above 1000Hz. The vocal formants of voiced speech are formed out of three to ten harmonics of a low frequency fundamental. This method of encoding information is not accidental. It is ideal for transporting information through a complex and noisy environment. The vocal formants are mostly above frequencies where nerve firings can synchronize with the carrier frequency. But nerves can synchronize with the fundamental frequency that created the spikes, and extract the amplitude information they carry from noise and similar peaks of a different fundamental pitch.

The fact that speech and music are largely a series of pulses is vital. The pulses stand out in the presence of noise, and formant information can be extracted from them. But evolution has done more than this. With some simple neural circuitry simultaneous pitched signals can be separated into independent neural streams. In 1951 J. C. R. Licklider proposed that the acuity of human pitch detection could only be explained if underneath the hair cells in the basilar membrane there were neural networks resembling autocorrelators. We have been modeling this type of structure for many years [2].

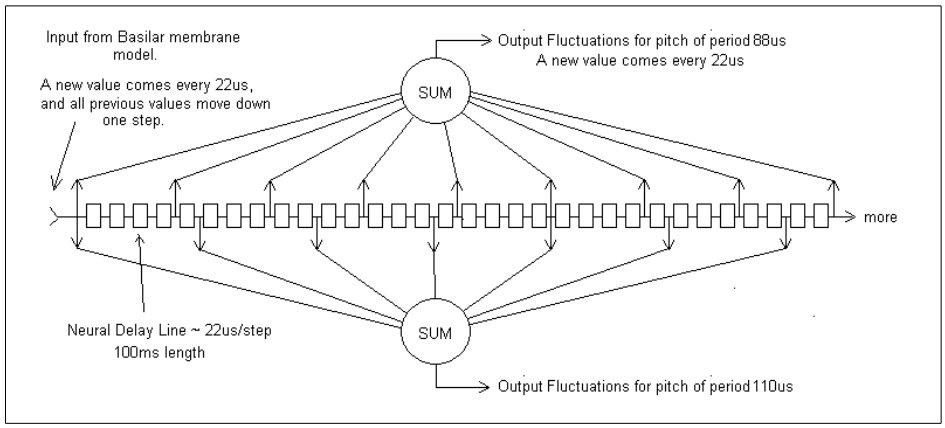

Figure 5: An autocorrelator formed from a pair of comb filters.

A comb filter autocorrelator similar to figure 5 is capable not only of determining pitch to high accuracy, but it can use that fine pitch discrimination to separate pitched signals from each other and from noise. Further – pulses with the same pitch in the two ears can be compared in the brain stem to find ILD and ITD, so pitched signals can be independently localized to high accuracy. In fact Blauert observes in Spatial Hearing that speech can be localized to high accuracy by ITD alone throughout the whole audio range [3].

Speech is not constant in pitch, and two people talking at the same time will sometimes overlap in time. To separate them cleanly you need a cue as to which pitch belongs to which talker. Localization is the best cue. We find a 1dB difference in ILD is sufficient to separate two talkers with a pitch difference of only one semitone. Lacking the ILD, timbre can be used, as we all know from listening to musical instruments.

It is mysterious that the ability to separate voices by pitch has been so little studied. The inner ear autocorrelator such as figure five gives a creature a large evolutionary advantage. We find the separation process depends on the regularity of the pressure peaks formed by the harmonics of low frequency tones. Noise-like signals cannot be separated in the same way. For example, a chorus produces multiple phases which do not form regular peaks on the basilar membrane. We can still detect pitch to high accuracy, but localizing each singer in the group precisely is usually not possible.

The author was once privileged to hear a concert of 40 fine Boston Symphony Orchestra string players under Ton Koopman performing a Haydn symphony in Boston Symphony Hall. They played without vibrato. The instruments sounded as if they were one. It was like a string quartet of solo instruments, sharply localized to the center of the section. When pitches from several instruments become the same to within a cycle or two, the phases of the upper harmonics can form regular peaks in the pressure waveform, and the brain interprets the source as close, single, and sharply localized. The sound was strongly engaging. In Haydn's time this sound may have been the rule, not the exception.

Comb filters also have the property that they respond similarly to the octaves of fundamentals, and respond to musical fourths and fifths exactly as we do, which strongly suggests that comb filters are the basis of an inner-ear autocorrelator. In our model an array of such combs tuned over the span of a low frequency octave is connected to each region of the basilar membrane. We model the membrane as overlapping 1/3rd octave second order filters with a 1/6th octave spacing. In our model the frequency resolution of the combs is better than 1%.

5 Phase and the intelligibility of speech with noise

There is also evidence of the effects of phase on human speech intelligibility in the presence of noise. Shi, Sanechi, and Aarabi (2006) [4] tested the intelligibility of speech with different degrees of phase randomization. Phase randomization was achieved by adding a random variable to the phase component of half-overlapping block fast Fourier transforms with a block size of 512 samples at 44.1kHz sample rate. This technique – sometimes called a decorrelator - is also used in music compression schemes such as MP3 surround to re-create the original degree of phase correlation in an encoded signal.

In their experiment they controlled the degree of randomization through a variable alpha that varied from 0 to 1. As aloha varied from 0 to 1, the original phase was reduced and the random phase increased. At an alpha value of 0.5, the phase was half original and half a variable that

varied randomly between +/- alpha /2. At alpha equals 1.0, the phase was totally random between +/- pi. They then added Gaussian (white) noise to standard speech word lists at signal to noise ratios (SNR) of -10, -5, 0, and +20 dB, and plotted the word error rate as a function of alpha. Note that except for the 0 dB and +20dB SNR case, the speech signal was lower in level than the noise, although since the noise was Gaussian most of the noise energy was at frequencies above 10,000 Hz. This would tend to obscure consonants more than noise, but phase randomization would not be expected to alter consonants.

6 Measuring proximity and the ability to localize speech

In figure 2 the voiced and whispered syllables have the same spectrum and sound power, but the voiced syllable has clearly higher peak amplitude. Figure 3 shows the same higher peak amplitude for the voiced syllable. Could the difference between the RMS amplitude and the peak amplitude of a signal be used as a measure for presence?

Figure 4 shows that the idea will not work for reverberant signals. Reverberation does not eliminate peaks, it creates random peaks in a vowel waveform. To measure presence we need to measure not just the height of the peaks, but also their regular spacing. Thus to measure the amount of proximity in a signal we need to use comb filters or a pitch-sensitive autocorrelator. We also need to use an accurate model of the outer, middle, and inner ear.

We have been working on such models for five years or more. They have yet to work as well as a human, but they are already useful for measuring proximity. For this purpose we measure the output of each basilar membrane filter before it enters the comb filter bank, and then measure the output of the comb filter that best matches the fundamental pitch of the harmonics. We express the result as a ratio in dB. When proximity is high the comb output can be as much as 10dB higher than the input. When the peaks are random, the ratio tends toward zero dB. Figure 8 shows values of proximity measured this way in a large classroom at Harvard.

7 Measuring proximity and localization with a binaural impulse response

The author has developed an algorithm for measuring proximity from an impulse response [5], [6].. The method is based on several concepts: 1. That the ear contains some form of correlator that analyzes sound over a time period of 80 to 100 milliseconds. 2. That the ear is concerned with the number of nerve firings that fall within that window, and not with the sound energy or the sound pressure. 3. That the rate of nerve firings is proportional not to the sound pressure or the sound power, but to the logarithm of sound pressure or power. And 4. That individual nerve fibers have a maximum firing rate, and a minimum firing rate. This implies that the signal to noise ratio in the nervous system is finite – somewhere between 20 and 30dB.

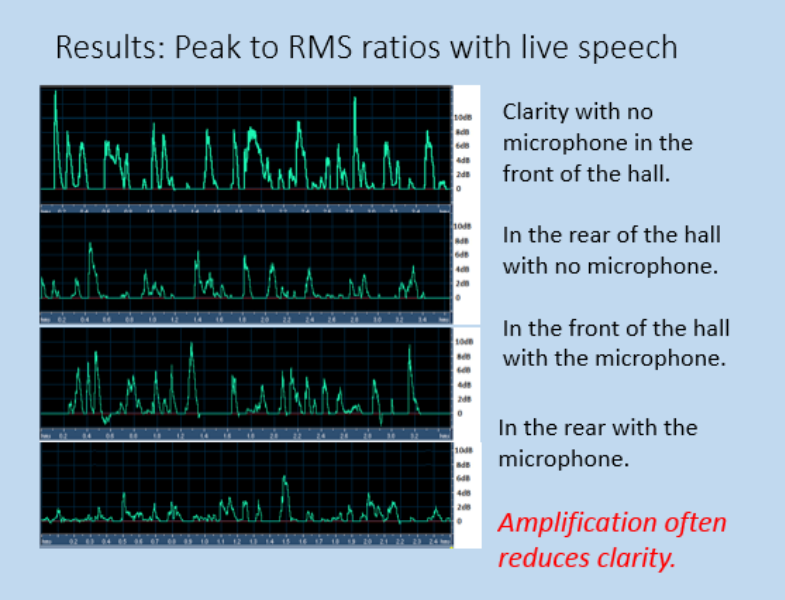

Figure 6: The ratio in dB of the output of the best-matched comb filter to the comb input for speech in a large classroom at Harvard. There is a measurable difference in this ratio which corresponds well to the perceived proximity. The rear of the hall with no microphone had very poor proximity. The sound was loud and intelligible, but the students were not listening.

These reasonable assumptions dictate a way to construct a measure. First we convolve the impulse response with a rectangular function that simulates a sound that starts at time zero and continues at constant amplitude for a length of time greater than 100ms. But we do it separately for the direct sound – the sound power in the first 5ms – and then for the reflections without the direct sound. The direct sound gives us a constant value. The convolution with the reflections will start at zero, and then rise as more and more reflections contribute. If we plot the logarithm of these two functions we will see the picture shown in figure 9. We assume the S/N of the nerve firings is ~20dB.

Matlab and C language scripts for calculating LOC can be found on the author's web-site. The Matlab scripts draw the picture shown in figure 9, so the user can see just how the ear hears a note played through a particular impulse response. We have been using LOC for hall, opera, and classroom analysis since 2008, and the correspondence between the LOC values and what I hear is quite good. Currently we calculate LOC separately for both ears of a binaural recording. We find that the perception of proximity is best predicted by the lower of the two values obtained.

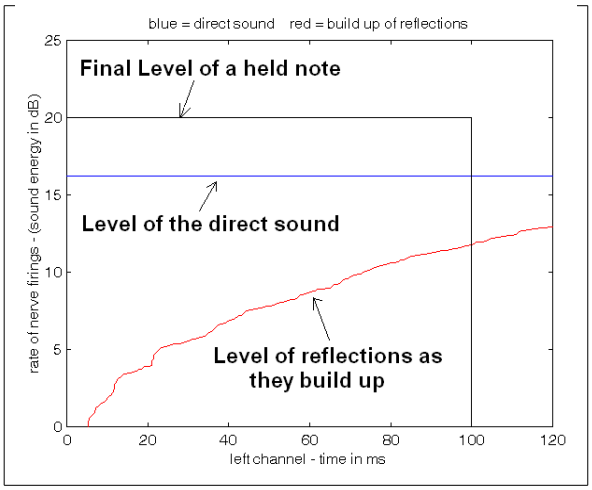

Figure 9: A diagram showing the logarithm of the direct sound level for a held note versus the buildup of the reflected energy as a function of time. To calculate LOC we simply find the ratio between the total area under the blue line inside the 100ms window, to the area under the red line inside the same window. A value of +3dB or more indicates that the number of nerve firings from the direct sound will outnumber the nerve firings from the reflections, and presence will be good. The LOC value for the diagram above – from one of the author's seats in Boston Symphony Hall - is +9dB, a satisfactory value. The sound in this seat is good.

8 The Limit of Localization Distance, or LLD

We have found, and the work of Aarabi shows, that the ability to detect the pressure peaks at the fundamental period is to some degree all or nothing. This can be demonstrated quite simply by walking away from a small ensemble (such as a string quartet) with eyes closed.

Up close the instruments are sharply localized, and the listener can tell which instrument played each note. As you walk away they remain sharply localized and proximate. Only the azimuth spread decreases. Surprisingly the impression of the hall is remarkably constant.

But at a particular distance the perception changes dramatically. The instruments blend together into a fuzzy ball, and the hall, instead of being a separate and highly enveloping perception, becomes part of the instruments, and is largely frontal.

We identify this distance as the limit of localization distance, or LLD. Aarabi's experiment shows a similar effect. Word error rate in noise is relatively low and constant until the random function blended with the phase of the speech signal reaches a peak value of +- pi/2. At this point the word rate dramatically increases.

We believe the number of seats in a hall that are within the region defined by the LLD is an important measure of hall quality. And it can be easily determined by walking around while listening with eyes closed, or eyes open if an electronic ensemble is used. Considerable

experience with other listeners has found that with a live string quartet the eyes dominate the aural perception. If you can see the instruments you will be sure you are hearing them precisely.

But we believe the aural perception of proximity is vital to attention. Music that is proximate has a visceral attraction that fuzzy music does not, regardless of what we are seeing.

9 Conclusions

We have presented evidence that the perception of proximity – the sense of being aurally closely connected to a speaker or performer – arises from the ear's ability to detect the sharp pulses in amplitude waveforms at vocal formant frequencies that are created when harmonics of low frequency fundamentals are formed.

These amplitude and regularity of these pulses are reduced by reflections, which randomize the phases on which the pulses depend. We find that to detect these pulses with the acuity of the human ear an autocorrelator of some form must exist in the inner ear. Our comb filter models of this autocorrelator perform moderately well.

We find that the earliest and the strongest reflections are the most likely to disturb the perception of proximity. We have developed a method called LOC of predicting from a binaural impulse response whether or not proximity will be heard. In work described in another preprint for this conference we find that it is not sufficient that just one ear has and adequate value of LOC. Values of LOC above 3dB are needed in both ears for clear localization and the sense of proximity to be perceived.

We also find in this other preprint that attenuating or redirecting the first few strong reflections – typically from the stage or the side walls – can greatly increase the limit of localization distance or LLD, and increase the number of seats where proximity can be perceived.

We believe that the LLD is easily determined in practice, and is an important and underappreciated aspect of hall quality.

9.1 References

[1] Lokki, T., Pätynen, J., Kuusinen, A., & Tervo, S. (2012). Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *The Journal of the Acoustical Society of America, 132,* 3148-3161

[2] Griesinger, D. What is clarity and how can it be measured J. Acoust Soc. Am 133, 3224-3232 (2013)

[3] Blauert, J. Spatial Hearing – MIT Press 1983 p153

[4] Shi, G., Sanechi, M., Aarabi, P., On the Importance of Phase in Human Speech Recognition. IEEE transactions on audio, speech, and language processing, vol. 14, no. 5, September 2006

[5] Griesinger, D. What is clarity and how can it be measured J. Acoust Soc. Am 133, 3224-3232 (2013)

[6] Beranek L. Concert hall acoustics: Recent findings J. Acoust. Soc. Am. 139(4) April 2016