Audio Engineering Society

Convention Paper

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

# Pitch, Timbre, Source Separation and the Myths of Loudspeaker Imaging

David Griesinger[1],

[1] David Griesinger Acoustics, Cambridge, Massachusetts, 02138, USA
dgriesinger@verizon.net

## ABSTRACT

Standard models for both timbre detection and sound localization do not account for our acuity of localization in reverberant environments or when there are several simultaneous sound sources. They also do not account for our near instant ability to determine whether a sound is near or far. This paper presents data on how both semantic content and localization information is encoded in the harmonics of complex tones, and the method by which the brain separates this data from multiple sources and from noise and reverberation. Much of the information in these harmonics is lost when a sound field is recorded and reproduced, leading to a sound image which may be plausible, but is not remotely as clear as the original sound field.

## 1. INTRODUCTION

Horizontal sound localization by means of the Interaural Level Difference (ILD), and Interaural Time Difference (ITD) has been extensively studied. It is also well understood that vertical localization is provided by the pinna, which alter the timbre of high frequencies.

ITD, ILD, and timbre are physical attributes of the sound pressure at the eardrum of a listener or a dummy head microphone, and thus can be studied. But they are only a part of at least seven processes.

1. Sound pressure is detected and converted to nerve firing rates.
2. Firings from sound events are separated from noise and reverberation.
3. Each sound event is separated from others.
4. The timbre and direction of each event is determined.
5. Using timbre and direction cues, events from individual sources are assembled into independent neural streams.
6. The streams are interpreted for meaning.
7. The meaning is stored in long-term memory.

The order of the processes is important. In the past the separation process has been assumed to come after the identification of timbre and direction, and sometimes it may. But when there are simultaneous multiple sounds and/or noise and reverberation there is an advantage to performing source separation first. Otherwise timbre information from multiple sounds will overlap, and meaning will be impossible to determine.

How can sounds that overlap each other in each critical band be separated? The secret lies in their pitch and the phase relationships of their harmonics. As social animals the sounds which have the most meaning for us (speech and music) consist largely of tones with relatively low frequency fundamentals and lots of high order harmonics. We are known to perceive the pitch of such tones with very high acuity – for a musician one part in a thousand. This acuity is far beyond the capability of the mechanical filters of the basilar membrane. It must reside elsewhere in the ear/brain system, and it is probably very old from an evolutionary standpoint. It is likely such acuity evolved to aid source separation, as it enables us to filter sounds one from another and from noise.

### 1.1. Sound separation

The first section of this paper presents some examples of the effects phase relationships of upper harmonics have on the quality of sound. They also affect the ease with which sounds can be separated. Hopefully the reader will be able to hear these examples by clicking on the links. With headphones or near-field speakers the differences in sound quality are very obvious. However we have found that room acoustics in even a good lecture hall can sufficiently muddle phases to the point where the differences in these examples become inaudible, demonstrating the point that clarity can be quite fragile.

We propose a mechanism based on pitch detection which can separate sounds from noise and each other. The mechanism is based on the physics of information and known properties of hearing. A model of the mechanism predicts not only our abilities to perceive pitch, but also our ability to instantly perceive whether a sound is near or far, and our ability to sharply localize sounds in a soundfield that contains multiple sources and noise and reverberation.

We propose that clarity of both speech and music depends on source separation. When separation is possible sounds are perceived as close to the listener and demanding of attention regardless of their visual distance. They are easier and quicker to parse, which makes them easier to remember. If we have to use grammar and context to interpret speech in poor acoustics we will often not be able to remember what was said. Classroom instruction becomes a nightmare.

Because source separation comes early in the neural chain it affects a great many of our abilities, all of which appear to degrade in the same way when the soundfield becomes too confused, noisy, or reverberant. The mechanism we propose utilizes information encoded in the phase relationships between the upper harmonics of richly harmonic tones, which create strong modulations in the motion of the basilar membrane. These phases are randomized both by loudspeakers and by noise and reflections. The degree of randomization can predict sound quality and the direct to reverberant ration that makes source separation impossible. Phase relationships can be measured, and measures based on phase can give us new insight into the acoustics of auditoria, classrooms, and music performance spaces. They may also give us insight into loudspeaker sound quality.

### 1.2. Localization of sound with Loudspeakers

The second section of this paper examines the reproduction of sound images through loudspeakers. The ILD, ITD, and timbre cues that allow us to precisely localize sounds in natural environments are not correctly reproduced by loudspeakers. There are only two stable image positions in two channel stereo – the positions of the two loudspeakers. If we pan a signal to the center, we perceive a “phantom image” but only if we are precisely in the sweet spot. The center phantom image is independent of frequency, because both the ITDs and the ILDs are zero. But if a signal is panned half way between center and left the ILDs we measure at the listener’s ears are strongly frequency dependent. Frequencies below about 500Hz behave as we would expect, being perceived at about 15 degrees azimuth in a +-30 degree loudspeaker basis. But we have learned through studying source separation that in most rooms the upper harmonics of sounds yield the sharpest localizations in natural hearing, but these produce much larger ILDs at a listener’s ears than the ILDs of natural hearing.

The brain is forced to make a “best guess” of the location of the source based on an average over all frequencies. The resulting image – which we perceive as sharp, is typically at least 7 degrees further to the left than we would expect from a sine/cosine pan-pot. Its perceived position depends on the frequency content of the signal.

We think we perceive a sharp sound image, but the accuracy is poorer than with natural hearing. The situation is far worse outside a +- 45 degree horizontal loudspeaker basis. Horizontal localization in stereo only works because the HRTF functions that determine timbre are very similar. Outside the front – for example to the sides, rear or overhead, the HRTF timbre cues are sufficiently different that interpolation between source positions is not possible. The brain tends to pick either one loudspeaker position or another. In-between positions are too implausible to accept.

## 2. EXAMPLES OF CLARITY, DISTANCE, AND PHASE

The first half of the speech example below was made by recording my voice with a close microphone. The second half of the example was modified by convolution with the impulse response shown in figure one. The impulse response is all-pass, which means that the frequency response measured over an approximately 40ms time window is completely flat. But the phase response of this impulse randomizes the phases of harmonics above 1000Hz.

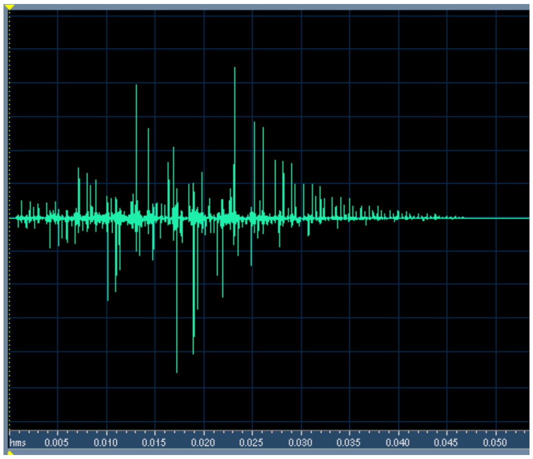

Figure 1 An all-pass impulse response with a total length of less than 50ms. The horizontal scale is in seconds.

Example of speech with intact phase followed by randomized phase

The change in sound quality is dramatic. When the phases are randomized the words are more difficult to understand and the voice sounds distant. The clear sound immediately grabs your attention, the unclear sound does not.

Try listening to this example from a distance – say ten feet. In many rooms the difference between the clear and unclear sections starts to disappear. Both sounds are perceived as distant because room reflections are randomizing the harmonic phases.

Although the change in sound quality is dramatic, there is no standard measure for this quality. Classrooms, lecture halls, and concert venues are specified and designed to meet clarity criteria such as C50 and C80. Both measures assume that reflections arriving sooner than 50ms are beneficial. In the above example C80 and C50 are infinite, implying perfect clarity. We have proposed a much better measure for clarity, LOC, which will be described below.

### 2.1. Sound separation by pitch

Information theory tells us that the number of bits per second a channel can carry is proportional to the bandwidth of the channel and the signal to noise ratio. The channels in the ear/brain system are the critical bands of the basilar membrane, and their bandwidth is roughly proportional to frequency. Frequencies at 1000Hz can carry roughly ten times the information as frequencies at 100Hz. Not surprisingly human hearing and speech exploit this basic physics.

Nearly all the information in speech is carried in frequencies above 1000Hz, the frequencies of the vocal formants. The dominant signals at these frequencies are harmonics of fundamentals with a definite pitch. It is the relative strength of these harmonics in different critical bands that determines the timbre of an instrument or the identity of a vowel.

It is not required for speech comprehension that the energy in these critical bands be composed of pitched harmonics. We can understand whispered speech. But if two people are whispering at the same time, or if the space is noisy, communication is impossible. The advantage of pitch is clear – but why is it so effective?

The author’s current work concentrates on the brain processes that enable our ears to separate simultaneous sounds into independent neural streams. At formant frequencies separation requires that sounds have a definite pitch, and multiple harmonics above 1000Hz.

Human pitch perception is different than what we would expect from the construction of the basilar membrane. We hear pitch as circular in octaves. A concert “A” at 440Hz is still a concert “A” if we play it at 220Hz or 880Hz. Circular pitch detection has an obvious advantage, in that harmonics of the fundamental do not need to be separately filtered and summed.

Remember that the information in speech is carried by the harmonics of the vocal fundamentals, and these harmonics are often above the frequencies that auditory nerves can fire. A typical critical band at formant frequencies contains five to ten of these harmonics, and they are not individually detectable.

But if their phases are intact they interfere with each other to create a strong modulation in the level of vibration at the frequency of the fundamental and the first few harmonics.

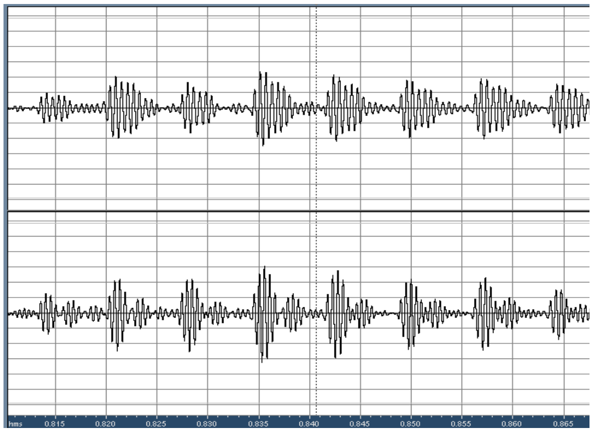

Figure 2 Top trace – the modulation pattern at 1600Hz of the syllable “two” when filtered by a critical band filter. Bottom trace – the same but filtered at 2000Hz. The carrier is visible, although the auditory nerves would not respond to it. They will respond with rate modulations proportional to the envelope of this signal, and these modulations are at the pitch of the fundamental. Note that the peaks of the modulation at the two traces are aligned in time.

When there are two or more sounds with different pitches at the same time, the modulations shown in figure 2 combine linearly, which means they can be separated by an appropriate (octave circular) pitch-sensitive filter. If we listen to the rectified and low-passed outputs of the filters shown in figure 2 it sounds pleasant. The fundamentals and harmonics are reproduced without distortion.

Once separated by pitch the amplitudes in each critical band can be compared to identify the timbre of the instrument or the identity a vowel. By comparing the separated signals between the two ears ILD and ITD can be determined for each event. Thus source separation, perceived distance, clarity of localization, and clarity of sound are ALL related to the physics of sound separation.

The regular, aligned modulations shown in figure 2 disappear when reverberation is added. If we listen to the outputs the sound is harsh and noise-like. This is the “mud” and distance we perceive when sound is unclear. The sound of the waveforms shown in figures two and three can be heard in the following link:

”One to ten” filtered at 1600Hz and 2000Hz, rectified – first clear and then garbled

In the first half of the example the fundamentals and first few harmonics are easily heard. In the reverberated section the sound is mostly noise.

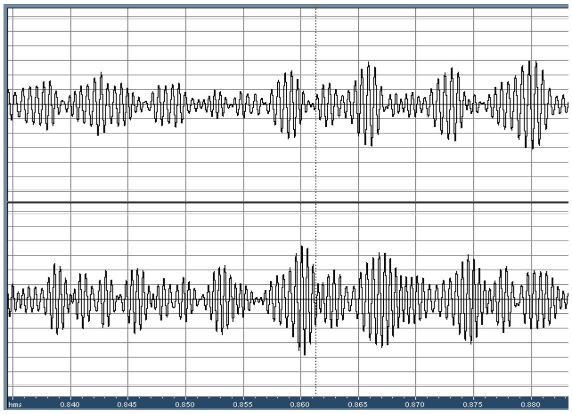

Figure 3 The same picture as figure 2, but with added reverberation. Note that the modulations are still present, but they are no longer at any particular period, and they are not aligned in time.

#### 2.1.1. Separation of monotone speech.

Broks and Noteboom (1983) [1] found that two simultaneous monotone speech signals in the same location can be separately understood if the pitch difference is only half a semitone, or 3%. The author finds 3% difficult. Here are some examples of simultaneous speech with a pitch difference of one semitone, or 6%.

We start with dry speech with full bandwidth at a pitch of C below middle C. Monotonone speech Then we filter away all frequencies below 1000Hz with a very sharp phase-linear filter. Filtered monotone speech. Note that the speech is actually easier to comprehend without the low frequencies.

Now we add a second voice at a pitch of C# below middle C. With dry acoustics these can be separately understood. Concentrate on listening to only the high pitches or the low pitches. The second voice is identical in timbre, which makes the task more difficult, but not impossible. Two filtered voices at the same time separated by one semitone.

But speech separation is only possible if the phases of the upper harmonics are unaltered. If they are randomized by acoustics or noise, separation and comprehension becomes impossible. As an example, we can convolve these examples with a measured binaural impulse response from a small auditorium. The measurement, made with an omnidirectional loudspeaker, was altered to raise the strength of the direct sound 6dB. Even so, he convolution severely reduces clarity. The speech is more difficult to understand and impossible to separate.

Speech at C in the room

Speech at C# in the room

C and C# together in the room

In a concert venue the ability to separate sources is often good in seats forward of a particular line. Just a few rows behind this line separation becomes impossible. As an example here are two binaural recordings of a live string quartet concert. The first is from row F, somewhat forward of the center of a 1500 seat shoebox hall. The second recording is of the same concert from row K, just five rows further back. The sound is quite different. In row F the instruments are clearly separated by timbre and localized, even though the group subtended an angle of only +-8 degrees. From row K the sound is far less clear, and is blended together into a sonic ball in front of the listener. The difference was difficult to perceive with eyes open, as visual localization takes precedence over auditory localization.

String quartet in row F

String quartet in row K

## 2.2. A measure for the threshold of localization

Data on the threshold for azimuth detection in the presence of reverberation was used to develop a measure based on a binaural impulse response for the ability to localize and to separate sources in a reverberant hall.

The measure is based on the idea that if we count the number of nerve firings (roughly proportional to the logarithm of the sound pressure) that result from the direct sound in the first 100ms, and compare that number to the number that result from reflections in the first 100ms, then the ratio of those two numbers predicts whether or not we will be able to localize a sound and perceive it as close to us. A ratio greater than 2 predicts implies good hearing, less than one predicts muddy sound. Details of the measure and Matlab code for calculating it can be found in reference [2].

Preliminary results from the measure in occupied and unoccupied halls have been surprisingly successful. It is hoped the measure will become more widely used as a predictor of clarity, ease of remembering, and the ability of a sound to hold attention.

### 2.2.1. Perception of envelopment

The goal of the ear/brain is to extract meaningful sound objects from a confusing acoustic field. To the brain reverberation is a form of noise. Where possible the brain stem separates direct sound from reverberation, forming two distinct sound streams: foreground and background. When separation is not possible reflections and reverberation are bound to the direct sound, and are perceived from the direction of the visual image, regardless of where they actually come from. When separation is possible reverberation is perceived as stronger and all around the listener.

Perceiving reverberation and envelopment is only possible when the direct sound can be separately perceived, and clarity of the front image is a vital part of this process. When the front image is muddy reverberation becomes a part of the front image, and all is perceived as a frontal ball of sound. Almost invariably a recording has a clearer front image than a typical concert seat, where “well blended” sound is all you can hear. It need not be so. With good acoustic design a concert seat can have better clarity and envelopment than any recording reproduced over loudspeakers.

## 3. LOUDSPEAKER REPRODUCTION

### 3.1 Localizing sounds in natural hearing.

It is well known that we localize sounds through the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) Experiments with sine tones show that ITD is not useful above 2kHz due to frequency limits on nerve firings, and that ILD loses accuracy below 1kHz as head shadowing decreases.

But high harmonics of low frequency fundamentals contain nearly all the information of speech, and provide timbre cues that identify musical instruments. When these harmonics are present we find that we can accurately localize tones above 2000Hz with both ILD and ITD. To understand our ability to localize speech and music we need to use signals that include harmonics! When harmonics are present our ability to localize can be extremely acute, +-2 degrees or better.

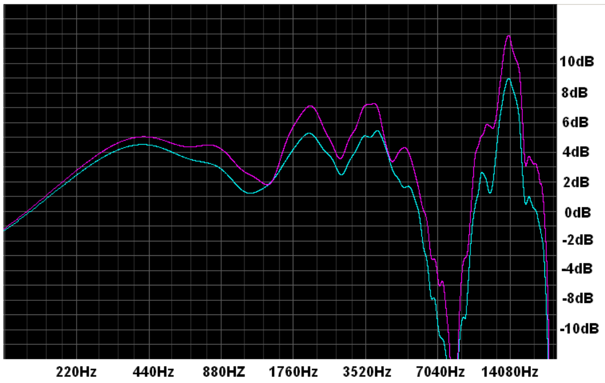

Figure 4 MIT Kemar data for 5 degrees azimuth. Note in the vocal formant range there is at least a 3dB frequency independent difference in ILD. If we assume a 1dB just noticeable difference (JND), this implies the ability to localize to ~1.5 degrees.

But with reproduction over loudspeakers the ILD is NOT frequency independent, but varies wildly as frequency rises from 500Hz to 4000Hz. In fact, if we set a pan-pot half way between center and left, filter speech or noise into 1/3 octave bands, and plot the perceived angle of the sound, we get the following result:

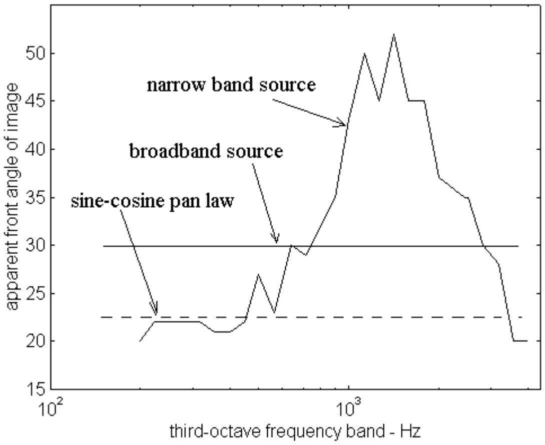

Figure 5 The perceived position of 1/3 octave filtered speech panned half-way between center and left (a level difference of 7.6dB).

Below 500Hz the perceived direction is what we expect from a pan-pot, but above this frequency the perceived angle moves strongly to the left. The result can be entirely predicted from the ILD. Sound from the left speaker that diffracts around the head is delayed sufficiently to interfere with the sound that travels directly from the right speaker. The sound pressure at the right ear is reduced, at about 1600Hz it is nearly reduced to zero.

When confronted with a broad band signal such as speech the brain must make a “best guess” as to the actual location of the image. A good approximation to the observed position can be found by weighting the incoming sound spectrum by an IEC equal loudness curve, and then averaging over the observations in figure 5. Having once made a decision about the location of a panned source the brain modifies it very reluctantly. You need to move a pan-pot nearly to the other side of the loudspeaker basis before a new position will be perceived.

Although the image is perceived as sharp, the accuracy and repeatability of the position is far poorer than in natural hearing. The only sharply localized positions in two channel stereo are left, center, and right, and that only in the sweet spot.

In classical music recording it is popular to combine time delay with amplitude panning, such as with the ORTF microphone technique. The combination is even more frequency dependent. The additional delay in the right loudspeaker from a source in the left increases the interference at the right ear and reduces the level further, and at lower frequencies. This increases the ILD, and moves the image even more strongly to the left.

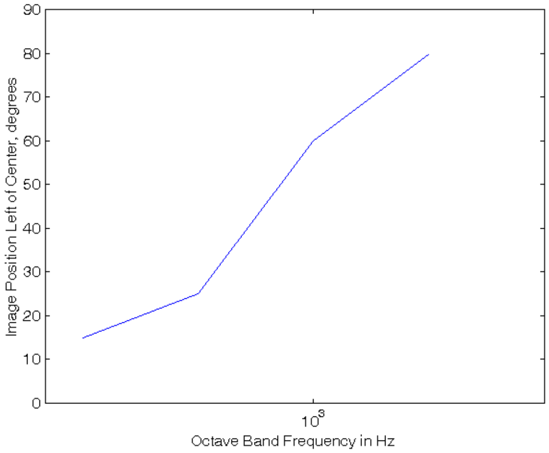

Figure 6 Apparent position of a sound image from a 200us delay plotted in octave bands.

The best that can be said about delay panning as used in the ORTF technique with cardioid microphones is that it widens an image that would otherwise be nearly monaural. Engineers compensate by bringing the microphones closer to the musicians. But at low frequencies the sound is still nearly monaural, particularly for reverberation. A far better solution is to replace the cardioid microphones with super cardioid or hypercardioid microphones, which deliver a wider image without delay panning, and pick up reverberation with little or no correlation.

The following links contain either noise or speech panned to the middle between center and left or right, assuming sine/cosine panning. The signals are filtered into third octave bands, so the variation of position with frequency can be easily heard either with loudspeakers or with headphones. The panning alternates between left and right, so the inherent tendency for an image to remain fixed in one position can be avoided.

amplitude panned noise in 3rd octave bands

amplitude panned speech in 3rd octave bands

200us delay panned noise in 3rd octave bands

It is quite interesting to listen to these examples with headphones or on a laptop with two loudspeakers. The speech example is particularly interesting. It illustrates that if phase is preserved fundamental pitches are much easier to hear from their harmonics than from their actual fundamentals. The fundamental of the speech is quite difficult to hear until at least the 500Hz band. As more harmonics enter each band the fundamental becomes clearer and clearer – all the way up to at least 8kHz.

In sum, localization in two channel stereo is an illusion based on very fuzzy data. The only stable locations are left, center, and right, and center is stable only in the sweet spot. Confronted with an image between center and left, or center and right, the brain must guess the location based on an average of conflicting cues. The result can be beyond the speaker axis. For example, try playing decorrelated (spaced omni) applause through a two channel system. If you are precisely in the sweet spot the applause will be perceived as all around you – well outside the loudspeaker basis. This is pleasant, but our perception of sharp images between center and left or right is an illusion generated by our brain’s desire for certainty, and it’s willingness to guess.

## 3.2 Localization over headphones

Localization of a panned image over headphones is not inherently frequency dependent, and a pan goes +- 90 degrees. But in practice localization is frequency dependent, as headphones do not couple identically to the left and right ears. It is very helpful if one is using headphones to make an on-location recording to play a series of 1/3 octave noise bands and adjust a 1/3 octave equalizer to center each band. It is even better to use noise bands to make an equal loudness curve for your favorite loudspeaker and then adjust an equalizer until your headphones give you the same equal loudness curve. In the following discussion I will assume this has been done – although this is highly unlikely.

When reproducing a binaural recording from a head that matches your own, with earphones matched to your ears as measured at your eardrums or with equal loudness, localization can be just as good as natural hearing. I have been amazed at the quality of a recordings made in a good concert seat in a great hall. They are more exciting and clearer than even the best stereo recording reproduced over loudspeakers. There are two reasons for the clarity. First, the localization accuracy of the human head is superior to the accuracy of typical main microphone arrays. Second, the head blocks reverberation from one side of the head from entering the other ear, and vice versa. This raises the direct to reverberant ratio over what would get with a microphone.

Careful use of close microphones and pan pots in combination with more distant microphones for reverberation can make a stunning recording when reproduced over headphones, as the localization is far better than when the same recording is reproduced over loudspeakers – which explains some of the success of mp3 players of all types. This is not true of common classical music microphone techniques. ORTF type techniques work better over headphones than they do over loudspeakers, but there are many differences from natural hearing between the ILDs and ITDs they produce

Engineers that monitor mostly over headphones sometimes swear by closely spaced omnidirectional microphones, which reproduce ITDs reasonably accurately but not ILDs. The result is peculiar over headphones, and disastrous over loudspeakers.

Coincident microphones of all types are not capable of the localization accuracy of natural hearing, and are always placed closer to the sound source than a typical listener. The 1dB JND for ILD of a coincident array such as a soundfield microphone is never better than about 3 degrees, and in practice is seldom better than 10 degrees.

### 3.3 Reproduction over multiple loudspeakers

Adding a center speaker to the front reduces the frequency dependent errors about a factor of two, and also makes a stable center image over a wide listening area. With hard panning to five frontal loudspeakers the imaging begins to approximate that of natural hearing.

Imaging to the sides, rear, and overhead requires hard panning between as many loudspeakers as possible, because the HRTF functions vary sufficiently that reproducing a direction through interpolation works poorly. The difference between a sound that emanates from a single loudspeaker and a sound that is panned somewhere between two or three speakers is quite large. Smooth panning in three dimensions can only be achieved by blurring the image.

Wave field synthesis systems depend on loudspeakers of finite size to attempt to re-create a two dimensional sound field. But the spatial aliasing of a typical array occurs just at the frequencies that are most important for source separation and localization in natural hearing. The usual result is that images are perceived as more distant than desired. First order Ambisonic reproduction suffers from the same problems as stereo, as frequencies in the formant range are reproduced with a power model and not a vector model, and a first order microphone has far less angular acuity than a human head. Third order Ambisonics begins to approach enough angular acuity to sound believable, but it needs to be reproduced with at least 5 frontal loudspeakers.

## 4. CONCLUSIONS

This paper defines “Clarity” as the ability to quickly and easily detect the timbre and meaning of multiple simultaneous sound sources in a noisy and reverberant sound field. The ability depends on source separation, which in turn depends on our amazing abilities to separate sounds by the pitch of their upper harmonics.

The pitch information in upper harmonics is carried exclusively by phase relationships between adjacent or nearly adjacent harmonics, and these phases are randomized by reflections and noise. The degree of randomization in the first 100ms becomes a measure for Clarity that can be useful in acoustic design.

That upper harmonics are so important in natural hearing gives insight into the process of image formation from loudspeaker arrays, which is poorer than natural hearing. The reason we always put microphones closer to musicians than we would listen is a consequence of the inaccuracies of recording and playback, not some magical phenomenon in the brain.

## 5. REFERENCES

[1] Bregman, A. “Auditory Scene Analysis” p560.

[2] Griesinger, D. “The Audibility of the Direct Sound as a key to Measuring the Clarity of Speech and Music” Presented at the Dublin Conference of the IOA, May, 2011. IOA Paper

[3] Griesinger, D. “Stereo and Surround panning in Practice” AES preprint Stereo and Surround Panning in Practice