Appendix

- Slides cut from the original presentation. They may – or may not, be helpful to understanding some of the concepts.

Number of taps

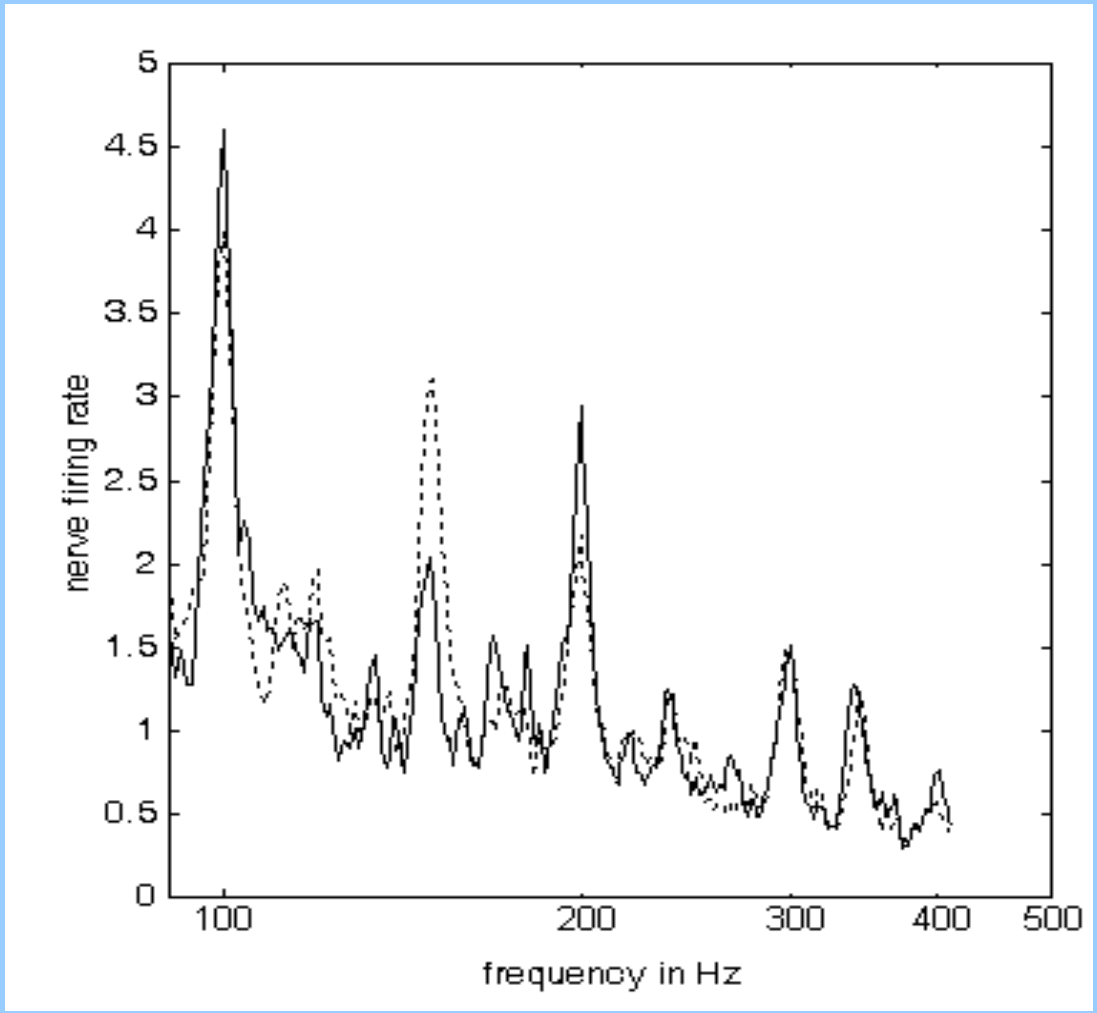
- In our model the number of taps for each frequency is simply the number of taps that fits in the delay line given the period we wish to measure.
  - If W is the length of the delay line in seconds, and f is the frequency in Hz, then N, the number of taps is:
- $N(f) = W*f$
- In our model the output of each summing neuron is divided by N, so all frequencies have roughly the same amplitude.

- In a comb filter of this type, the frequency sharpness of each summing node (full width at half maximum - fwhm) is just the frequency divided by N.
- $fwhm = f/N$
- In practice the sharpness is higher, as the input to the filter is not sinusoidal, but sharply peaked at the frequency of the fundamental.
- Our model achieves a frequency discrimination (+-3dB) of 1%.

Detection of harmonics and sub-harmonics

- Comb filters are sensitive to harmonics of the tap period.
  - The full width at half maximum is sharper for the harmonic than for the fundamental.
  - Thus a modulation that contains harmonics will be more accurately detected.
- A tap sequence intended to detect 100Hz will also detect 200Hz and 300Hz.
  - This property produces what appear to be sub-harmonics of the input modulations, since a 200Hz modulation will also excite the 100Hz tap sequence.
- This artifact has interesting consequences for harmony, as the pitch pattern produced by both major and minor triads has a strong output at the root frequency
  - Regardless of the inversion of the notes in the triad.

Minor triad in two inversions



Solid line: output of the pitch detector with a minor triad, 200Hz, 235Hz, and 300Hz.

Dotted line: The same triad with the fifth lowered by one octave.
200Hz, 235Hz, and 150Hz.

Notice the acuity of the pitch detection is better than 1%, or 1/6th of a semitione.

Each peak represents a separate neural data stream, which can be further analyzed for timbre and azimuth.

Parameters

The model contains several parameters:

- 1. The choice of a log-linear model for the hair cells, not a fully logarithmic detector. (the choice is clear…)
  - 1600,2000Hz bands after hair cell: Source - Logarithmic Log-linear



- 2. The choice of 10ms as the response time of the logarithmic adaptation.
  - This causes the onsets of sounds to have a higher amplitude than the body of the sound, which is probably an advantage
- 3. The choice of a 100ms window for the frequency filter
  - This choice is consistent with earlier work on localization.
- 4. The use of equal weighting for all the taps
  - Chosen for simplicity. Further work is needed. But equally weighted taps work rather well…

Why pitch is robust in the presence of reverberation

- Although pitch acuity is reduced by reverberation, reverberation is not capable of completely scrambling the phases of the harmonics.
- Some 1/3 octave bands will still contain significant modulation – but which have modulation, and which do not, is random, and varies from note to note.

- If we are relying on the amplitude of the modulations in each band to determine timbre – then timbre will be scrambled.
- In fact the major perception of a distant source is that the timbre has become muddy.

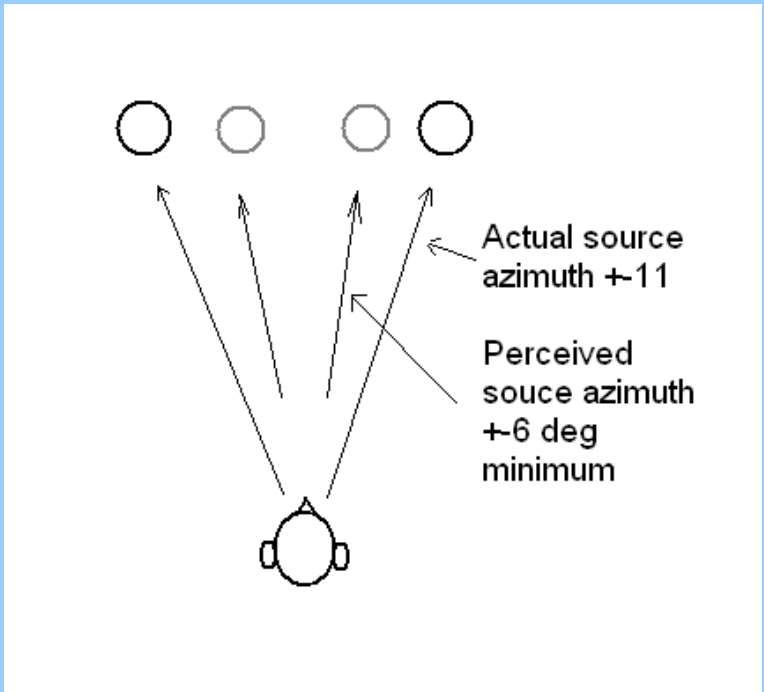
What is "Auditory Engagement"

- "Engagement" is the perception that you are not just watching a scene from distance, but present in the middle of it.
  - Thus lack of distance is a critical component of presence.
- Auditory engagement is the perception that you are acoustically close to the sound sources.
  - Distance is perceived directly through harmonic coherence – but experiments to directly measure it with subjects are difficult. However it correlates both with the ability to *localize* sound sources, and the perception *presence*, or musical clarity.
  - To perceive presence you must be able to localize sound sources nearly all the time,
  - and be able to distinguish them from one another nearly all the time.

- Clear localization and the ability to hear most of the notes are key components of audience engagement.
  - Although particularly important in drama and opera, it should be (and often is not) a part of the emotional experience of music.
  - Being able to hear all the notes and localize the players draws the audience into the performance. They don't just watch it.

- This view of clarity is different from the one that equates clarity with intelligibility. Perhaps we need a new word for it.

Experiment for threshold of Azimuth Detection in halls



A model is constructed with a source position on the left, and another source on the right

Source signal alternates between the left and a right position.

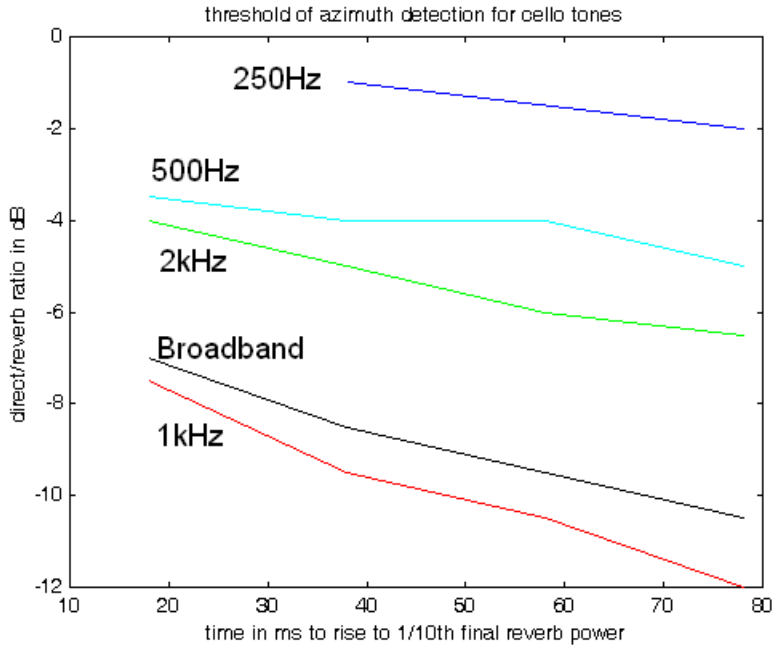
When the d/r is less than about minus 13dB both sources are perceived in the middle.

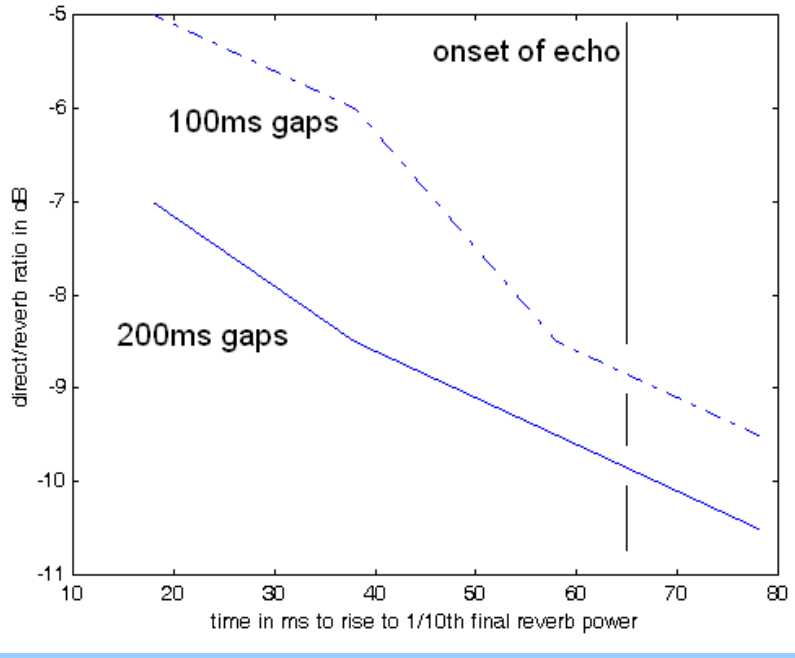
Subject varies the d/r, and reports the value of d/r that separates the two sources by half the actual angle.

This is the threshold value for azimuth detection for this model

(Above this threshold the subject also reports a decrease in subjective distance)

Threshold for azimuth detection as a function of frequency and initial delay





As the time gap between the direct sound and the reverberation increases, the threshold for azimuth detection goes down. (the d/r scale on this old slide is arbitrary)

As the time gap between notes increases (allowing reverberation to decay) the threshold goes down.

To duplicate the actual perception in small halls I need a 50ms gap between notes.

An important caveat!

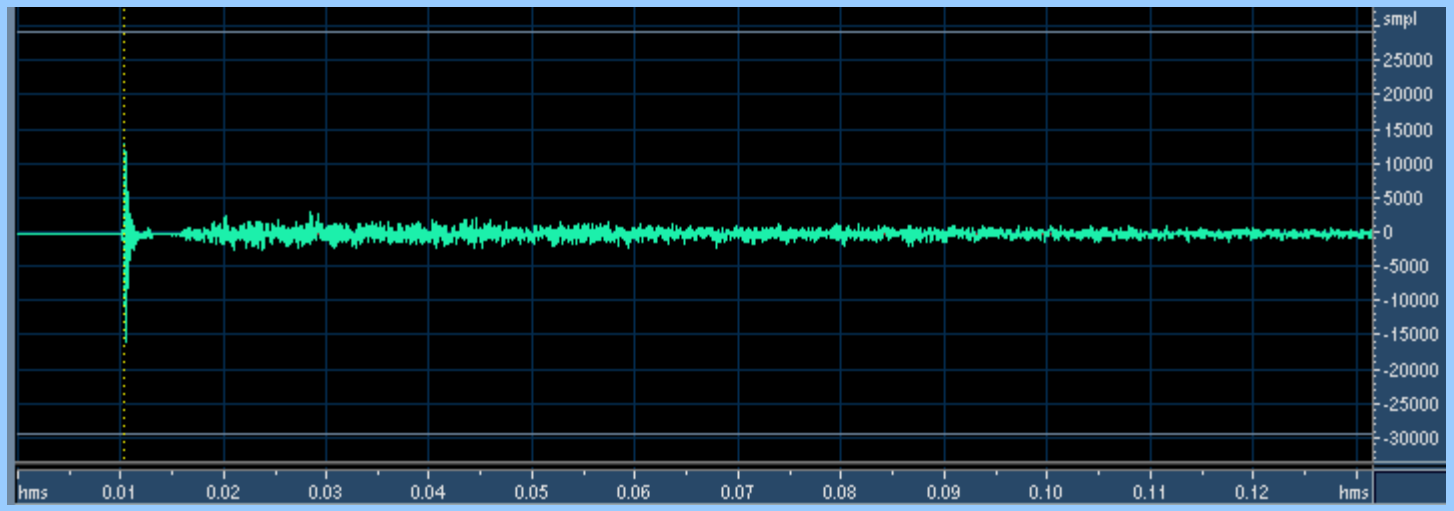
- All these thresholds were measured without visual cues
- The author has found that in a concert (with occasional visual input) instruments (such as a string quartet) are perceived as clearly localized and spread.
- When I record the sound with probes at my own eardrums, and play it back through calibrated earphones the sound seems highly accurate, but localization often disappears!
  - Without visual cues when the d/r is below threshold the individual instruments are localized and spread when they play solo, but collapse to the center when they play together.
  - My brain will not allow me to detect this collapse when I am in the concert hall – even if I close my eyes most of the time!
  - With eyes closed it is more difficult to separate the sounds of the individuals, such as the second violin and the viola. This difficulty persists in the binaural recording.

Localization

- For this paper we assume sound sources are localized by the direct sound.
  - In some cases localization is aided by early reflections – but these vary strongly from seat to seat, and are too complex to consider here.
- For localization to be successful the direct sound must be perceived.
  - Prompt strong reflections can – and do – mask the direct sound.
- Let's propose that the brain detects the loudness of – and the presence of – sounds by integrating nerve firings over a period of time.
  - If the integrated nerve firings from the direct sound exceed the integrated nerve firings from the reflections inside this time window, the direct sound will be perceived – and localized.
- We can calculate the threshold of perception by double integrating the impulse response over a fixed time window.

The ear perceives notes – not the impulse response itself.

- Here is a graph of the ipselateral binaural impulse response from spatially diffuse exponentially decaying white noise with an onset time of 5ms and an RT of 1 second. This is NOT a note, and NOT what the ear hears!



D/R = -10dB

RT = 2s:

C80 = 3.5dB
C50 = 2.2dB
IACC80 = .24

RT = 1s:

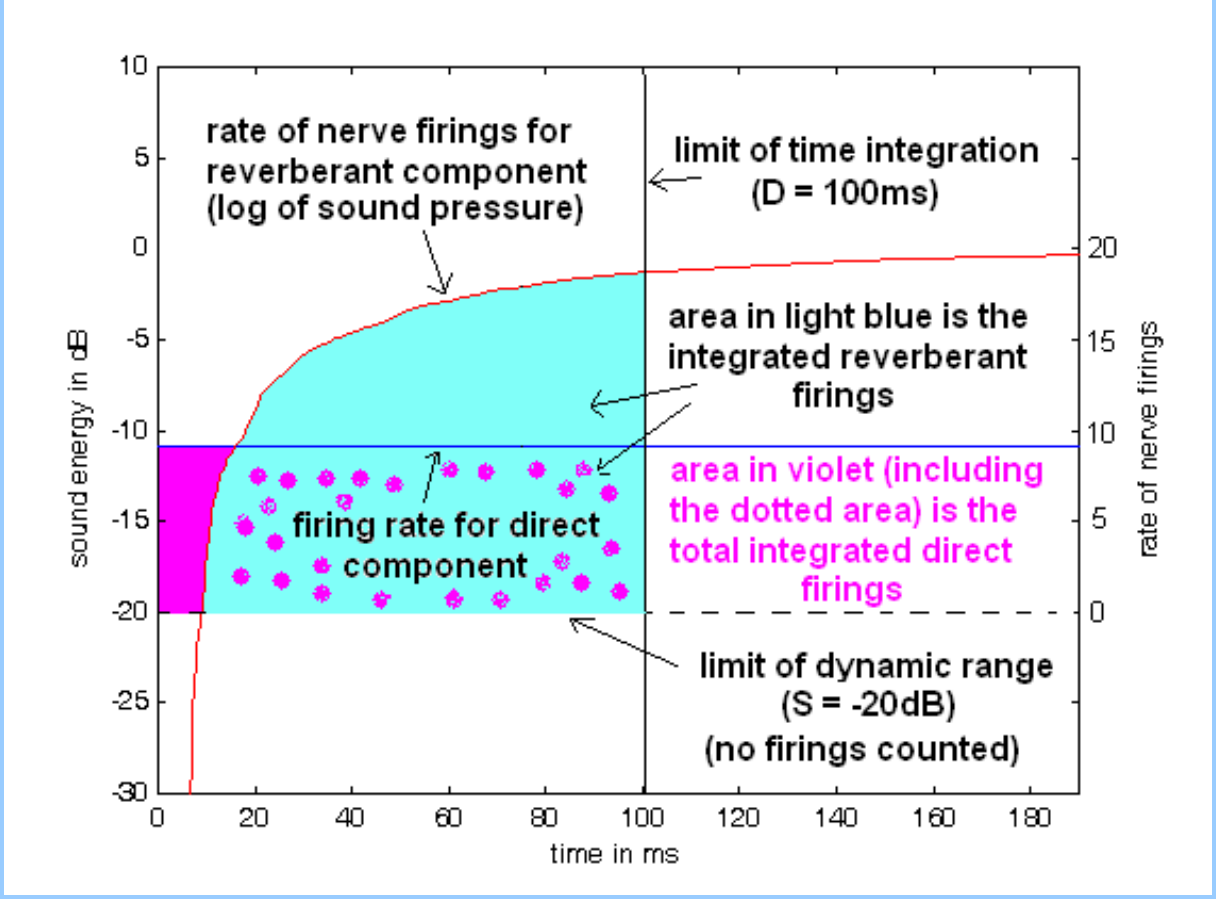
C80 = 6.4dB
C50 = 4.1dB
IACC80 = .20

- To visualize what the ear hears, we must convolve this with a sound.
  - Let's use a 200ms constant level as an example.

- The nerve firings from the direct component of this note have a constant rate for the duration of the sound.

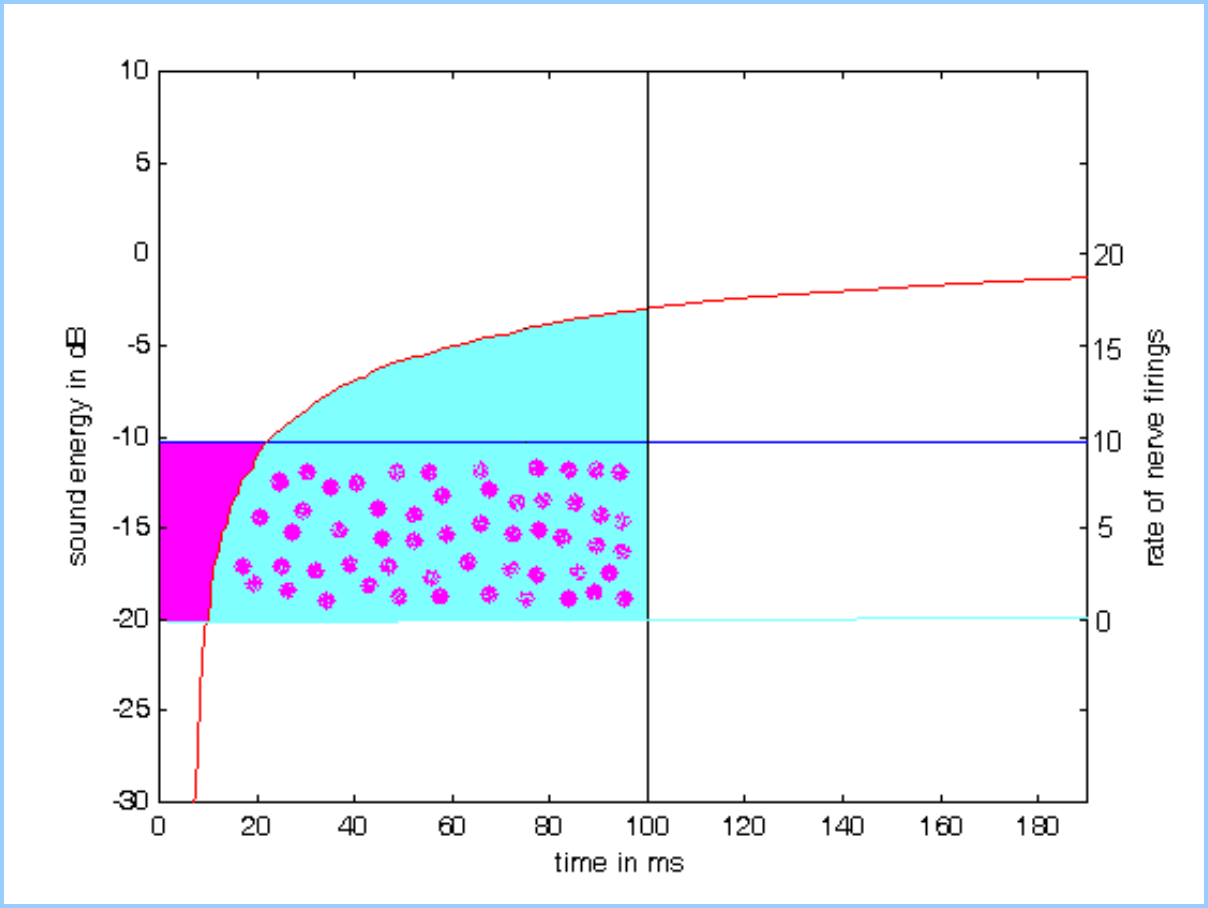
- The nerve firings from the reverberant component steadily build up until the note ceases and then slowly stop as the sound decays.

Direct and reverberation for d/r = -10dB, and RT = 1s



The blue line shows the rate of nerve firing rate for a constant direct sound 10dB less than the total reverberation energy. The red line shows the rate of nerve firings for the reverberation, which builds up for the duration of the note. The black line shows a time window (100ms) over which to integrate the two rates. In this example the area in light blue is larger than the area in pink, so the direct sound is inaudible.

Direct and build-up RT = 2s



If we hold the d/r constant, when the reverberation time is two seconds it takes longer for the reverberation to build up, so the light blue area decreases, while the pink area stays constant. This makes the direct sound more audible. In a large hall the time delay between the direct sound and the reverberation also increases, further reducing the area in light blue. The direct sound would be even more audible.

Equation for Localizability – 700 to 4000Hz

- We can use this simple model to derive an equation that expresses the ease of perceiving the direction of direct sound as a decibel value. *p(t)* is the sound pressure of the ipselateral channel of a binaural impulse response. With the previous simple assumptions, we propose the threshold for detection would be 0dB, and clear localization would occur at a localizability value of +3dB.

- Where *D* is the window width (~ 0.1s), and *S* is a scale factor:

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

S is the zero nerve firing line in the previous two slides. It is 20dB below the maximum loudness. *POS* means ignore the negative values for the sum of S and the cumulative log pressure.

- Localizability (LOC) in dB =

$$S - 1.5 + 10 * \log \int_{0}^{.005} p(t)^2 dt - (1/D) * \int_{0}^{D-.005} POS\left(S + 10 * \log \int_{.005}^{\tau} p(t)^2 dt\right) d\tau$$

- The scale factor *S* and the window width *D* interact to set the slope of the threshold as a function of added time delay. The values I have chosen (100ms and -20dB) fit my personal data. The extra factor of +1.5dB is added to match my personal thresholds.

Some explanation of the equation

- The equation as written in the previous slide simply calculates the ratios of the pink and blue areas shown in the previous pictures.
- The first integral on the left in LOC is the “pink” area – the sum of the nerve firings for the direct sound. This area is the product of the normalized sound pressure times the length of the window D.
  - However here we have divided through by D – so this factor is not shown.
- The next two integrals represent the total nerve firings for the reverberation – the “blue” area.
  - Since we have divided by D, a factor of 1/D is included at the beginning.
- The second of the two integrals is the physical sum of the sound pressure that would exist if the impulse response was convolved with a steady excitation. The first integral finds the area under this curve. In the second integral we have excluded the direct sound – assuming this will be in the first 5 milliseconds.
- The limits of the integrals have been adjusted to account for this exclusion. Thus the second integral goes from .005 seconds to the end, and the first integral is from zero to the window width minus .005.
- I have included the -1.5dB adjustment for my personal thresholds.

Matlab code for LOC


```
% load in a .wav file containing a binaural impulse
response – filter it and truncate the beginning

upper_scale =20;  % 20dB range for firings
% proposed box length
box_length = round(100*sr/1000);  % try 100ms
early_time = round(5*sr/1000);

D = box_length;       %the window width

ir_left = data1;   % the binaural IR
ir_right = data2;

clear data1 data2  % filter the Irs
wb = [2*1000/sr 2*4000/sr];
[b a] = ellip(3,2,30,wb);

ir_left = filter(b,a,ir_left);
ir_right = filter(b,a,ir_right);
clear data1 data2
wb = [2*1000/sr 2*4000/sr];
[b a] = ellip(3,2,30,wb);

ir_left = filter(b,a,ir_left);
ir_right = filter(b,a,ir_right);

for il = 1:0.1*sr
                    if abs(ir_left(il)) > 500
                                        break
                    end
                    if abs(ir_right(il)) > 500
                                        break
                    end
end

ir_left(1:il) = [];
ir_right(1:il) = [];
```

```
% ir_left is an ipselateral binaural impulse response,
%truncated to start at zero and filtered to 1000-4000Hz.
% early_time is 5ms in samples, D is 100ms in samples.

% here starts the equation on the slide:

S = 20-10*log10(sum(ir_left.^2));

early = 10*log10(sum(ir_left(1:early_time).^2));

% first integral is a cumsum representing the build up in
%energy when the IR is excited by a steady tone:

ln = length(ir_left);

log_rvb = 10*log10(cumsum(ir_left(early_time:ln).^2));

% look at positive values of S+log_rvb  only

for ix = 1:ln-early_time
                    if S+log_rvb(ix) < 0
                                        log_rvb(ix) = -S;
                    end
end

LOC = S-1.5+early -(1/D)*sum(S+log_rvb(1:D-early_time))
```

Use of the localization equation

- Just as RT or C80, LOC uses a measured impulse response as an input, with the direct sound starting at time zero. This is the only data a user needs to supply.
  - The measure is calibrated for a front facing binaural impulse response.
    - An omnidirectional impulse response will give lower values of LOC for the same seat position, due to the lack of head shadowing.
- The localization equation appears more complex than most current measures for room acoustics, but it has a simple, physiologically based interpretation.
  - It is the ratio in dB of the number of nerve firings received by the brain from the direct sound in a 100ms window, divided by the number of nerve firings received from all reflections in the same time period.
  - It contains three experimentally based parameters: the window width *D*, the dynamic range of the nerve channels *S,* and the time window for separating direct sound from reflections (*5ms*). These parameters are not intended to be adjustable without further experimental work.
  - Matlab code for calculating LOC is simple, as can be seen above.

Interpretation of LOC

- LOC was developed and verified as a method for predicting when a sound will be accurately localized when the direct sound is much lower in total energy than the sum of all reflections.
- Like C80, IACC80, and similar measures, LOC is based on a time window that begins with the onset of the direct sound.
  - In practice, syllables or notes that will be affected by any of these measures will depend on the rise time (onset time) of the sound.
  - If the sound starts gradually the precise moment of onset becomes indeterminate, and separating direct sound from reflections becomes impossible.
  - Thus LOC – and other such measures – are accurately predictive only for signals with sharp onsets.
  - Additionally, if the direct sound from a note or syllable is masked by reverberation from a previous sound, the direct sound will not be audible.
- LOC predicts the audibility of the direct sound for a syllable or note with a rapid rise-time when there is sufficient freedom from masking from previous sounds.
  - Although musical signals often do not meet these criteria, in practice there are enough occasions that do meet the criteria that the LOC equation is useful.
- Remember that for the purposes of this talk Localization is only a proxy for the main goal – predicting when the direct sound is sufficiently audible to produce *engagement*.
  - Preliminary results suggest LOC achieves this goal.

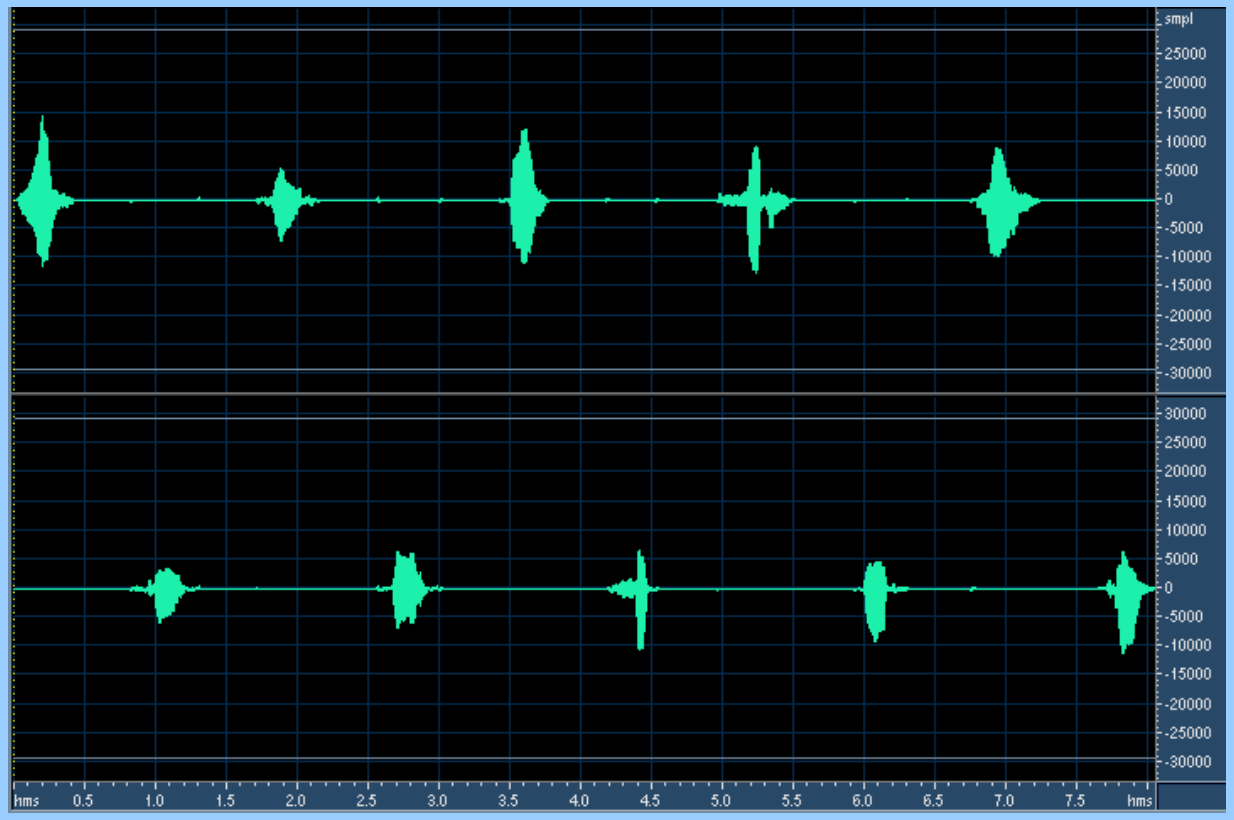
Localization Equation Setup

- The Localization Equation was developed and tested using binaural impulse response generated using the author's own HRTFs.
  - The source position was 15 degrees to the left (and right) of center. Only the ipselateral channel was analyzed.
  - Male speech alternated from left to right with a time gap of 400ms, to allow for complete decay of the reverberation between each word.
  - The reverberation was generated using an independent decaying noise signal convolved with each of 54 HRTFs spaced equally around the listening position.
  - The HRTFs were equalized so that the azimuth zero elevation zero HRTF was flat from 40Hz to about 4kHz. The elevation notch at 7.8kHz was not equalized away, but was left in place.
  - Playback was done through headphones equalized to match a loudspeaker placed in front of the listener – again not equalizing the 7.8kHz notch from the listener's frontal HRTF of the loudspeaker.
- Because my data show that the perception of both localization and near/far is mostly a high frequency phenomenon, the impulse response was bandpass filtered between 700Hz and 4000Hz before being analyzed for localization.
  - If a measured binaural impulse response is used as an input, care should be taken to insure the dummy head is equalized as described above.
  - Because of the importance of upward masking in localization, if the low frequencies in the room signal are significantly stronger than those in the frequency range from 700 to 4000Hz, localization is likely to be poorer than the equation would predict.

Comments on LOC

- LOC is based on the *LOG* of the build-up of reverberant energy.
  - This follows directly from the physiological model.
  - Current measures integrate the sound energy rather than the log of sound energy. But our physiology works differently. One of the consequences is that reflections that arrive early have more influence than reflections that arrive later.
    - As energy builds up additional reflections are not counted as strongly.
  - Reflections later than 100ms are ignored in calculating LOC.
- This is very different from C80 or C50, which count the earliest reflections a part of the direct sound, and compare the energy sum to the energy sum of all the later reverberation.
  - In a small hall most of the energy arrives before 80ms regardless of the relative strength of the direct sound, so C80 and C50 are usually high.
  - But small halls can have high C80 or C50, poor localization, and a lack of clarity.
- LOC depends strongly on the delay between the direct sound and the build-up of the reverberation.
  - late reverberation does not impair localization of short notes.
  - The principle difference between the localizability in small halls and large halls is the rate at which reflected energy builds up after the start of a note.
- LOC is NOT related to EDT – even if Jordan's original definition of EDT is used.
  - EDT is relatively independent of the initial time delay
  - When D/R < -10dB, EDT and RT are the same, as there is insufficient direct sound to be detected in a reverse integrated impulse response.
- LOC correlates with IACC80 – but IACC is not sensitive to medial reflections.
  - IACC is sensitive to the sum of reflected energy – not the log of energy, and thus is insensitive to when the reflections arrive

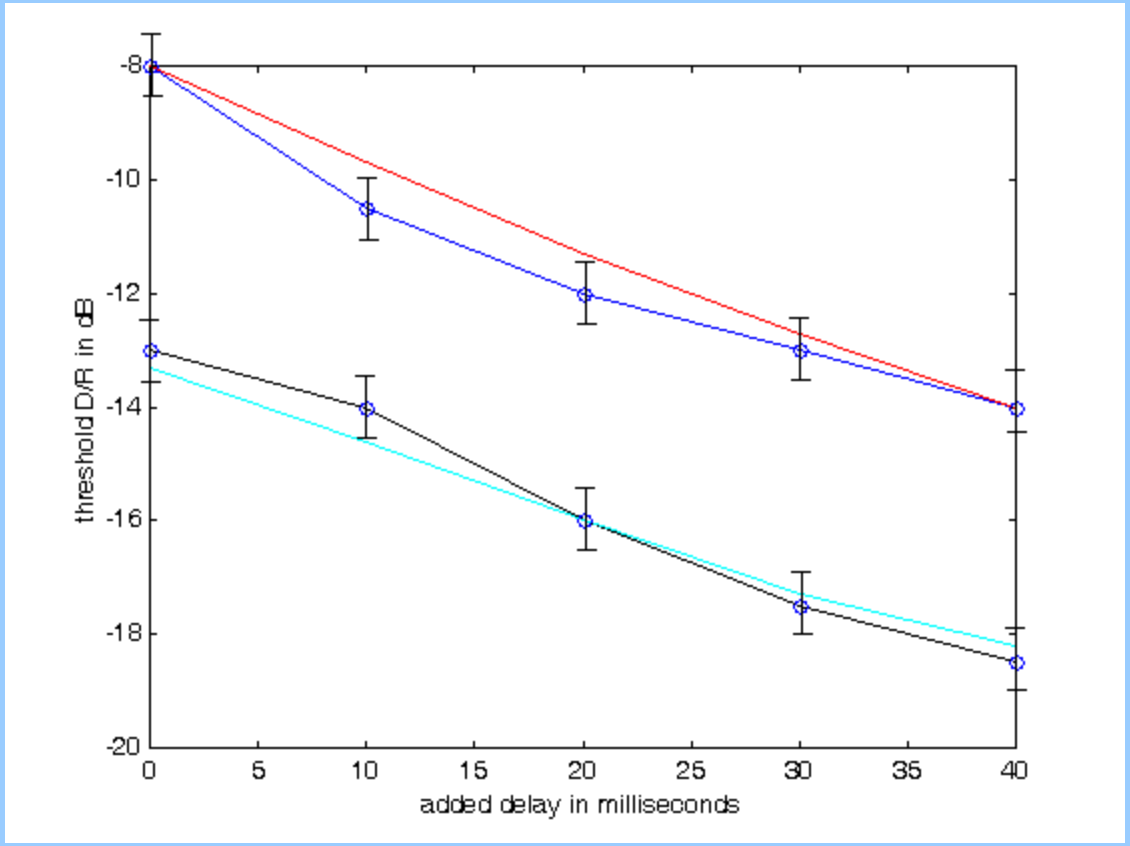
Tests with speech



A speech signal was convolved with a pair of binaural impulse responses, such that the sound appears to come from +-15 degrees from the front.

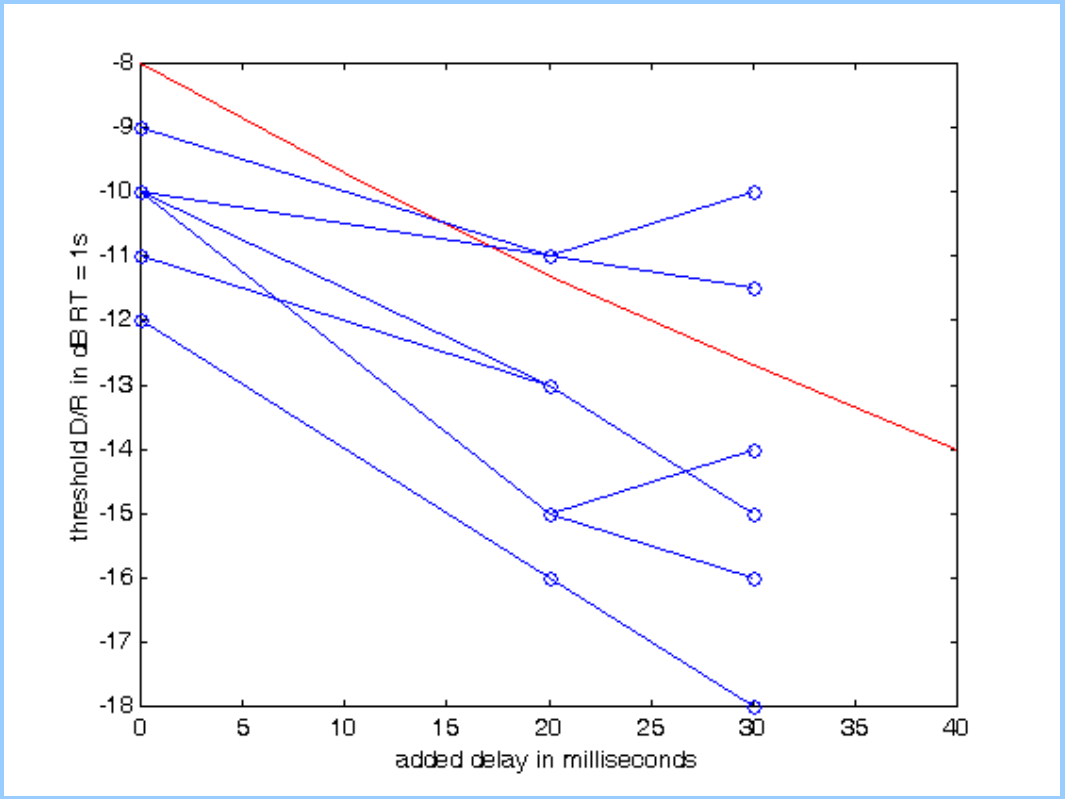
Then a fully spatially diffuse reverberation was added, in such a way as the D/R, the RT, and the time delay before the reverberation onset could be varied.

Broadband Speech Data



Blue – experimental thresholds for the alternating speech with a 1 second reverb time. Red – the threshold predicted by the localization equation. Black – experimental thresholds for RT = 2seconds. Cyan – thresholds predicted by the localization equation.

Threshold Data from Other Subjects – 1s RT

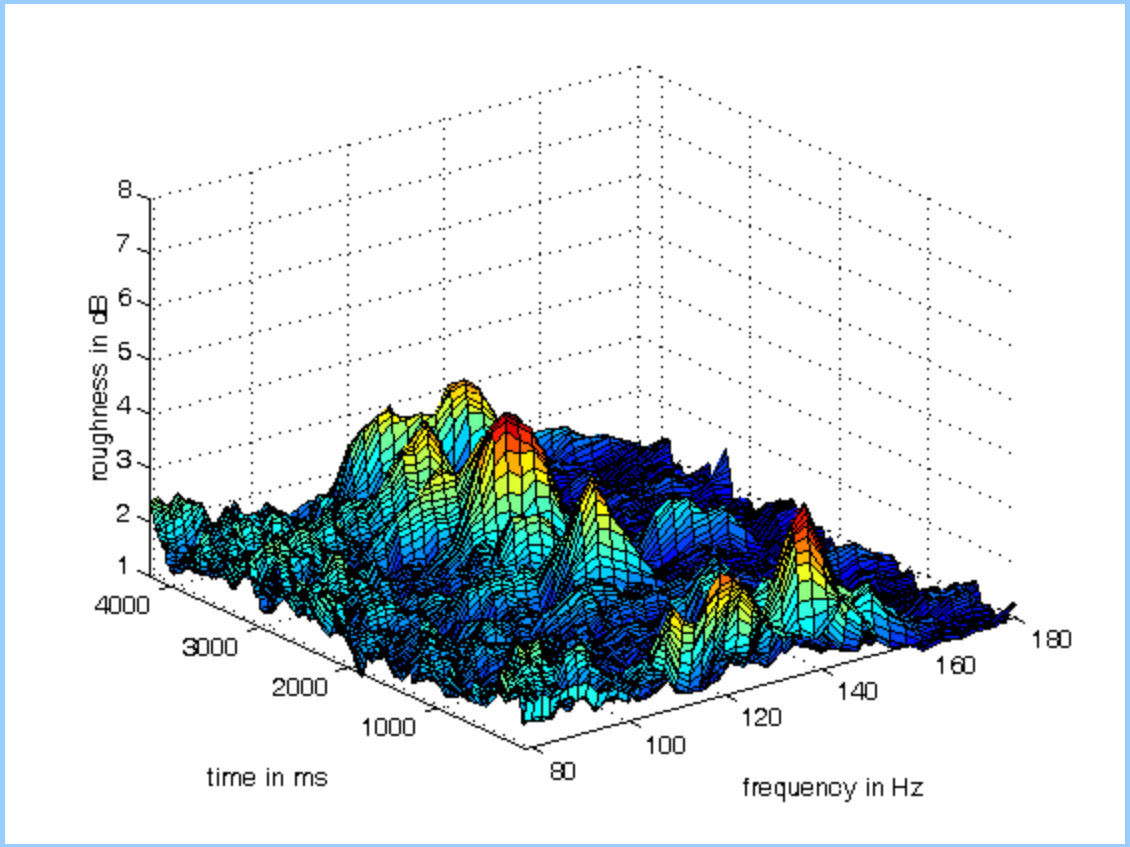


Blue – new data using absence of any localization as a criterion for threshold.

Red – the author's previous data based on a half-angle criterion.

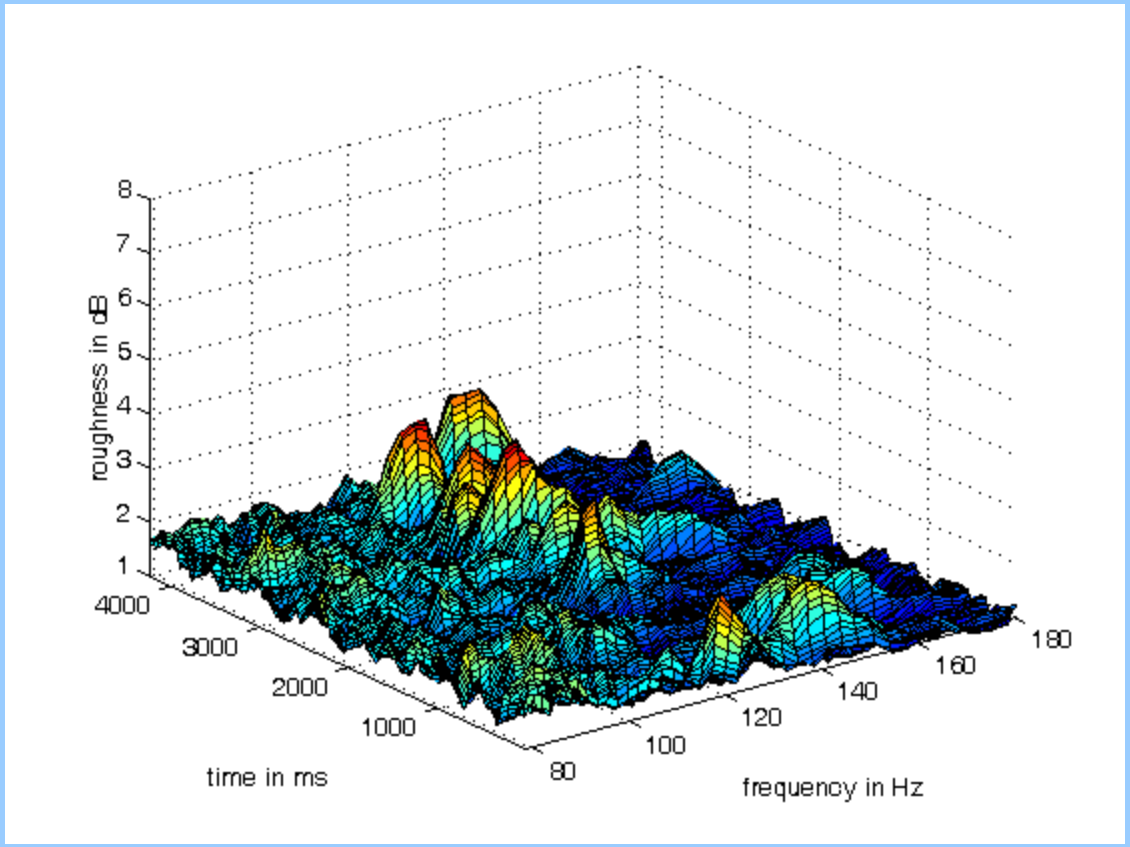
- Seven subjects participated in a threshold experiment at Kyushu University.
  - In these experiments the threshold was defined by the extinction of localization , not by the reduction of angle by a factor of two.
  - Consequently the thresholds are lower than they were in my previous experiment, and they have more variation.
- However, the data is consistent to within 3dB

Analysis Add reverb at 2s RT -10dB D/R



Two second RT at -10dB D/R does not seriously interfere with pitch.

Add reverb at 1s RT -10dB D/r



RT = 1s is much more damaging to harmonic coherence in the 100ms window.

An existing small hall – pictures

Note the highly reflective stage and side walls, deeply coffered ceiling, and relatively low internal volume per seat.

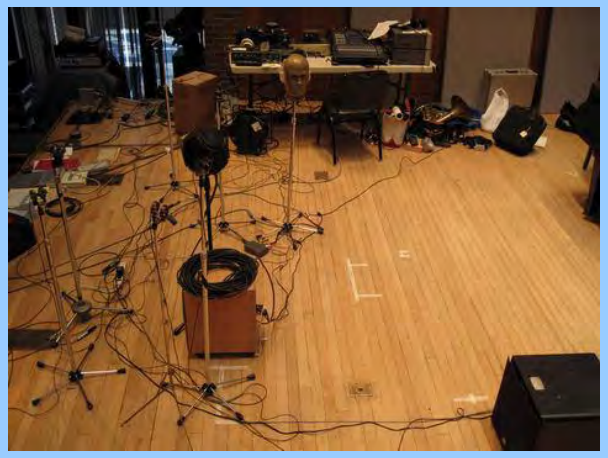
The sound in many seats is muddy. Adding reflections or decreasing absorption only increases the muddiness.

Hall data

- The pictures show a recital hall of 65000 cubic feet (1840 cubic meters). Designed for 350 seats, it has currently 300 seats, giving a volume/seat of 6 cubic meters. There is 1400 square feet of carpet under the seats on the floor.
- Reverberation Time (RT) unoccupied is 1.1 seconds from 1000Hz to 63Hz. C80, dominated by the reverberation time, is ~+5dB everywhere.
- The parallel side walls of the stage provide little diffusion.
- The hall is generally liked by the audience and players, although there are reports of loudness and balance problems on stage.
- Musicians desire more resonance and greater clarity in the middle of the hall.

Experiments with absorption and acoustic enhancement

- Measurements and experiments involving various combinations of fiberglass panels and electronic reverberation enhancement were conducted in January 2009.
  - Measurements were made with three loudspeakers, three dummy heads, and a Soundfield microphone.
  - All musical performances were recorded with the same microphones, and with an array of close microphones on stage.
- About 30 musicians participated, including faculty, staff, students from all three divisions, and musicians from the wider community.
- The goal was to improve the instrumental balance on stage, reduce excess stage loudness, and to increase resonance and the ability to hear individual instruments throughout the hall.
- With both panels and enhancement in place comments from the participants were favorable. Players and singers found balancing with piano was easier, and the middle registers of the piano were more easily heard both by the musicians and in the hall.



The absorptive curtains at the rear of the stage could be rapidly withdrawn. The blankets that simulated audience could be removed in 5 minutes, along with the panels on stage. This allowed prompt A/B comparisons. Some of the 25 LARES enhancement speakers are visible

Results from the experiments

- The experiments in January showed that adding fiberglass panels around the stage increased clarity and the ability to localize instruments in the hall, raising the measured value of LOC from an average of minus 1.5dB to +3dB or more.

- Localization and clarity in the balcony were additionally improved by adding panels to the upper audience right side wall, which eliminated the strong lateral reflection from that surface.
  - The lower surface of this wall was already absorptive.

- The electronic enhancement successfully compensated for the loss of resonance due to the panels. Without the enhancement the perceived resonance was reduced.

- In a subsequent experiment with a violin-piano combination and no enhancement we found that just 12 fiberglass panels each 2'x6'x2" around the bottom of the stage noticeably improved the clarity on the floor of the hall, and also improved the balance for the players on stage. For this music the reduced resonance was not a problem.
  - These panels absorbed the first-order reflection from the back of the stage, which has the highest level and the shortest time delay. Absorbing this reflection contributed strongly to increasing LOC.

Usefulness of the measure LOC

- LOC informs us that the primary contribution to difficulty in localization are the first strong reflections, regardless of the direction they come from.

- We initially thought that since the floor of the hall is not a significant source of these reflections, it is would be likely that removing the carpet under the seats would raise the RT without decreasing LOC significantly.
  - However LOC is also sensitive to reverberation which arrives before 100ms, and this would be increased by removing the carpet.
  - A few later experiments suggest that removing the carpet will increase the reverberant level sufficiently to eliminate the improvement in LOC provided by the absorption on stage.

- The existence of a LOC as a physical measure can help to answer these questions in advance – or at least suggest that an experiment is needed before drastic alterations are undertaken.

Small shoebox halls can be OK

- If the client insists on a shoebox it can work by building a large hall and installing a small number of seats.
  - I was just in such a small hall in Helsinki, and at least half the seats were OK.
- But this is not the ideal solution.
  - With a different shape nearly all the seats could have been OK – and it might have been less expensive.

Light models

I ran across these pictures while cleaning out my office. The top one is a too-simple model of the Philadelphia Academy of Music.

The bottom is intended to be BSH, but with a single balcony.

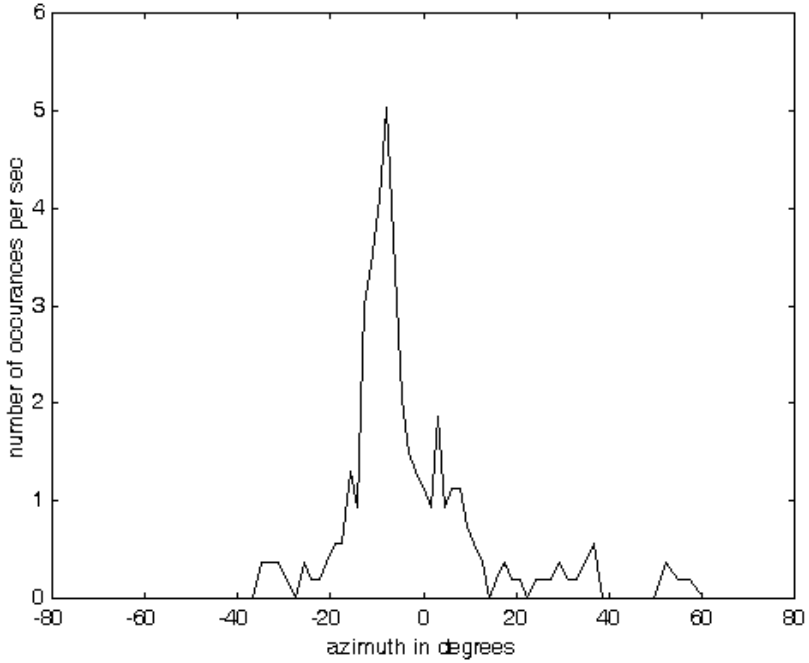
I abandoned light modeling because it does NOT provide any information about the time delay gap – nor information about the effects of note length on D/R.

But it DOES provide information about the total reverberant energy compared to the direct. And very complex hall shapes can be quickly modeled.

A few slides from earlier measures of azimuth from recordings, using running IACC in a 10ms window.

They are followed by an earlier measure of harmonic coherence.

Localization

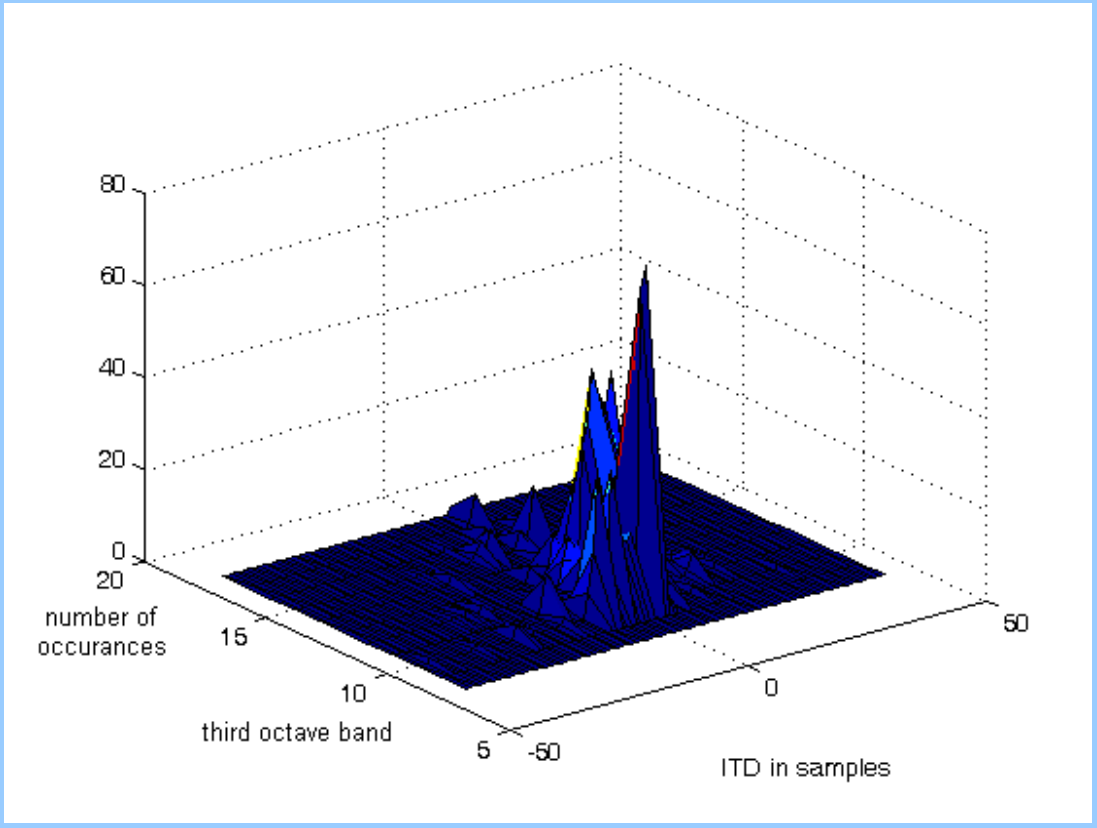


The figure shows the number of times per second that a solo violin can be localized from row 4 of a small shoebox hall (~500 seats) near Helsinki.

It also shows the perceived azimuth of the violin

As can be seen, the localization – achieved at the onsets of notes – is quite good, and the azimuth, ~10 degrees to the left of center, is accurate.

Localization – surface1

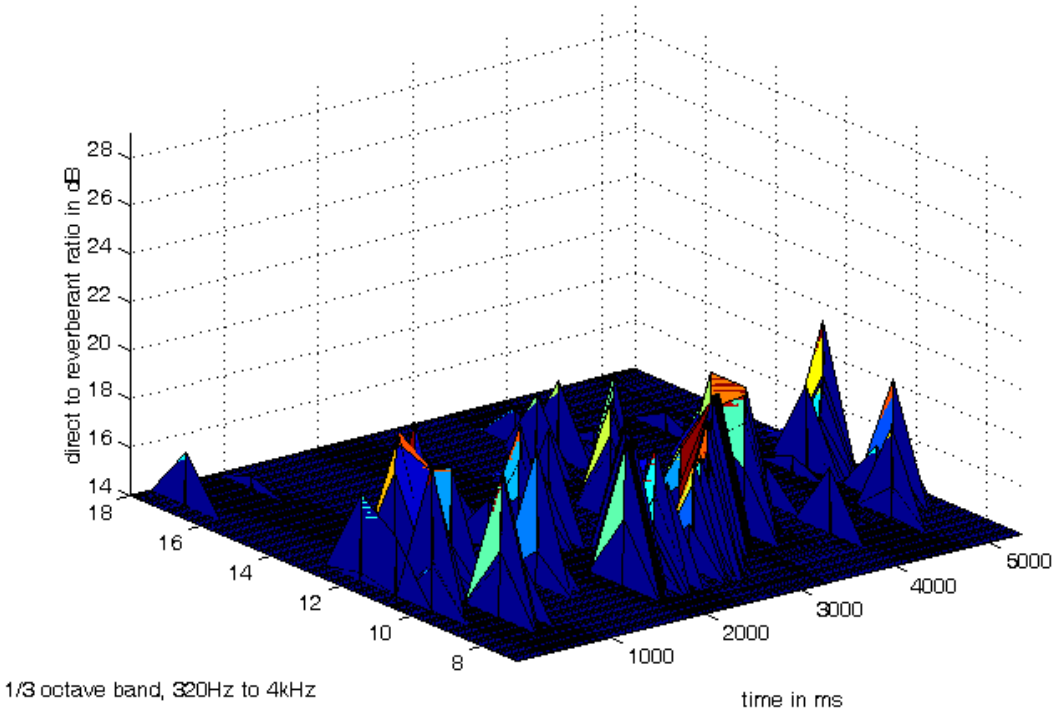


Here we plot the same data for the violin as a function of (inverse) azimuth, and the third octave frequency band.

As can be seen, for this instrument the principle localization components come at about 1300Hz.

Interestingly, Human ability to detect azimuth, as shown in the threshold data, may be maximum at this frequency.

Localization, Surface 2

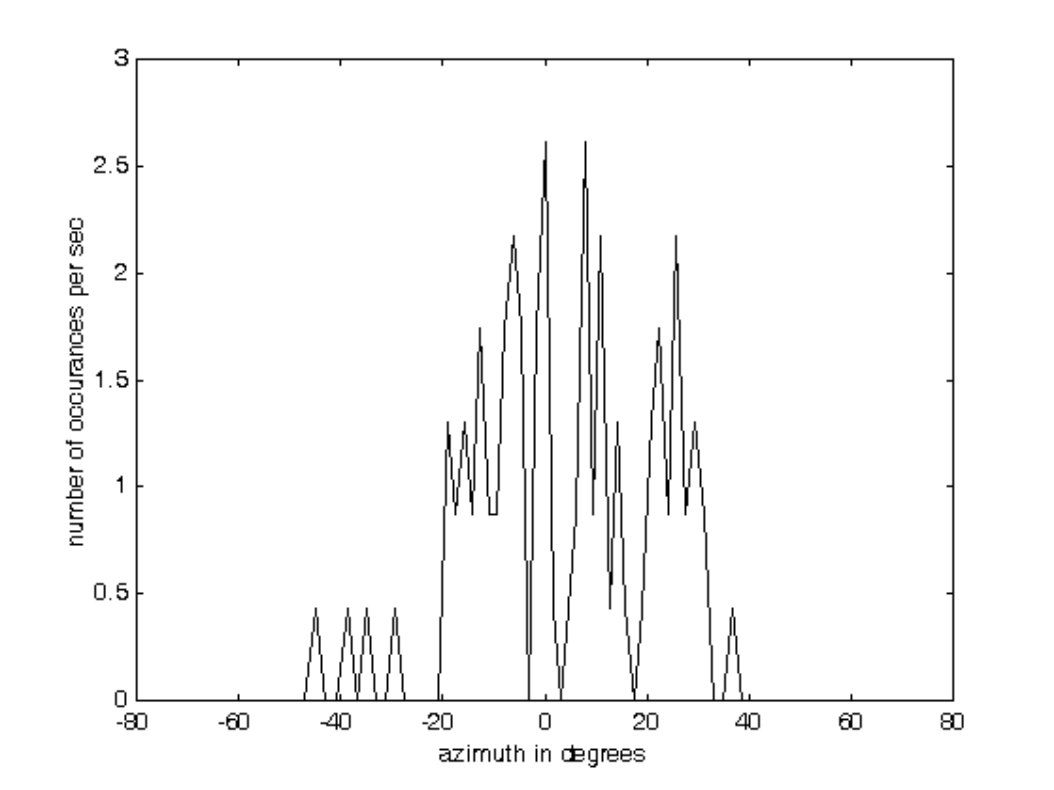


Here we plot 1/(1-IACC) as a function of time and third octave band.

Note that the IACC peaks at the onset of notes can have quite high values for a brief time.

This happens when there is sufficient delay between the direct and the reverberation, and sufficient D/R.

Localization – a poor seat



Here is a similar diagram for a solo violin in row 11 of the same hall. The sound here is unclear, and the localization of the violin is poor.

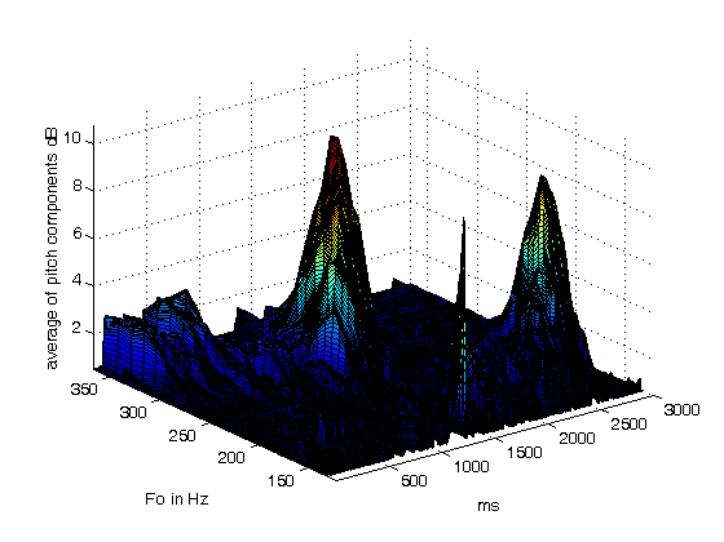
As can be seen, the number of localizations per second is low (in this case the value really depends on the setting of the threshold in the software).

Perhaps more tellingly, the azimuth detected seems random.

This is really just noise, and is perceived as such.

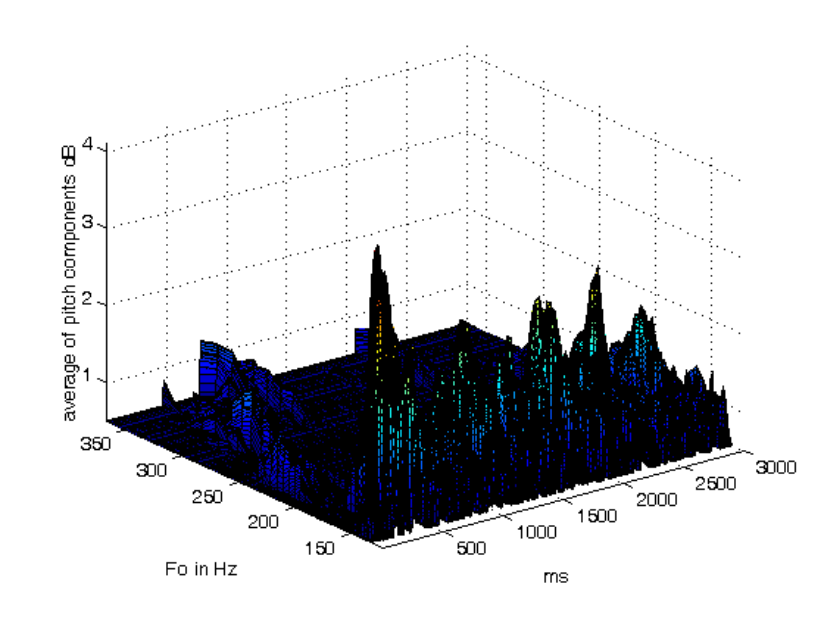
Measures based on harmonic coherence

- In the absence of reflections the formant frequencies above 1000Hz are amplitude modulated by the phase coherence of the upper harmonics. This modulation is easily heard, creating the perception of “roughness” (Zwicker).
  - Reflections randomize the phase of these harmonics.
- The result is highly audible, and is a primary cue for the distance of an actor, singer, or soloist.
- This effect can be measured with live recordings, and is sensitive both to medial and lateral reflections.



This graph shows the frequency and amplitude of the amplitude modulation of a voice fundamental in the 2kHz 1/3 octave band. The vertical axis shows the effective D/R ratio at the beginning of two notes from an opera singer in Oslo to the front of the third balcony (fully occupied.) The sound there is often muddy, but the fundamental pitch of this singer came through strongly at the beginning of these two notes. He seemed to be speaking directly to me, and I liked it.

Another singer



From the same seat the king (in Verdi's Don Carlos) was not able to reach the third balcony with the same strength.

Like the localization graph shown in a previous slide, this graph seems to be mostly noise.

The fundamental pitches are not well defined. The singer seemed muddy and far away.

His aria can be heart-rending – but here it was somewhat muted by the acoustics. We were watching the king feel powerless and forlorn. But we were not *engaged.*