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

David Griesinger

David Griesinger Acoustics, Cambridge, Massachusetts, USA
www.davidgriesinger.com

Introduction: What is Clarity?

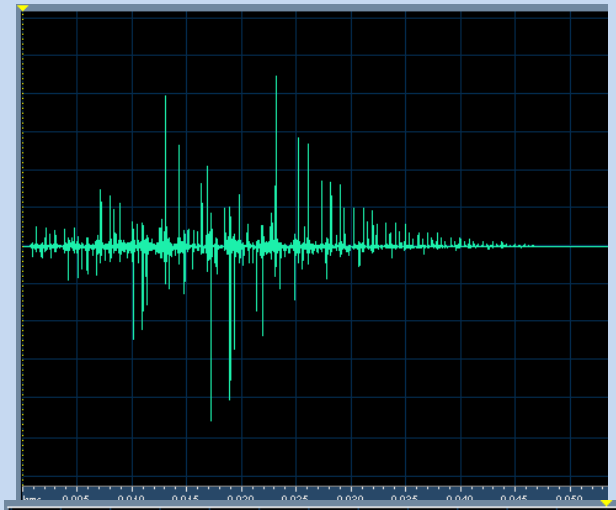
- Clarity and direct sound are key to this talk, but I propose:
  - But we don't know how to define clarity.
  - And we don't know how to measure it.
    - If we wish to design the best halls, operas, stages, and classrooms, we must break out of this dilemma.
- We will propose a solution based on human abilities to separate simultaneous sound sources.
  - This is one of several abilities that all depend on the same physical mechanisms.
- The conclusions we draw are surprising and can be uncomfortable:
  - Too many early reflections from any direction can eliminate clarity.
  - The earlier a reflection comes (>10ms) the more damaging it is.
  - Adding absorption to a stage area can greatly increase clarity for the audience.
  - When clarity is poor absorbing or deflecting the strongest first-order reflection can make an enormous improvement.

C80 and C50 may be somewhat related to intelligibility

- But Clarity is NOT the same as intelligibility .
- When sound is unclear words may be recognizable, but it may not be possible to remember what was said.
  - Working memory is limited. When grammar and context are needed for recognition, there is no time left to store the meaning. (SanSoucie)

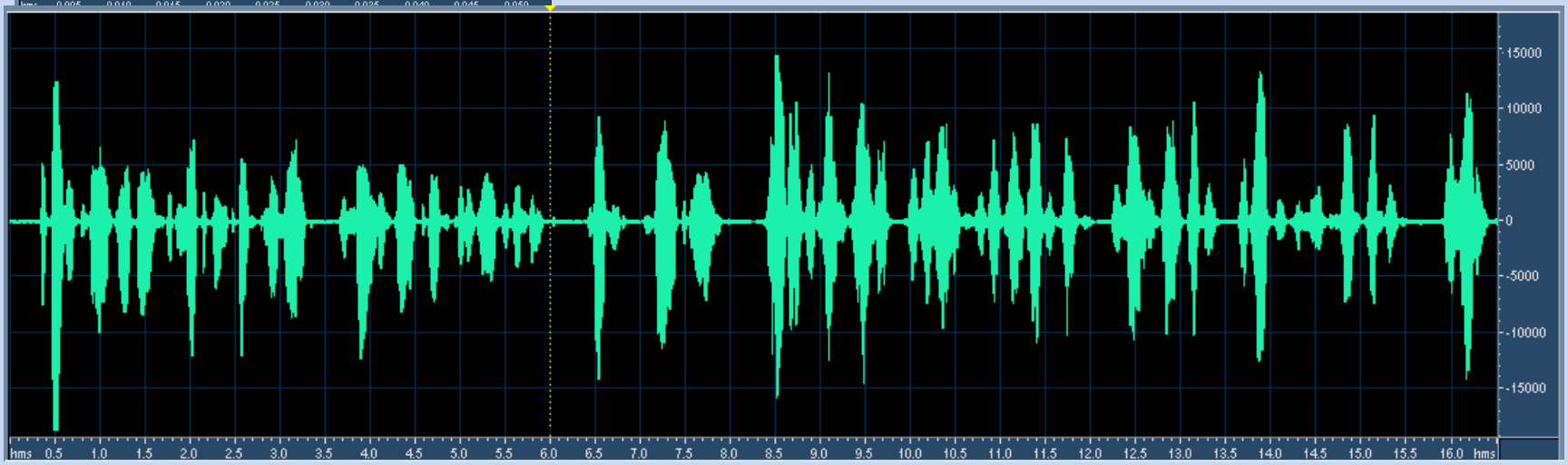
Example of Clarity for Speech

- This impulse response has a C50 of infinity
  - STI is 0.96, RASTI is 0.93, and it is flat in frequency.



In spite of high C50 and excellent STI, when this impulse is convolved with speech there is a severe loss in clarity. The sound is muddy and distant.

The sound is unclear because this IR randomizes the phase of harmonics above 1000Hz!!!



So What is Clarity? And what is “direct sound”

- Why does the previous impulse response affect clarity so strongly?
  - The speech in the previous example is not just difficult to understand.
    - It sounds distant
    - It is difficult or impossible to localize in a reverberant field
    - And it is difficult or impossible to separate from another example of unclear speech spoken simultaneously.
  - All these perceptions depend on the same ear/brain mechanism.
    - And all are dependent on the presence of high-order harmonics of complex tones.
- We claim that clarity is perceived when harmonics in the vocal formant range retain their original phase relationships
  - At least for sufficient time at the onset of a sound that the brain can decode them.
- The “direct sound” is the component of sound that retains the original harmonic phase relationships.
  - Very prompt <~5ms reflections do not alter phases!
  - But a 10ms or more reflection can be damaging, and the sooner a reflection comes the more damaging it is.

A little history

- At RADIS in 2004 I presented a paper showing that our perception of near and far depends on the presence of harmonic tones!
  - If loudness is controlled you cannot perceive near and far with noise-like sounds or whispered speech.
  - But with speech or music in a hall or room the perception of near or far is nearly instantaneous.
  - I found that the perception of “near” depends critically on the phase coherence of harmonics in the vocal formant range.
    - Coherent harmonics are produced by solo instruments.
      - Once every fundamental period the harmonics are in phase.
      - The ear easily detects the peak in sound pressure – and the perception of “near” results
    - Reflections randomize the phases – and the ear perceives “far”.

Audience Engagement

- A few years later I connected the perception of “near” with the ability of a sound to demand, and hold, the attention of a listener.
  - I presented papers on this subject at the ICA in Madrid, and the following conference in Seville.
  - The only result I could detect was severe audience confusion. “Engagement” does not translate into other languages, and there is no standard measure for it.
  - And no one seems to know what “harmonic coherence” might mean.

- But to me the ability to precisely localize sound sources is strongly correlated with engagement.
  - So I studied the threshold localization of sound sources in a diffuse reverberant field.
  - The data was fascinating, and begged for an objective measure.
  - Using this data I developed the measure called LOC.

Localizing three instruments playing simultaneously

- During a quartet concert in January of 2010, fascinated that I could hear three instruments at the same time, I had a revelation:
  - Near/far,
  - The localization of sound sources in a highly reverberant field,
  - The ability to identify by timbre and localization simultaneous musical lines,
  - Stage acoustics,
  - and classroom acoustics
- ALL depend on the ability to separate simultaneous sounds into separately perceivable sound streams. (the cocktail party effect.)
  - ALL depend on the presence of harmonic tones.
  - And all are degraded in similar ways by reflections.

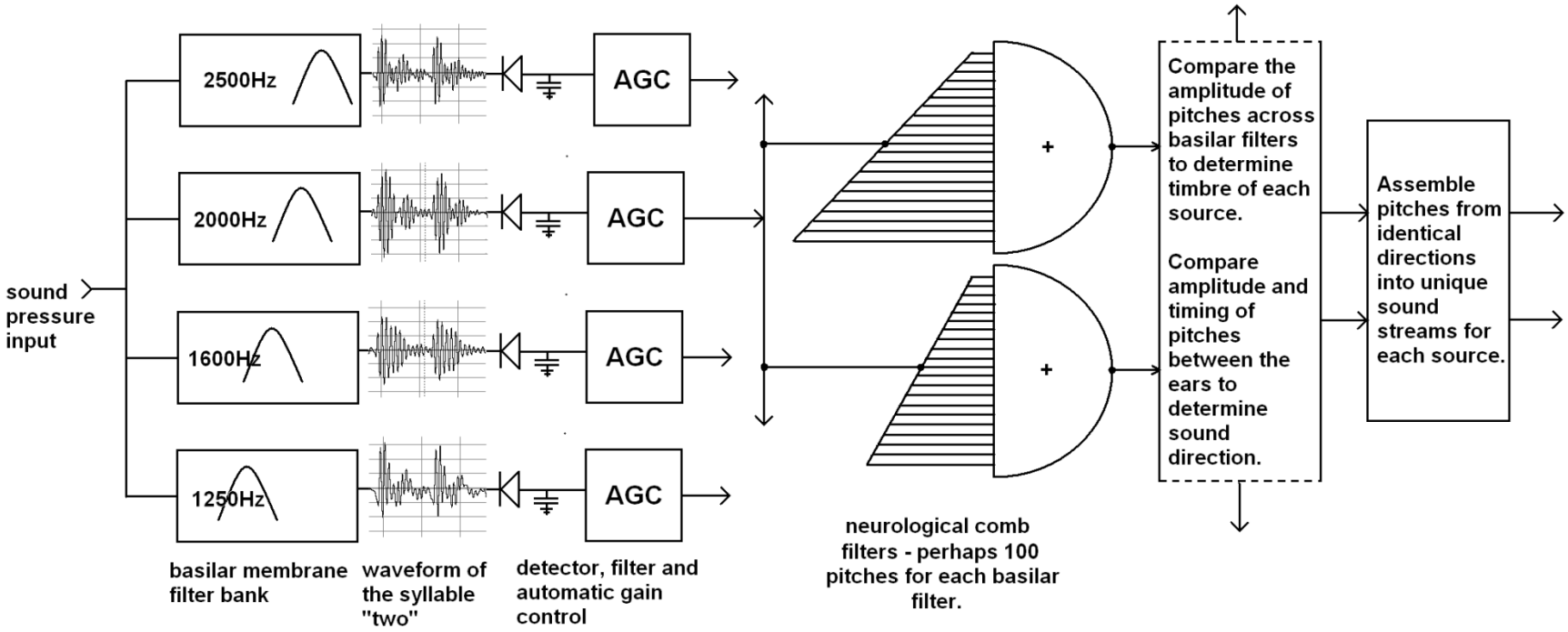
- It should be possible to define and measure “CLARITY” by the ease with which we can perceive the distance, timbre, and location of simultaneous sound sources.

Measures from live music

- Binaural impulse responses from occupied halls and stages are very difficult to obtain!
  - But if you can hear something, there must be a way to measure it.
- So I developed a model for human hearing!

- The sound is the Pacifica String Quartet playing in the Sala Sinfonica Puerto Rico – binaurally recorded in row F
- This sound is the same players as heard in row K, just five rows further back. The sound is very different – distant and muddled together. The ability to perform the cocktail party effect has been lost due to an excess of reflections.

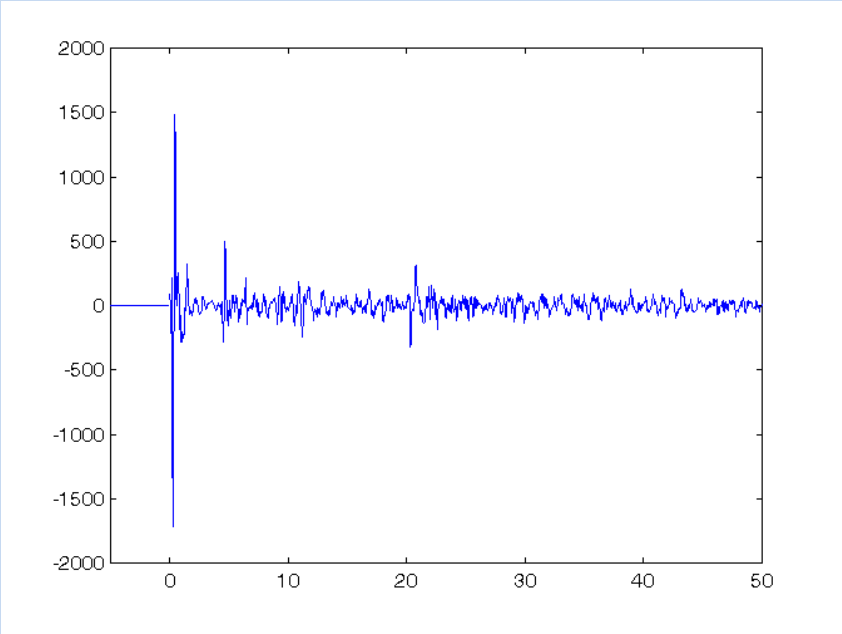
The Model

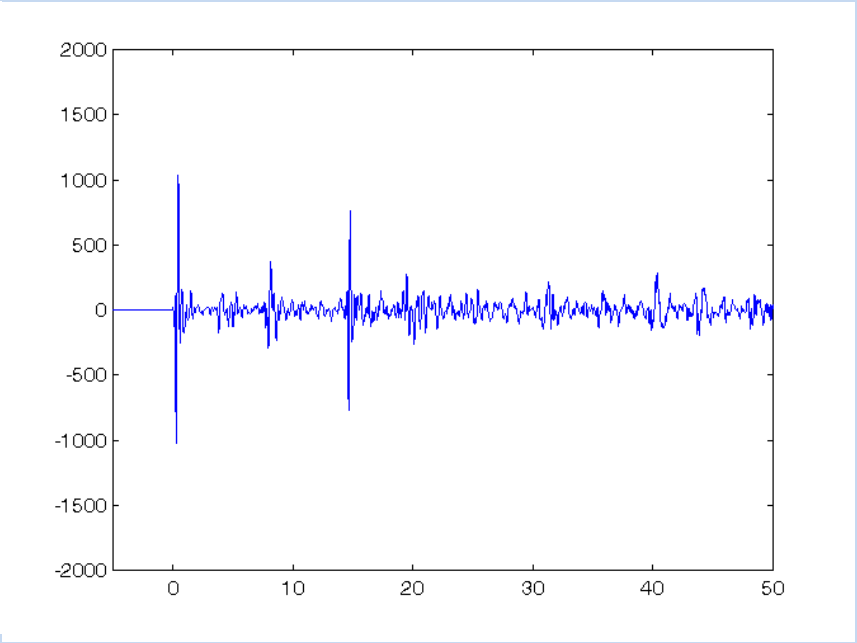


An explanation of this model is in the preprint and on my web-page.

We do not need to understand it to develop a useful measure for Clarity.

As an example, here are two impulse responses from Boston Symphony Hall.



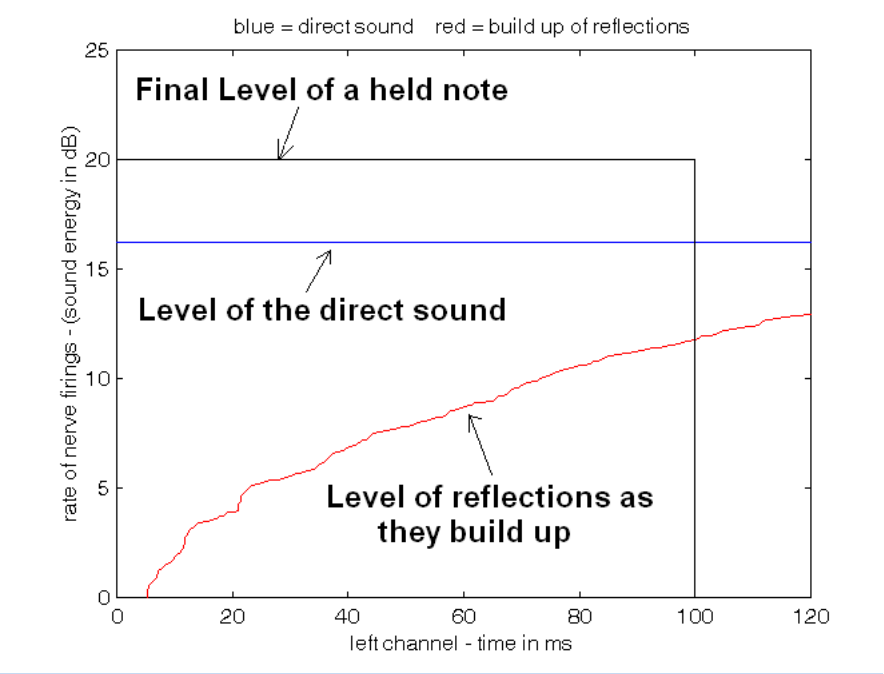


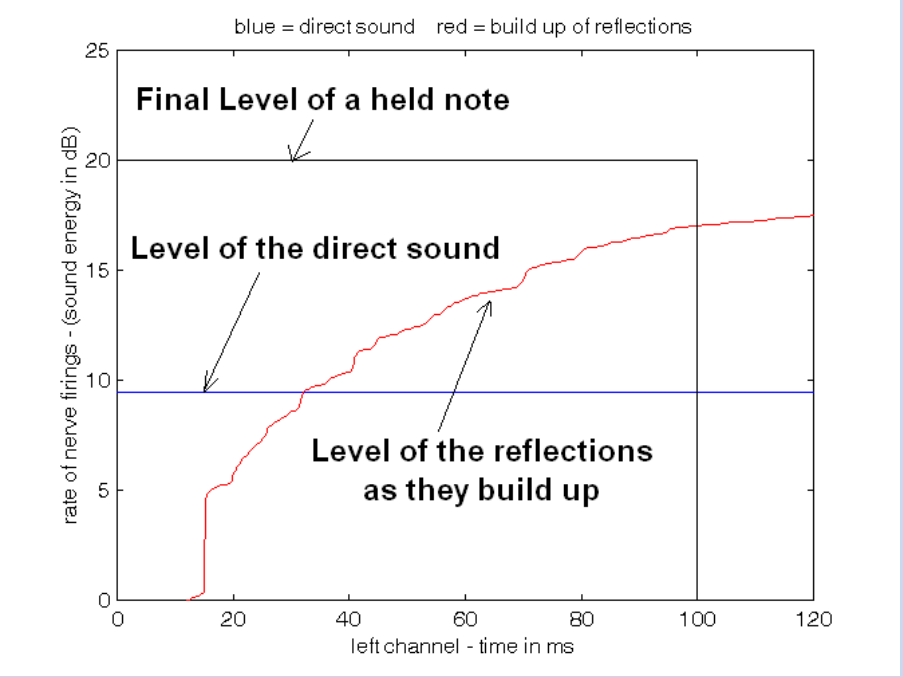
Binaural impulse response BSH row R seat 11
C80 = 0.85dB IACC80 = .68 LOC = 9.1dB

Same, Row DD, seat 11 C80=-0.21
IACC80 = 0.2 LOC = -1.2

C80 is nearly the same for both seats – but clarity is excellent in row R, and nearly absent in row DD. LOC clearly identifies the better seat.

These two impulse responses lead to a simple diagram:





Boston Symphony Hall row R seat 11 from the podium. The left channel of a binaural impulse response. LOC = 9.1dB

Same, row DD, seat 11. The final sound level is almost the same, but in this seat it is mostly reflections. LOC = -1.1dB

Note the window defined by the black box. We propose that if the area under the direct sound is greater than the area under the red line, the sound will be CLEAR. The ratio of these areas is LOC (in dB).

And the following equations:

- We can use this simple model to derive an equation that gives us a decibel value for the ease of perceiving the direction of direct sound. The input *p(t)* is the sound pressure of the source-side channel of a binaural impulse response. (700-4000Hz)
  - We propose the threshold for localization is 0dB, and clear localization and engagement 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. It is 20dB below the maximum loudness. *POS* in the equation below means ignore the negative values for the sum of S and the cumulative log pressure.

- Localizability (LOC) in dB =

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

- 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.
- Further description of this equation is beyond the scope of this talk. An explanation and Matlab code are on the author's web-page..

LOC was *not* derived from a hearing model, but from a few well-known facts.

- Humans can detect pitch to about one part in a thousand – (~3 cents).
- It takes a structure – either physical or neurological – of ~100ms length to measure a 1000Hz signal to that precision. And determination of loudness also requires an integration time of about 100ms.
- Our ears are sensitive to the integrated *logarithm* of sound pressure, NOT to the integral of sound energy.
- Our ears are acutely attuned to the onsets of sounds, and not to the way sound decays.

Note Onsets

- The ear is attuned to sound onsets, not sound decays:
  - Consider reverberation forward and reversed:

Forward

Reversed

These Facts Predict:

- We need a structure for integrating sound about 100ms long
- We need to analyze NOTES or SYLLABLES – short bursts of harmonic tones, not clicks or infinitely long noise that suddenly stops.
- We need to integrate the LOGARITHM of sound pressure – not pressure squared.
- We need to look at note ONSETS, not decays.

www.akutek.info

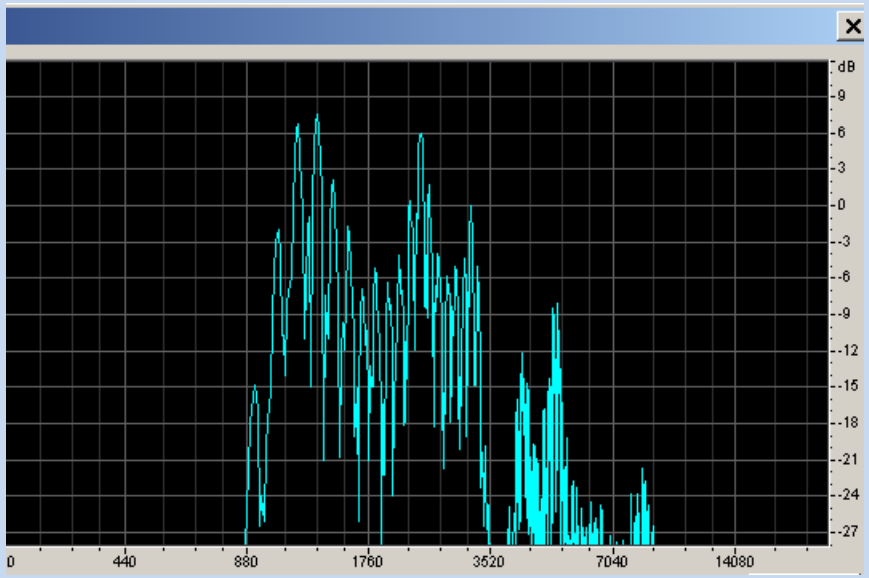
Demonstration

- The information carried in the phases of upper harmonics can be easily demonstrated:

Dry monotone Speech with pitch C

Speech after removing frequencies below 1000Hz, and compression for constant level.

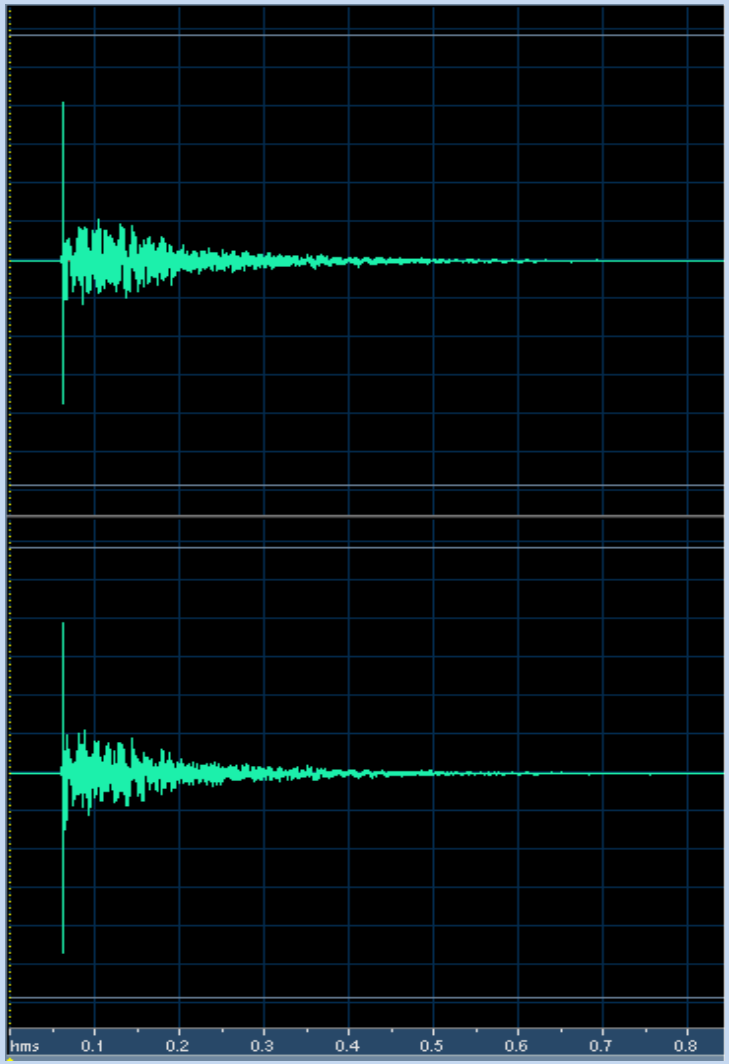
C and C# together



Spectrum of the compressed speech

It is not difficult to separate the two voices – but it may take a bit of practice!

What happens in a room?



Measured binaural impulse response of a small concert hall, measured in row 5 with an omnidirectional source on stage. The direct level has been boosted 6dB to emulate the directivity of a human speaker.

RT ~ 1s

Looks pretty good, doesn't it, with plenty of direct sound.

But the value of LOC is -1dB, which foretells problems…

Sound in the hall is difficult to understand and remember when there is just one speaker. Impossible to understand when two speakers talk at the same time.

C in the room

C# in the room

C and C# in the room together

The Cocktail Party Effect and Classrooms

- The ability to separate sounds by pitch is not just an advantage when there are multiple speakers.
  - Pitch acuity also separates meaningful sounds from noise.
  - Recognizing vowels is easier when the direct sound is easily detected and analyzed.
  - When the brain must devote working memory to decoding speech, there is not enough memory left over to store the information.

Localization and Envelopment

- The ability to precisely localize sound sources changes the apparent direction of reflections and reverberation.
- Reverberation and reflections without precise localization of sources is perceived as in front of a listener.
  - In nearly all halls it ***is*** in front.

- When direct sound is added just above the threshold of audibility reverberation is perceived as louder and all around the listener.
- The effect is perceived at all frequencies, even if the direct sound is band-limited to the 1kHz or 2kHz octave bands.

- When the pitch, timbre, location, and distance of a source can be perceived at the onset of a sound we perceive these properties as extending through the sound, even if later reverberation overwhelms the data in the direct sound.
- When – as in a recording – the reverberant level is low, we perceive the reverberation as continuous, even if the direct sound overwhelms it.

Conclusions

- We have proposed that amplitude modulations of the basilar membrane at vocal formant frequencies is responsible for
  - Making speech easily heard and remembered,
  - Making it possible to attend to several conversations at the same time,
  - And making it possible to hear the individual voices in a music performance.
  - A model based on these modulations predicts a great many of the seemingly magical properties of human hearing.
- Although some of the consequences of this research for hall, stage, and classroom design might seem controversial or disturbing, they can be and have been demonstrated in real rooms.

- The power of this proposal lies in the simple physics behind these hearing mechanisms. The relationships between acoustics and the perception of timbre, direction and distance of multiple sound sources becomes a physics problem .
  - How much do reflections and reverberation randomize the phase relationships and thus the information carried by upper harmonics.
- A measure, ***LOC,*** is proposed that is based on known properties of speech and music.
  - In our limited experience LOC predicts – and does not just correlate with – the ability to localize sound sources simultaneously in a reverberant field. It may be found to predict the ease of understanding and remembering speech in classrooms, the ease with which we can hear other instruments on stages, and the degree of envelopment we hear in the best concert halls.
- A computer model exists of the hearing apparatus shown in the model slide.
  - The amount of computation involved is something millions of neurons can accomplish in a fraction of a second. The typical laptop finds it challenging.
  - Preliminary results indicate that a measure such as LOC can be derived from live binaural recording of music performances.