Audio Engineering Society

Convention Paper 9659

Presented at the 141st Convention
2016 September 29–October 2 Los Angeles, USA

The physics of auditory proximity and its effects on intelligibility and recall

David Griesinger[1]

[1] *David Griesinger Acoustics, 221 Mt Auburn St., Cambridge, MA 02138, USA*

Correspondence should be addressed to David Griesinger (dgriesinger@verizon.net)

ABSTRACT

Cutthroat evolution has given us seemingly magical abilities to hear speech in complex environments. We can tell instantly, independent of timbre or loudness, if a sound is close to us, and in a crowded room we can switch attention at will between at least three different simultaneous conversations. And we involuntarily switch attention if our name is spoken. These feats are only possible if, without conscious attention, each voice has been separated into an independent neural stream. We believe the separation process relies on the phase relationships between the harmonics above 1000Hz that encode speech information, and the neurology of the inner ear that has evolved to detect them. When phase is undisturbed, 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. This talk will detail the many ways ears and speech have co-evolved, and recent work on the importance of phase in acoustics and sound design.

1 Introduction

The importance of phase at frequencies above 1000Hz has been widely dismissed. But Blauert in "Spatial Hearing" [1] writes that "localization by ITD alone can be done with increasing accuracy over the entire audible range." How is this possible, and why did the ability evolve? An answer can be found as early as 1951 with Licklider's seminal paper "A duplex theory of pitch perception" [2]. Nearly all mammals, and many other creatures, communicate using information encoded in the upper harmonics of complex tones. These harmonics are created by a burst of pressure when the vocal cords open – causing a sharp peak in sound pressure. This forces all the harmonics of the tone to align in phase once every period, recreating an image of the original burst. These peaks can be easily seen in the envelope of any speech signal. They tend to cut through noise, and the ear is sensitive to them. When they are present the sound has a close "proximate" sound that is attention-grabbing. But reflections – especially early reflections - randomize these phases, and the envelope at the formant frequencies becomes noise-like. The sound becomes distant and muddy. Recent work by Arrabi et al [3] has shown that this type of phase randomization has a dramatic effect on speech intelligibility in the presence of noise. Licklider's paper explains why. If we postulate that the inner ear contains an autocorrelator tuned to the speech fundamentals, it not only explains our octave-circular precise sense of pitch, it explains our ability to separate particular voices from noise and other voices.

1.1 Intelligibility is not sufficient for recall

A major theme of this preprint is that word intelligibility does not necessarily measure the ability to communicate by speech. Or, for that matter, to perceive the full artistic content of music. It is not sufficient to recognize individual words. There must be sufficient working memory left in the brain to store their meaning. In short, intelligibility is necessary for recall, but it is not sufficient. The clarity and speed of the speaker, the hearing ability of the listener, and the noise or reflections in the acoustic channel will all affect the time it takes for the brain to recognize each word.

Many studies have investigated the effects of noise masking on the ability to recall speech. For example, Tepring Piquado [4] found that masking of individual words by speech babble at a level of -2dB results in 88.8% accuracy of single word detection, but if only one word in a sentence is masked, there is near complete loss of recall of that word and the preceding word in the sentence.

This is because word recognition is only the first step in communicating information. The meaning of the words must be found, assembled into sentences, the rules of grammar applied, and finally the meaning must be held long enough in working memory to store in long-term memory.

1.2 Proximity and attention

One of the most critical factors in this process is attention. If the listener is partly thinking about something else, they might think they are following the speech perfectly, but will be unable to recall it later.

We believe when the ear detects the peaks in the envelope of speech and music our brains automatically and sub-consciously pay attention. The source is close enough to be dangerous.

This is one of the reasons that drama directors insist on close contact between actors and audience, and theaters that are acoustically dry. Cinema directors are also aware of the attention grabbing properties of direct speech. From the beginning of talking pictures dialog has been reproduced from a highly directional phase-linear speaker behind the center of the screen.

If dialog is panned between two speakers interference between the two sound sources randomizes the phase of upper harmonics, and some of the dramatic intensity is lost. The sound can still be intelligible, but it loses its sense of proximity, and its attention grabbing properties.

A strong sense of proximity has been demanded of drama theaters since historic times. Lokki has made an extensive study of the Greek ampitheater at Epidaurus. [5] Speech intelligibility and a sense of dramatic connection to the actors is present in every seat, even though the sound pressure is low. There is a gentle reverberation from the steps to the left and right of a listener that adds a sense of community, but is never stronger than the direct sound. These are ideal conditions for drama.

1.3 Loss of proximity by reflections

Panning dialog between two or more speakers is not the only way to randomize phase. Early reflections can do this effectively if they are too early and too strong. Typically the earlier they come the stronger they are, and the larger effect they have on the onset of vowels and notes. So smaller venues are particularly subject to the loss of proximity.

We made image-source models of two typical classrooms, and binaurally auralized the sound in different seats. The sound varied a lot from front to rear. Figures one and two show two classroom models, both of which when measured by STI, the current standard for measuring intelligibility, should have excellent acoustics. However LOC, which we are proposing as a measure of the ability to sharply localize and perceive proximity, is quite low except in the first row we tested. To bring the acoustic engagement we expect in theaters requires twice as much absorption. The difference in the sound quality of a spoken voice is dramatic. [6]

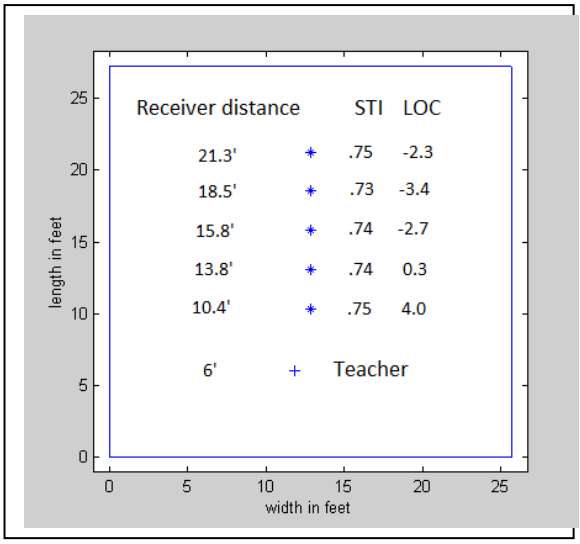

Figure 1: A 27'x25x'10' classroom model with an average absorption of 0.15 and a reverberation time of 0.34 seconds. STI is uniform and high throughout. But LOC, a measure of attention, is adequate only in the first four of five feet from the teacher.

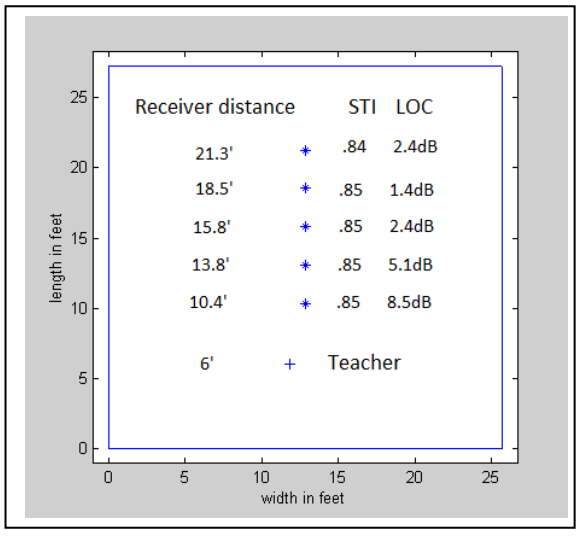

Figure 2: The same classroom model with an average absorption of 0.3, and an RT of 0.26. The number of seats with LOC greater than 2dB has gone up enormously.

1.4 Proximity and music

Most musical instruments also produce these pressure peaks, and they have the same attention grabbing power of speech if acoustics are favourable. In most halls the sound close to the musicians does have this quality, but many if not most seats do not. This is easy to demonstrate during a rehearsal. If you walk backwards away from a small ensemble like a string quartet while looking at the floor, at first each musician is sharply localized, and it is easy to tell which player played each note. But there is a distinct distance at which localization and separation suddenly disappears. We call this distance the Limit of Localization Distance, or LLD. The ability to localize and detect proximity seems to be all or nothing – at some point early reflections become too strong, and the sound collapses into a fuzzy ball.

The difference is not subtle, and there is very little difference in the LLD for different individuals. Binaural recordings made in front of and behind the LLD sound very different, even on loudspeakers. The LLD appears to be a property of the sound field, and how the ear and brain system has evolved to decode it. We propose that the presence or absence of proximity can be predicted by the LLD.

This presentation will discuss how these properties of sound affect classrooms, music venues, and sound reinforcement systems.

2 Proximity, Localization, and Phase

We have been studying what we now call proximity for more than ten years. We propose that it is the direct sound, the component of a sound-field that travels directly from a source to a listener,that contains the information needed to localize a source and to perceive its closeness. We believe that our ability to localize must result from abilities our ears and brains have evolved to distinguish the direct component of a sound field from reflections that follow.

How is this done? We find that the direct component and the following reflections have very different amplitude envelopes. The direct sound from speech and most musical instruments is created by a

repeating impulse, the opening of the vocal cords, a reed, rosin on a bow, etc. As a consequence once in each fundamental period the harmonics are forced to align in phase, creating peaks in the pressure amplitude at the fundamental period.

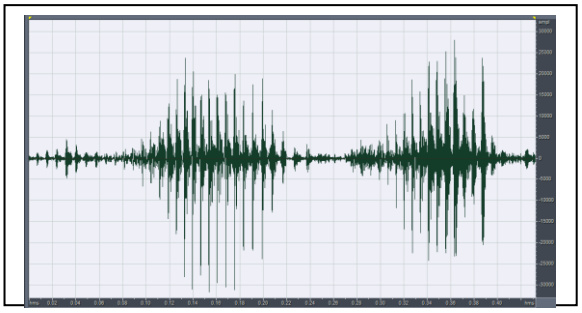

Figure 1: The syllable "one" first filtered through a 2kHz 1/3 octave 2nd order Butterworth filter, and then though a 4kHz filter. Note the prominent envelope modulation at the fundamental period, with peaks more than 10dB above the minima between peaks. Although the ear is not sensitive to the phase of the carrier at these frequencies, it is highly sensitive to these peaks. When they are present such a source can be sharply lateralized by interaural time differences (ITD) alone. If you listen to these filtered waveforms there is also a prominent perception of the fundamental tone. The horizontal scale is 0 to 0.44 seconds. (figures created by the author)

Licklider's autocorrelation mechanism in the basilar ligament not only explains our acute sense of pitch, in combination with the envelope it can explain our ability to detect ITD above 1000Hz, and enables the separation of several pitched signals into independent neural streams.

But reflections from all angles interfere with the phases of harmonics in the direct sound. The phase alignment is lost, and the sharp peaks at regular intervals become random noise. The ability to separate the direct sound from other signals, reflections, and noise is degraded, and the sense of proximity is lost.

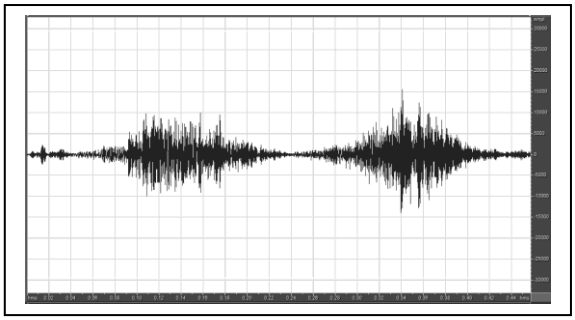

Figure 2: The same signals as figure 1, but altered in phase by a filter made from three series allpass filters of 371, 130, and 88 samples and allpass gains of 0.6. Notice that the peaks at the fundamental period have largely disappeared. When you listen to these signals no fundamental tone is heard. There is garbled low frequency noise instead.

We believe a primary mechanism for both source separation and the perception of proximity resides in the spiral ganglia just below the hair cells in the basilar ligament, and that the mechanism relies on the phase alignment of the upper harmonics of music and speech. The properties of the spiral ganglia are only just beginning to be studied. [7], [8] They are now known to be extremely fragile to continuous loud noise, and are thought to be vital to our perception of speech in noisy environments. Arrabi et al find that randomizing the phase of speech in the presence of noise can easily double the word error rate. [3]. Why has this mechanism and its effects on our perception of acoustics been unknown or forgotten for so long?

It is well known in the audio field that it is nearly impossible to tease out differences in sound quality without instant A/B comparisons. But to do this for concert halls requires that the sound in different halls and seats be exactly reproduced in a laboratory, with the ability to instantly switch from one sound to another. To detect the presence or absence of proximity we need a laboratory system that accurately re-creates the phase relationships between the harmonics of each individual instrument, along

with the reflection field that follows it. The current ISO 3382 measurement techniques were developed primarily through listening tests that used two channel anechoic recordings of orchestras played back through two loudspeakers on a concert stage. The phase information of each instrument was already lost, and proximity was not there to be heard. The same thing happens if a single instrument is reproduced through multiple loudspeakers, as happens with Wave Field Synthesis (WFS), or most Ambisonic systems.

Lokki [9] uses an electronic orchestra with a single pair of speakers for each instrument, one pointing up and one pointing out. Each pair plays independent anechoic recordings of orchestral parts. Separate three-dimensional intensity impulse responses from each instrument are recorded at the seat position under test. The direct sound from each instrument is reproduced through a single loudspeaker as close as possible to the original azimuth of the instrument, and the reflections are reproduced in a similar way without panning. His system is the first I have heard that reproduces the sense of proximity. Lokki's system is a big step forward. [10] But does it really reproduce the sound of a particular seat in a particular hall? How can we know?

3 Binaural technique and halls

We have performed a series of experiments that use a version of Lokki's virtual orchestra to study the effects of early reflections through binaural technology. With this method it is possible to take existing binaural impulse response data from the stage to a particular seat, and use it along with Lokki's anechoic recordings to synthesize the sound of a musical ensemble. It is also possible to compare the sound from an electronic ensemble to live musicians. Binaural technology, if it can be made accurate and convincing, can be a reference by which other methods of orchestral reproduction can be compared. But to work properly, it is absolutely necessary to have accurate equalization, all the way from the sound outside the microphone *to the eardrum* of the listener.

Experiments at IRCAM [11], as well as the author's work, has shown that if the equalization is correct at the eardrum external frontal localization can be achieved without head tracking. But to do this it is essential to measure the eardrum pressure directly or indirectly for each listener. For many years the author has been recording data and live music with probe microphones at his eardrums. Equalizing headphones with the same probes at the eardrums allows the author to play back these recordings with startling accuracy. The experience can be stunningly beautiful – recreating the experience of the performance better than any commercial recording technique.

Two of our three dummy head microphones are fitted with silicone castings of my own pinna and ear canals, and one is a standard Neumann KU-81. All of them are equalized for flat response from a frontal plane wave up to a frequency of about 6kHz. The author finds the difference between them to be minimal. The author's personal eardrum recordings are equalized the same way. This type of equalization makes the dummy head microphones and my own eardrum recordings similar in timbre to a high-quality studio microphone, but with a far different directivity. An additional advantage of this equalization is that our binaural recordings can sound natural when played on loudspeakers.

But do recordings from my ears played through headphones equalized for my ears work for other listeners? In general, the answer is no. But we have found that most of the variation between individuals comes from the transfer function from the headphone to the eardrum, which varies enormously between individuals. If we individually calibrate a pair of headphones to a particular listener the result with my personal recordings can be surprisingly robust. Almost all listeners report frontal localization and a realistic timbre. The presence or absence of proximity can be easily determined.

The most obvious method of matching these recordings a listener is to use probe microphones at the eardrums, and match the headphone response to the response at the eardrum to a frontal loudspeaker. The procedure is fast and accurate, but invasive.

We have developed a computer application that uses equal loudness methods similar to ISO 226 to match a pair of headphones to a listener using their own eardrums as microphones. This method is described in another preprint for this conference. The method results in the perception of accurate timbre, and almost always permits frontal localization without head tracking. Once fitted with individually equalized headphones most listeners find our binaural recordings to be realistic.

We believe that when you combine binaural recording and individual headphone equalization at the eardrum of the listener with Lokki's electronic orchestra the acoustic properties of a given seat in a particular hall can be accurately evaluated by different individuals. The process becomes an inexpensive and powerful tool for disentangling hall acoustics.

4 Measuring proximity

Our first measure, LOC, uses the ability to sharply localize sound as a proxy for proximity. We determined the threshold of localization of voiced speech in reverberation as a function of D/R and pre-delay. RT values of 1 second and two seconds were tested. For a 2s RT the threshold of localization can be as low as -17dB! We developed the function LOC to predict this data. The red and cyan lines show the accuracy of the fit. The LOC measure has been tested in real rooms and halls with useful results.

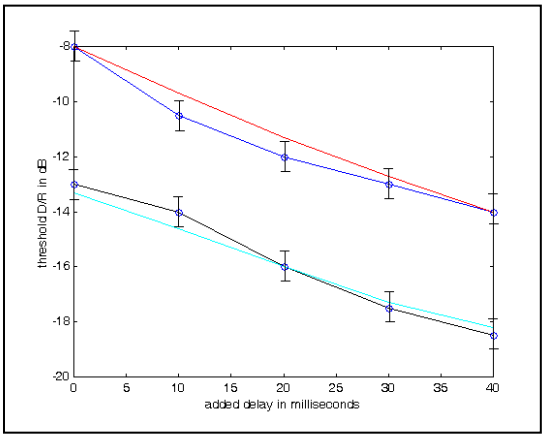

Figure 3: Data from localization experiments by the author and students of Professor Omoto in Fukuoka Japan. The data points indicate the threshold of localization of a male speaker counting from one to ten alternately at +-20 degrees azimuth in the presence of a three-dimensional reverberant field generated from exponentially decaying random noise. The top curve is for a 1 second reverberation time, the bottom curve is for a 2 second reverberation time. The red and blue curves are values calculated by LOC from the impulse responses used in the experiment.

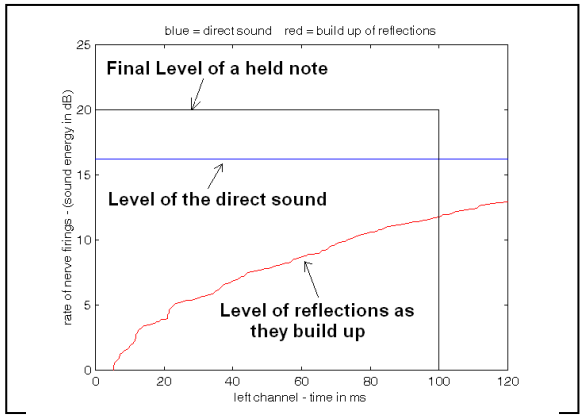

Figure 4: The diagram which explains the calculation of LOC. In this diagram we show data from an impulse response from Boston Symphony Hall seat R11. We assume a constant sound starts at time zero, and holds steady for at least 100ms. The level of the direct sound is shown by the blue line. As reflections arrive the reverberant level builds up. Its level is shown by the red line. The value of LOC is the ratio of the area under the blue line inside the 100ms box to the area under the red line, expressed in dB. The 100ms size of the box and the -20dB level for the base line of the box were set experimentally to fit the observed data.

LOC assumes the ear detects the direct sound using a mechanism that involves a ~100ms comb filter or autocorrelator. The mechanism searches for the regular peaks in the amplitude envelope that characterize the direct sound. The vertical axis can be understood as measure for the rate of nerve

firings in the basilar membrane, divided into two parts: firing from the direct sound, and firings from the reverberation. LOC in dB simply predicts whether the firings from the reverberation are less than or greater than the firings from the direct sound. If the direct sound firings are greater, the sound is localizable. [12], [13] Distinct proximity may require LOC values of +2dB or more. For the seat shown in figure 4 LOC is +9dB, C80 is 0.85dB, and IACC80 is 0.68. Proximity in this seat is very good.

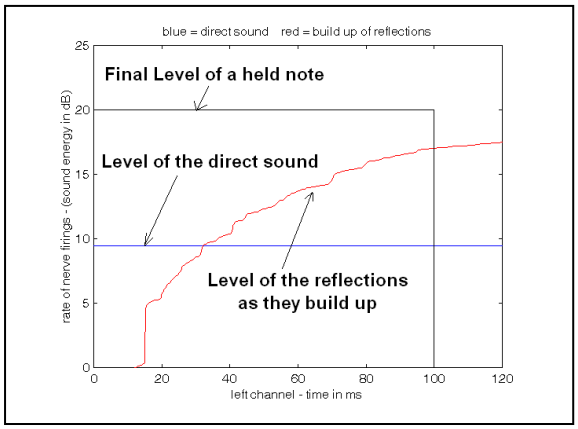

Figure 5: A similar graph for seat DD11. The sound in this seat is loud and fuzzy. Localization of individual instruments is impossible. This seat has C80 = -0.21, IACC80 = 0.2, and LOC = -1.2dB. If the strong lateral reflection from the side wall at 14ms is deleted in the impulse response LOC rises to above +2dB, and the sound improves markedly.

We are working on measures that use our hearing model to determine localization and proximity from live speech. They seem promising, but need a lot more work.

5 Conclusions

1. The human ability to instantly perceive that a sound source is "near" has been ignored by current acoustic science.

2. The neural mechanism and the physics by which it works depend on phase relationships of harmonics above 1000Hz. The mechanism not only provides distance information, it provides the ability to separate harmonic signals into independent neural streams.

3. There is substantial evidence that the perception of proximity, or "near" has consequences for the ease of comprehension, the ease of recall, and the focusing of attention.

4. The perception of proximity, and the ability sharply localize sound sources, and separate sound streams from competing streams and noise requires careful control of early reflections. Too many reflections coming too soon destroys the sense of presence, the ability to localize, and the ability to separate sources.

5. The ability to localize, and separate sources is predicted reasonably well from an impulse response by the acoustic measure LOC.

6. To the greatest extent possible sound systems intended for acoustic research should strive to preserve phase, and to have separate loudspeakers for each source.

References

[1] Blauert, J. (1983). *Spatial hearing*. Cambridge, Mass. MIT Press

[2] J. Licklider, (1951) "A duplex theory of pitch perception." *Experientia*, Vol VII/4 128-134 (1951)

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

[4] T. Piquado, (2010) "Effects of Cognitive Effort in Speech Comprehension and Recall in Younger and Older Adults." *PhD Thesis, Program in Neuroscience, Brandeis University*, (2010)

[5] T. Lokki, A. Southern, S. Siltanen, L. Savioja, "Studies of Epidaurus with a hybrid room acoustics modelling method." *The Acoustics of Ancient Theatres Conference Patras, September 18-21*, (2011)

[6] Links to sound files of the sounds in different seats can be found in the power point on www.davidgriesinger.com that was given in 2004 at the ASA convention in Indianapolis.

[7] S. Kujawa, M. Liberman, "Adding insult to injury: Cochlear Nerve Degeneration after "Temporary" Noise-Induced Hearing Loss." *Journal of Neuroscience*, 29(45):14077–14085 (2009).

[8] S. Kujawa M. Liberman, "Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss." *Hearing Research* http://www.sciencedirect.com/science/article/pii/S037859551500057X (2015)

[9] J. Pätynen, T. Lokki, (2011). "Evaluation of Concert Hall Auralization with Virtual Symphony Orchestra". *Building Acoustics*, ***18***, *349-366* (2011)

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

[11] Personal communication from Eckhard Kahle

[12] L. Beranek, "Concert hall acoustics: Recent findings" *J. Acoust. Soc. Am. 139(4*) (2016)

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