FAQ's and further information on Kunchur's research on temporal resolution of human hearing and audio reproduction

My three papers on this subject (published in *Technical Acoustics*, *Acta Acustica*, and *POMA*, which can be freely downloaded from the web site http://www.physics.sc.edu/kunchur/Acoustics-papers.htm)were discussed on various internet forums (stereophile, gearslutz, hydrogenaudio, etc.) and generated some very interesting questions. These questions were collected and forwarded to me and I have prepared this FAQ page to answer them. Journal papers have a page limit and style restrictions, and contain an implicit assumption of a certain background of the underlying subject on the part of the reader. Long tutorial sections are prohibited (except in the case of review papers). Naturally, misinterpretations can result when statements are read out of context or only fragments of the papers are read. Also definitions of terms may differ in different fields; so two people may use the same expression or terminology to mean two different things. This especially applies to the present papers because they intersect many different disciplines. In the present article, I want to fill in the blanks, so to speak, and provide more background and context for the information presented in the papers. Rather than answer every question individually, I have lumped together some of the overlapping conceptual types of questions, and address these in the sections that follow. After that I address some of the other more specific questions. I hope this discussion will help to clarify the subject.

Temporal resolution and digital signals

In most fields of science, "to resolve" means to "substantially preserve the essence of the original signal" and in particular to be preserve enough information in the signal so that it can "become separated or reduced to constituents" (e.g., please see *v.tr.* [11] and *v.intr.* [2] under http://www.thefreedictionary.com/resolve). If the constituents cannot be separated and have merged together, the signal's essence has been killed. However, a certain other definition exists which pertains to the smallest time shift that produces a difference in the final digital code; this resolution allows noticing differences in the "degrees of death" of the killed signal rather than the system's ability to preserve sonic details and convey them to the ear. In psychoacoustics and auditory neurophysiology, the former definition applies. Below I give optical and audio examples to explain this further.

Optical example: A binary star system is imaged through a telescope with a CCD. First, there is the analog optical resolution that is available, which depends on the objective diameter, the figure (optical correctness) of the optics, and seeing (atmospheric steadiness). This optical resolution is analogous to the "analog bandwidth". Because this resolution is limited, a point source becomes spread out into a fuzzy spot with an intensity profile governed by the point spread function or (PSF). Next we are concerned with the density of pixels in the CCD. To avoid aliasing, the pixel spacing L must be finer than the optical resolution so that the optics provides "low pass filtering". If the pixels and their separation are larger than the separation of the centers of the two star images, the two stars will not be resolved separately and will appear as a single larger merged spot. In this case the essential feature (the fact that there are two separate stars and not an oblong nebula) has been destroyed. This is usually what is meant by "resolution" or the lack of it. The number of bits N that can differentiate shades of intensity ("vertical resolution") has little to do with this – no number of vertical bits can undo the damage. However, details of the fuzzy merger do indeed depend on N: if the star images are moved closer together, the digital data of the sampled image will be different as long as the image shift exceeds L/N. This L/N definition of resolution applies to the coding itself and not to the system's ability to resolve essential features in the signal as described above (otherwise, the average 6" backyard telescope with a 12 bit CCD would have a resolution that is < 0.001 arc seconds, which is better than the ~0.1 arc seconds resolution of the research grade telescopes!).

Digital audio recording: In my papers, statements related to "consumer audio" refer to CD quality, i.e., 16 bits of vertical resolution and a 44.1 kHz sampling rate (when the work for these papers was begun around 2003, 24bit/96kHz and other fancier formats were not in common use in people's homes for music reproduction). For CD, the sampling period is 1/44100 ~ 23 microseconds and the Nyquist frequency f_N for this is 22.05 kHz. Frequencies above f_N must be removed by anti-alias/low-pass filtering to avoid aliasing. While oversampling and other techniques may be used at one stage or another, the final 44.1 kHz sampled digital data should have no content above f_N. If there are two sharp peaks in sound pressure separated by 5 microseconds (which was the threshold upper bound determined in our experiments), they will merge together and the essential feature (the presence of two distinct peaks rather than one blurry blob) is destroyed. There is no ambiguity about this and no number of vertical bits or DSP can fix this. Hence the temporal resolution of the CD is inadequate for delivering the essence of the acoustic signal (2 distinct peaks). However this lack of temporal resolution regarding the acoustic signal transmission should not be confused with the coding resolution of the digitizer, which is given by 23 microseconds/2^16 = 346 picoseconds. This latter quantity has no direct bearing on the system's ability to separate and keep distinct two nearby peaks and hence to preserve the details of musical sounds. Now the CD's lack of temporal resolution for complete fidelity is not systemic of the digital format in general: the problem is relaxed as one goes to higher sampling rates and by the time one gets to 192 kHz, the bandwidth and the ability to reproduce fine temporal details is likely to be adequate. I use the word "likely" rather state definitely for two reasons. In our research we found human temporal resolution to be ~5 microseconds. This is an upper bound: i.e., with even better equipment, younger subjects, more sensitive psychophysical testing protocols, etc., one might find a lower value. The second reason to not give an unambiguous green signal to a particular sampling rate is that the effective bandwidth that can be recorded is less than the Nyquist frequency because of the properties of the anti-aliasing filter, which is never perfect in real life. One more thing I want to add is that one forum poster inquired whether the blurring is an analog effect and not a digital one ("… this isn't a sampling-rate issue, it's a simple question of linear filtering…"). But the two are not separate. While it is true that the smearing may take place in the analog low-pass filter circuitry before the signal reaches the ADC, the low-pass filter cutoff is dictated directly by the sampling rate. The exact amount of smearing and other errors will depend on the slope and other details of the filter, but the big-picture conclusion is still the same.

Digital synthesis and DSP

Some questions were asked pertaining to the limitations of digital synthesis versus analog waveform generation for use in the experiments. Digital synthesis and signal processing are vast subjects that go beyond the scope of the experiments and are in fact not directly relevant for the results of the experiments since digital sources were not used. However in the papers, I made certain observations on this subject mainly for one purpose: a referee who reviewed my paper asked why some simpler methods involving certain "typical" types of digital sources could not be used. My comments pertain to those types of digital sources – in particular output from a sound card when fed by a square waveform *wave file* generated by a software such as Sound Forge and secondly the output from certain *arbitrary waveform* generators intended for (non-audio) scientific research. Here I will give a brief description of the digital synthesis process and the potential (sometimes subtle) problems as they pertain to the experiments. [For more information, please see the following references: "Communication in the presence of noise", C. E. Shannon, Proc. Institute of Radio Engineers, vol. 37, 10–21 (1949); "The Shannon sampling theorem—Its various extensions and applications: A tutorial review" by A. J. Jerri, Proc. of the IEEE Vol. 65, 1565 – 1596 (1977); and "Multirate Signal Processing for Communication Systems", by F. J. Harris (Prentice Hall).]

During the recording/digitizing process, the low-pass filtered signal (with components above the Nyquist frequency removed) is digitized so that the level at each sample point in time is represented as a binary number (which for the 16-bit CD can have 2^16 = 65536 distinct values). During playback a smooth continuous waveform (which again should not have frequency components above the Nyquist frequency) must be reconstructed. If the DAC simply spat out the discrete voltage values for each sample, a staircase like output signal results with a lot of ultrasonic content. This is in fact what certain arbitrary waveform generators that are intended for (non-audio) scientific research do. For one model that was tested, the signal never became sufficiently smooth despite the use of built-in interpolation and filtering functions. Soundboards and other kinds of DACs intended for audio purposes use band limited interpolation to produce a smoother output waveform and thereby avoid such undesirable (and artificial) ultrasonic content. With optimum reconstruction (at least in theory – if certain conditions are satisfied) the exact original waveform can replicated from the set of discrete samples. This brings us to the Whittaker–Shannon interpolation formula: If one takes an arbitrary continuous signal V(t) and replaces it with a set of discrete samples V(n) that are periodic in time with a period T, then one can reconstruct the original continuous waveform *exactly* by placing a Sinc function at each sample point whose amplitude is equal to the V(n) value at that sample point. In other words:

$$V(t) = \sum_{n=-\infty}^{\infty} V(n)\ \mathrm{sinc}\left(\frac{t-nT}{T}\right)$$

This Whittaker–Shannon interpolation is exact as long as the following conditions are met:
(1) The discrete levels represented by V(n) are correct in the first place. (Some error will result because of the finite vertical resolution. Other, more serious, errors can result because of the way the V(n)'s are synthesized).
(2) V(t) is band limited and does not contain frequencies above the Nyquist value 1/[2T]. (This can be a problem when one is trying to reproduce or generate square waveforms, unless special tricks and techniques are used. The problem can be exacerbated when the square waveform and sampling rates have incommensurate periods; sine and other smoother waveforms are more forgiving in this regard as explained further below.)
(3) The sum stretches from minus to plus infinity. (In practice the sum is always finite).
(4) The interpolation function corresponds exactly to the sinc function, where sinc(x)=sin(πx)/πx. In the frequency domain, this corresponds to an exact "brick-wall" low-pass filter. (In practice this is impossible to realize exactly.)
(5) There is no error in the timing (jitter) at any point in the process: the voltage samples are taken exactly on time at the sampling frequency during the analog-to-digital conversion process and similarly the digital samples are converted to analog voltages exactly on time during the DAC process. (In practice no clock is perfect; but in modern high-end audio equipment jitter seems to have been reduced to negligible levels.)

In reality, *all* of these conditions are violated to some extent. How tolerable the errors are and whether the final result is acceptable depends on the application. For a product that is to be commercially marketed, the engineering requirements are such that the customers are sufficiently satisfied so as to buy the product. In a research experiment the standard is higher and the goal is to try to maintain mathematical purity to the extent possible and presuppose as little as possible regarding which errors won't be heard.

One can test the Whittaker–Shannon interpolation with a "mathematical experiment" using Mathematica™ software, avoiding physical instrumentation. Then conditions (1) and (4) can be met exactly. In this case if you construct a waveform that satisfies condition (2) then the Whittaker–Shannon procedure does a fabulous job of interpolation even in the case of incommensurate periodicity (condition (3) is still violated and so there are some errors at the beginning and end). The two plots on the graph below show:
(a) the original continuous function which is a blend of angular frequencies 1, 2, and 3 rad/s:
f(t) = sin(t) + 0.3*sin(2*t) + 0.2*sin(3*t)
and
(b) the reconstructed waveform where f(t) was first sampled at 5 rad/s (which is incommensurate with the original periodicity) and then each sampled point was multiplied by the scaled sinc function and these were then added together: g(t)= Sum[f[n*0.2]*sinc[Pi*(x - n*0.2)/0.2]

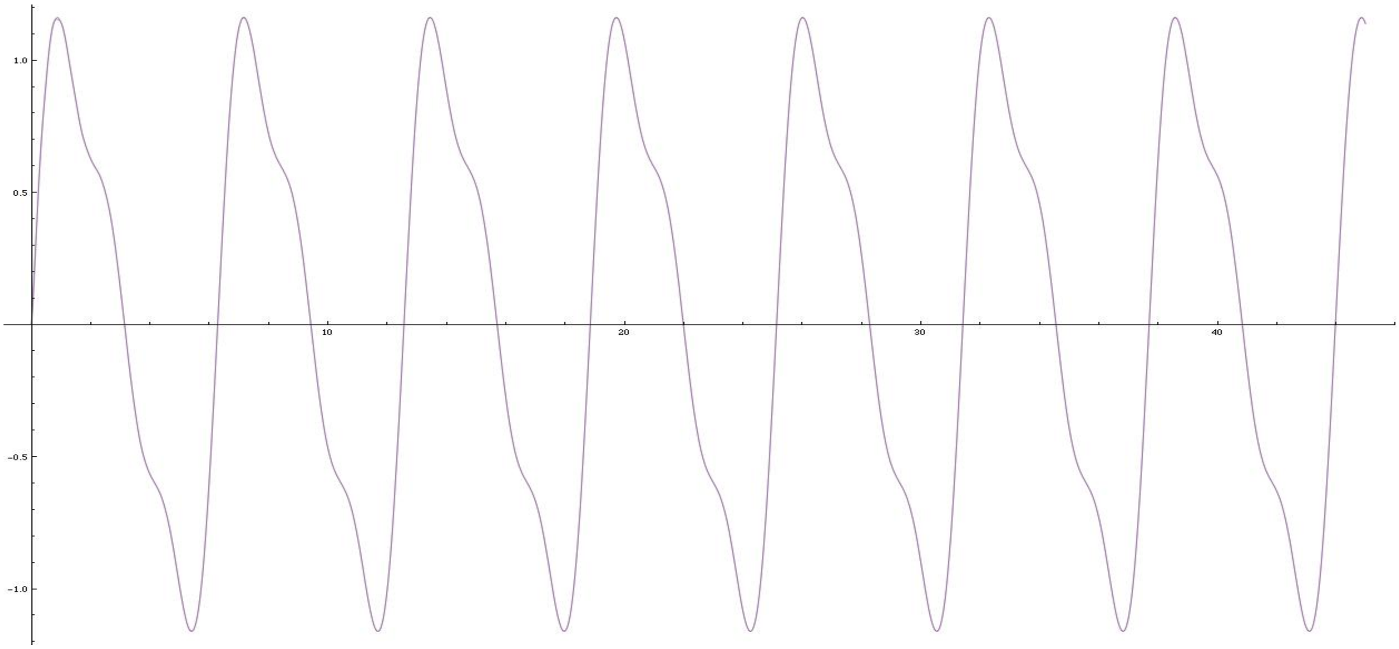

Fig. 1: Original continuous signal f(t) and reconstructed signal g(t) from sinc interpolation of samples. The two curves are essentially indistinguishable proving the exactness of the interpolation procedure.

In reality, the low-pass filter will not be an exact brick-wall filter. This introduces some error in the interpolation, which can be illustrated by adding some "impurity" to the interpolation function and re-plotting the two curves:
h(t) = Sum[f[n*0.2]*exp^(-(x - n*0.2)^2/0.04)*sinc[Pi*(x - n*0.2)/0.2]
The graph below shows the corresponding original (f(t)) and reconstructed (h(t)) functions. Notice that there is now some disagreement and that, in particular, the various cycles are slightly different thereby destroying the periodicity in the case when the signal and sampling periods are incommensurate (not exactly divisible without a remainder):

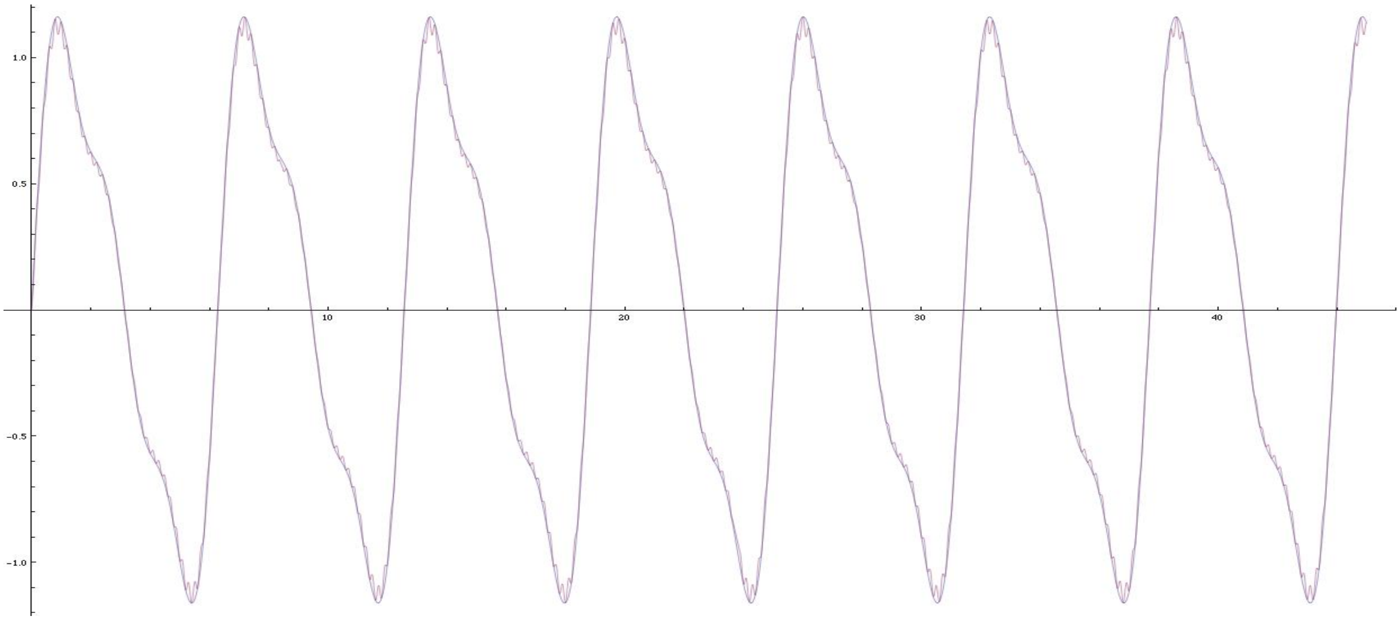

Fig. 2: Original continuous signal (ft) and reconstructed signal (h(t) from imperfect sinc interpolation of samples.

The above demonstrations are purely mathematical exercises and devoid of additional instrumental errors. Now let's see what happens in the case of one particular simple implementation of digital synthesis: using Sound Forge software to construct the waveforms and taking the output from a computer sound card. We will synthesize 8bit-8kHz wave files of 1850 Hz sine and square waveforms (in this case the signal frequency is incommensurate with the sampling frequency since 8000/1850 gives a remainder). Shown below are the wave files as seen in Sound Forge windows. Obviously this is not what actually comes out of the soundcard – such an assertion was neither stated nor implied.

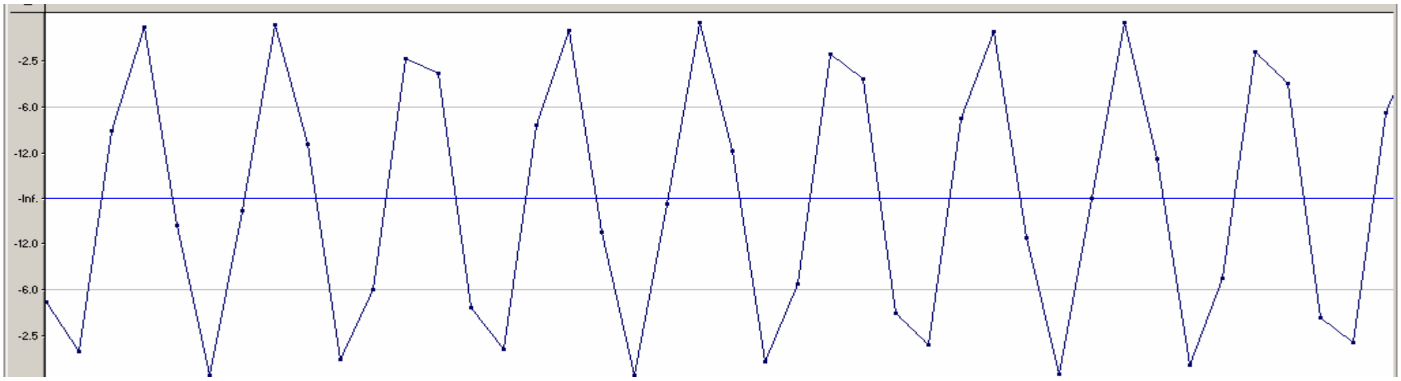

Fig. 3: *Wave file* of 1850 Hz sine waveform (lines joining points simply clarify ordering and *do not* imply output waveform!).

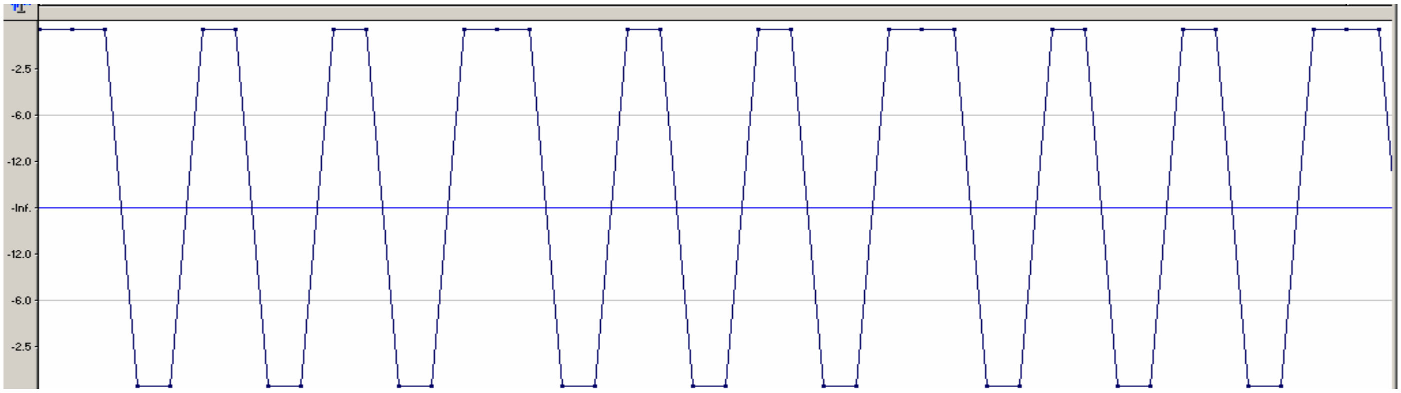

Fig. 4: *Wave file* for 1850 Hz square waveform (again lines joining points are only present to clarify point ordering and do not imply output waveform).

Here are the actual electrical signals that come out of the sound card for the above two respective wave files:

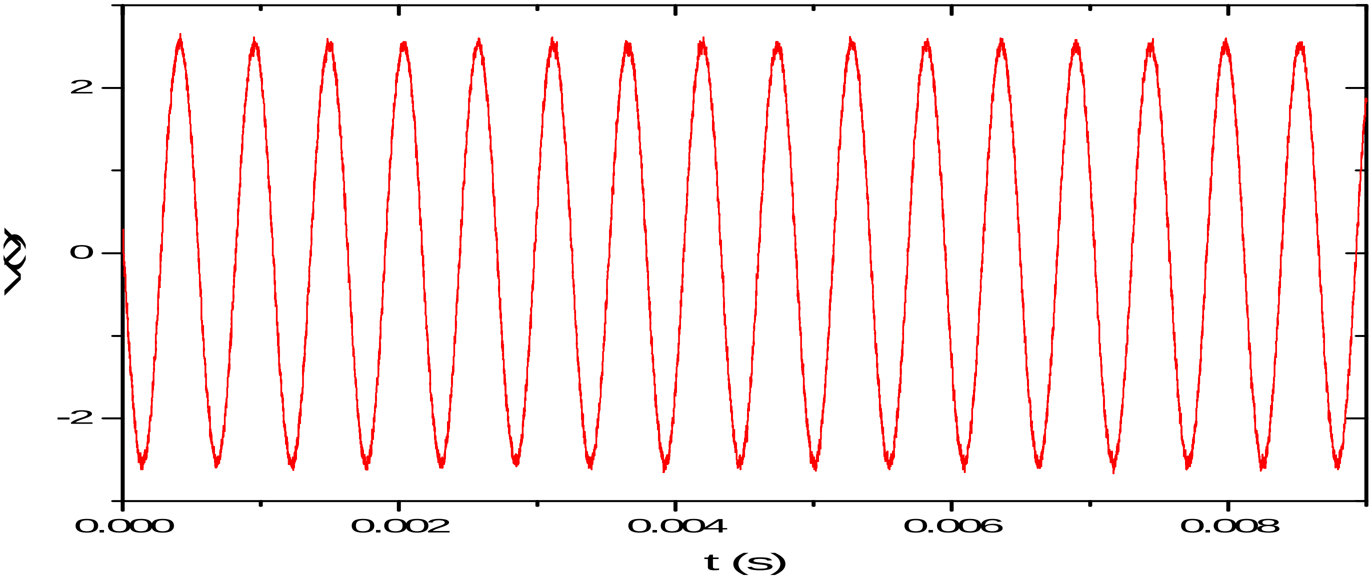

Fig. 5: Oscilloscope trace of electrical sound-card output for 1850 Hz (8 bits/8000 Hz) sine waveform.

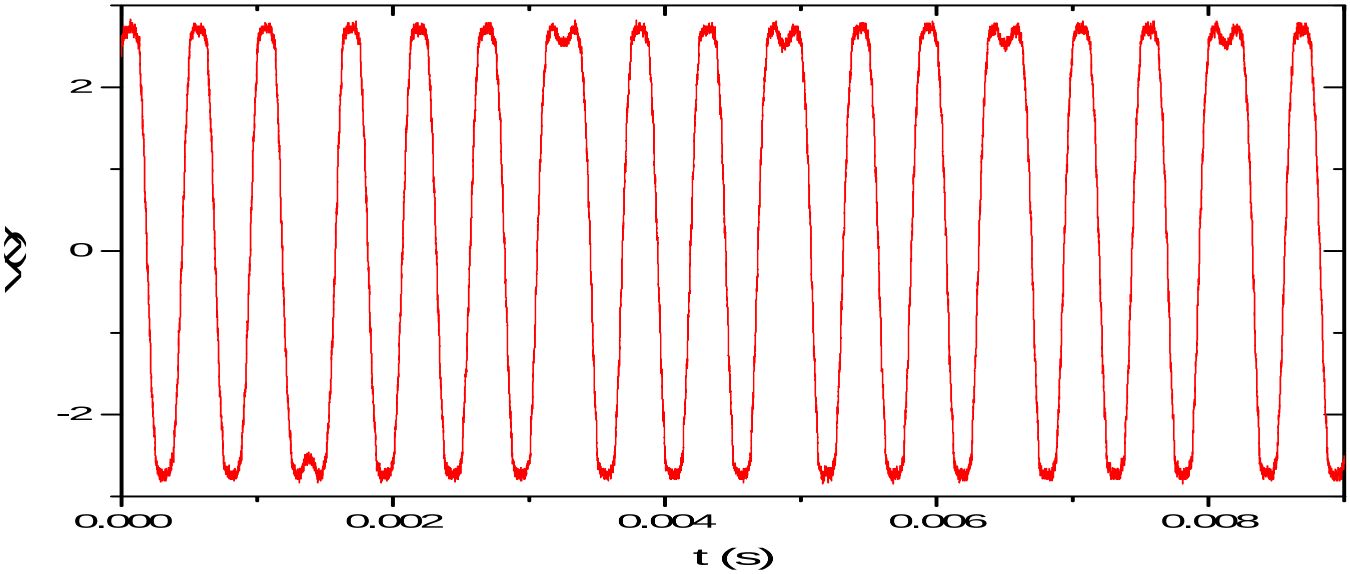

Fig. 6: Oscilloscope trace of electrical sound-card output for 1850 Hz (8 bits/8000 Hz) square waveform.

Notice that the sine waveform does not look too bad – at least not to the casual eye. The square waveform has obvious cycle-to-cycle variations and thus violates the original periodicity. The reason for this is that its construction procedure violates conditions (1) and (2) (please see above) for the Whittaker-Shannon interpolation scheme. Sound Forge generated the samples without first band limiting the square waveform. Thus the sampled waveform in the wave file (Fig. 4) has the aperiodicity mathematically trapped in it and the subsequent DSP did not fix this. The sine waveform contained in its wave file (Fig. 3) on the other hand does not violate condition (2) (i.e., it has no frequency content above Nyquist) and hence its sampled data V(n) is correct (at least within the vertical quantization error) and hence the reconstruction worked a lot better. Because the actual electronic interpolation carried out by the sound card is not an exact implementation of the Whittaker-Shannon scheme, there will still be some errors and as a result there are very small cycle-to-cycle variations causing aperiodicity even for the sine function (these errors go away for commensurate frequencies). As sampling rates and vertical resolution are increased, these errors become progressively smaller but in principle will always be present. One forum poster referred to figures posted at the web site:
http://www.hydrogenaudio.org/forums/index.php?showtopic=47827&st=100&p=427738&#entry427738
Looking visually at a few cycles on an oscilloscope is not sufficient. In our experiments a listener typically listens to ~70,000 cycles of each tone before making a judgment. How consistent are all of those cycles? Errors that might be acceptable for consumer audio are not necessarily acceptable for research. The quality of the interpolation will depend on the particular sound card used (the electrical signals shown in Figs. 5 and 6 were generated with a ~$400 Digital Audio Labs card, which had ¼" TRS phone sockets). For this reason, the combination of a software such as Sound Forge and an average sound card were judged to not be adequate for the experiment and this was the answer to the referee's questions. Now are there more elaborate schemes to better synthesize square and other waveforms of arbitrary frequency while reducing these errors (possibly to negligible levels)? Of course there are – for example, please see http://ccrma-www.stanford.edu/~stilti/papers/blit.pdf. But this was not the referee's question. Therefore, in the end, the analog square-wave generator provided a suitable and straightforward source.

Jitter

Jitter (cycle-to-cycle variability and errors in the timing of samples) must be considered when dealing with the periodic signals used in the experiment – whether the synthesis is analog or digital. Excessive jitter is audible and the threshold was shown by Pollack (1969 and 1971) to be as little as 100 ns (a variation of 0.1% in the period). In the present work, a LeCroy DSO was used to acquire a large number of individual traces of the square waves. These were then mathematically analyzed to obtain the jitter; it was found that our analog source had inaudible jitter (68 ns or <0.05% of the period). This sufficiently low jitter combined with the relatively short (20 ns) rise time made this analog source suitable for the experiment. This obviously does not imply that *all* digital audio sources, as a rule, have excessive jitter or are otherwise unsuitable audio sources. But my papers do question the "typical digital signal sources" that were used in past psychoacoustic literature (these are usually not of high-end audio standards). While digital setups do exist in which jitter levels and other problems have been sufficiently suppressed, jitter is something to be aware of. It should be remembered that it is the jitter for the entire chain that matters, not just for one stage in the process (please see condition (5) for the Whittaker-Shannon interpolation scheme above). Apparently the jitter problem is less serious in single-chassis CD players versus digital separates where the digital signal is read off the CD in a "transport" and converted to analog in an "outboard DAC" unit. In my own current listening setup, I have an older Generation Theta DSPre converter fed by a G&D Transforms transport. It was found that the digital interconnect between the two components profoundly influenced the sound. Preliminary blind tests (using music) were conducted on two subjects to assess the audibility of the different digital cables and the subjects' ability to distinguish the them. The total combined score for the two subjects was 36/36. Subsequently the single digital interconnect was replaced by two cables and a jitter-buster box (Monarchy Audio DIP Classic), which buffers, re-clocks, and reshapes the signal; this box got rid of a lot of the harshness and improved the resolution. (I understand that the current model Generation VIII Theta DSPre has a "jitter jail" at the input, which fixes this problem internally.)

There is a certain dichotomy between audio engineers and HiFi enthusiasts on the one hand (who are more knowledgeable of high-level audio gear and advanced techniques) and the formal psychoacoustics research community on the other hand (who are responsible for establishing the official standards and thresholds). The first group does not seem to conduct controlled blind tests of publishable standards (of all the JAES papers I read on this subject, I didn't find a single one that reported controlled listening tests). I wanted to bridge the different disciplines together and point out that the some of the earlier conclusions regarding the limits of hearing may actually be more closely related to limitations in the equipment. Of course if higher bandwidth drivers were used in the present work, we might also get lower thresholds. Hence, as stated in the papers, the result should be viewed as an upper bound rather than the final word on human temporal resolution. But the point is that the threshold was brought down from 12.5 to 4.7 microseconds. Also, more importantly, this is the first report of a threshold below the trivial $1/[2\pi\ f_{max}] \sim 9$ microseconds (the significance of this is explained further in the papers).

JND (just noticeable difference) for temporal resolution versus JND for intensity level (loudness)

Any audio reproduction system has a finite bandwidth. One basic question is how high does the bandwidth cutoff need to be for there to be no audible alteration; this cutoff frequency will naturally depend on the details of the filter slope, etc. A second question that follows is what type of auditory mechanism is involved in the perception – whether it is detecting differences in intensity levels (Symbolized by L or L_I, the *intensity level* is closely related to the *sound pressure level* L_p or SPL, and both are measured in dB.) or detecting differences in timing/phase. In the present experiment, both the relative harmonic phases as well as the harmonic levels are altered. It was found that the level changes in the experiments (~0.2 dB) were subliminal (four times smaller than the published level JND) making it likely that the discrimination depended on more than just level changes. My papers also propose quantitative neurophysiological models to explain what might be happening in the timing/phase domain.

One forum poster asked why I did not re-measure the level JNDs. In scientific research we have to start with what has already been published and cannot go back to the beginning of time and re-measure and reprove everything ever published (unless there is a special reason to doubt the previous results) otherwise it will be impossible to move forward. The present work took about five years. To redo the level JND thresholds properly will take at least two years. However, as far as the first question in the previous paragraph is concerned, the experiments unambiguously establish that temporal alterations on a ~5 microsecond level can be heard (regardless of which type of auditory mechanism is operative) thus establishing the need for ultrasonic bandwidth in music reproduction. These results provide a concrete basis for audiophile claims and have shown that the previously reported temporal and/or loudness thresholds are considerable overestimates and hence human hearing is more sensitive in one/both domains than previously believed.

Non-linear mixing as a mechanism for relative harmonic phases sensitivity

Some forum posters asked me to give more details about the non-linear mixing mechanism that leads to the sensitivity to relative phases of the ultrasonic harmonics. A brief description of this was given in my paper

"Temporal resolution of hearing probed by bandwidth restriction" (http://www.physics.sc.edu/kunchur/temporal.pdf). Here I give an expanded description that includes figures of the gamma-chirp modeling of the inner ear.

Time-domain models (e.g., Meddis and Hewitt, 1991; Patterson *et al.*, 1992; Patterson, 1994; Patterson *et al.*, 1995; Meddis and O'Mard, 1997; Irino and Patterson, 1997; Krumbholz and Wiegrebe, 1998; Wiegrebe and Krumbholz, 1999; Irino and Patterson, 2001; Irino and Patterson, 2006) seek to trace the evolution of the signal as it progresses from the initial acoustic stage through the basilar-membrane motion (BMM) and hair-cell transduction to an internal neural representation. The outer- and middle-ear transfer functions are modeled as broadband filters: as second-order butterworth filters with cutoff frequencies of 450 and 8000 Hz (Meddis and O'Mard, 1997; Wiegrebe and Krumbholz, 1999); or as an inversion of the equal level contours (ELC), minimal audible field (MAP), or minimal audible pressure (MAP) curves (Glasberg and Moore, 1990). This broadband filtered sound then reaches the basilar membrane, whose tonotopy can be modeled as a bank of bandpass filters. The number of filters, their bandwidth, and their response functions differ between different specific models, such as the gammatone filter (Boer, 1975; de Boer and de Jongh, 1978; Patterson et al., 1992), dynamic-compressive gammachirp filter (Irino and Patterson, 2006), etc. The model should also allow for the generation of combination tones such $f_1 - n(f_2 - f_1)$ that can arise from non-linearities in the cochlear mechanics and then propagate to their appropriate frequency channel (Patterson et al., 1995). The BMM is transduced in the inner hair cells (IHC) into a receptor potential. These IHCs have a limited temporal speed as evidenced by loss of phase locking around 3–4 kHz (Johnson, 1974; Shamma, 1989) which leads to a smoothening of temporal fine structure. This effect can be incorporated in a model by low-pass filtering the output of the filters (Carney, 1993). Early work (e.g., Viemeister, 1979) assumed the lowpass cutoff to be 60 Hz whereas Palmer and Russel (1986) showed that cutoffs in the 600–2000 Hz range are more realistic; Krumbholz and Wiegrebe (1998) used a second order lowpass cutoff of 1.1 kHz in their analysis. The hair cells are contacted by auditory nerve fibers whereby the continuously variable IHC receptor potential is converted into stochastic nerve impulses (spikes). The nerve firing only takes place on positive half cycles of the IHC potential, which aspect is modeled as a half-wave rectification step. The next step of a model is to compute the spike probability for each filter channel as a function of that filter's output intensity. This conversion from the BMM to a neural activity pattern (NAP) may be based on IHC simulation (Meddis, 1988) or a functional (e.g., adaptive thresholding) mechanism (Patterson et al., 1995). IHC/auditory-nerve adaptation that results from depletion of neurotransmitter is taken into account in models such as Sumner (2002). In addition, the auditory periphery undergoes adaptation on longer time scales, which has been incorporated by feedback loops with time constants in the 5–500 ms range (Kohlrausch and Puschel, 1988; Kohlrausch et al., 1992; Dau et al., 1996). Once the NAP has been computed for two stimuli, the corresponding discriminability index can be estimated from the correlations and variations between different NAP instances for each stimulus and a template (average over several instances of the control pattern). Variability between different NAP instances for the same stimulus arises from the spontaneous discharge rate and stochasticity of nerve firing. This randomness can be incorporated into a model through Gaussian noise (Dau et al., 1996). For certain temporal tasks (e.g., edge detection) it may be appropriate to sum over different channels of an NAP; it is known from physiology that such synchronous cross-frequency comparisons take place in the octopus neurons in the posteroventral cochlear nucleus (Golding et al., 1995; Ferragamo and Oertel, 1998; Golding et al., 1999; Oertel et al., 2000). Some examples of recent experiments that probed temporal integration (temporal window Δt over which signal energy is summed up) and temporal resolution (the ability to distinguish quick fluctuations in the instantaneous signal amplitude), in which the results could be well fitted by models, are the works by Krumbholz and Wiegrebe (1998) and Wiegrebe and Krumbholz (1999) respectively. They found that their obervations could be fit by modeling the peripheral auditory filtering with a gammatone filterbank.

In the present experiment, the two acoustic stimuli being compared are both long-duration steady complex tones whose essential compositions are well approximated by:
$K''[0.98 \cos(2\pi 7000t) + 0.18 \cos(2\pi 21000t + \varphi''_A)]$ for Tone A (4.7 μs filtered) and
$K''[\cos(2\pi 7000t)+0.22 \cos(2\pi 21000t+\varphi''_B)]$ for Tone B (unfiltered).
By the time the signals arrive at the cochlea, external- and middle-ear filtering change these signals to (applying an "ELC" correction as per Glasberg and Moore, 1990):

$K'[0.98 \cos(2\pi 7000t)+0.15 \cos(2\pi 21000t+\varphi'_A)]$ for Tone A (4.7 μs filtered) and
$K'[\cos(2\pi 7000t)+0.19 \cos(2\pi 21000t+\varphi'_B)]$ for Tone B (unfiltered).

The weak 21 kHz component is far outside the bandwidth of the highest filterbank channel and so we will assume that it will not directly contribute to the NAP. However nonlinearities in cochlear mechanics and the preceding mechanical chain can generate an audible 14 kHz component. For a nonlinear response represented by $y \alpha x+bx^2$ (with $b < 0$ for a compressive nonlinearity), an input consisting of a fundamental and third harmonic mixture $x \alpha \cos(\omega_0 t) + a\cos(3\omega_0 t+\theta)$ will give rise to a response $y \alpha \cos \omega_0 t + {}^{b}\!/\!{}_{2} \cos(2\omega_0 t) + ab \cos(2\omega_0 t+\theta)$ (keeping oscillating terms up to $2\omega_0$ in frequency). The second term, with $2\omega_0$, comes from doubling the fundamental and maintains the same phase; the last term with $2\omega_0$ arises as a difference tone between ω_0 and $3\omega_0$ (in the input) and maintains their original phase difference θ. Applying this nonlinearity to the previous middle-ear filtered signals gives the effective signals feeding the BMM filterbanks:
$K[0.98 \cos(2\pi 7000t) + b\{0.5 \text{ x } 0.98^2 \cos(2\pi 14000t) + 0.15 \text{ x } 0.98 \cos(2\pi 14000t + \varphi_A)\}]$ (Tone A) and
$K[\cos(2\pi 7000t) + b\{0.5 \cos(2\pi 14000t) + 0.19 \cos(2\pi 14000t + \varphi_B)\}]$ (Tone B). The phase difference
$\Delta\varphi = \varphi_B - \varphi_A = \varphi''_B - \varphi''_A = \{\tan^{-1}(-2\pi 7000\tau) - \tan^{-1}(-2\pi 21000\tau)\} = 20^\circ$ is preserved during the nonlinear mixing as explained earlier. The maximum difference in levels of the nonlinearly generated second harmonic between the tones A and B can now be calculated:
$\Delta L_p(14 \text{ kHz}) = 10 \log([0.5+0.19 \sin(\Delta\varphi/2)]^2/[0.5 \text{ x } 0.98^2 + 0.15 \text{ x } 0.98 \sin(-\Delta\varphi/2)]^2) = 1.4$ dB.

This level change is eight times larger than the subliminal $\Delta L_p(7 \text{ kHz}) = 0.18$ dB of the fundamental, thus allowing a listener to discern the phase shift of the 21 kHz component (and hence waveform shape) indirectly through an interference between the two (quadratic and difference) 14 kHz nonlinear products. There is a third potentially interfering 14 kHz contribution that has not been considered up to now—this arises from the fact that the 14 kHz filterbank channel will receive some excitation directly from the 7 kHz signal because of the finite bandwidth of the channel. To analyze this quantitatively, we will take the impulse response of each filterbank channel to be given by the gammachirp function (Irino and Patterson, 1997; Irino and Patterson, 2001; Irino and Patterson, 2006):
$g_c(t) = At^{(n-1)} \exp(-2\pi b \text{ ERB}(f_r)t) \cos(2\pi f_r t + c \ln t + \varphi)$, where $A, n, b, f_r, c,$ and φ are parameters. ERB(f_r) is the equivalent rectangular bandwidth given by ERB(f_r) = $24.7+0.108f_r$; the level dependence of the bandwidth is contained in the parameter b. For the present experiment, the applicable values can be taken as $n \approx 4$, $b \approx 1.3$, $\varphi \approx 0$, and $c \approx -3$ (Irino and Patterson, 2001). Fig. 7(a) shows the impulse response of this filter, for f_r=7760 Hz, and panel (b) shows its amplitude spectrum: $|G_c(f)| = |A\Gamma(n+ic)|e^{c\theta}/|2\pi b \text{ ERB}(f_r)+i2\pi(f - f_r)|^n$, where $\theta = \arg\{2\pi b \text{ ERB}(f_r) + i2\pi(f - f_r)\}$ and $i =\sqrt{-1}$ (Irino and Patterson, 1997). Note that the peak (center) frequency of $f_c = 7$ kHz occurs lower than f_r and that the filter skirts are asymmetric, cutting off more sharply on the high frequency side. Panel (c) shows the time-domain response to a sinusoidal signal (with $f = f_c$) that started at $t = 0$:
$A \int_{0 \text{ to } t} \{\sin(2\pi f_c s)(t - s)^{(n-1)} \exp(-2\pi b \text{ ERB}(f_r)(t - s)) \cos(2\pi f_r(t - s) + c \ln(t - s) + \varphi)ds$.

As can be seen the output settles to a constant level in ~1 ms. Thus for the long-duration steady tones of the present experiment, the filter outputs will be steady amplitudes (i.e., the chirp and other time dependence will have settled down). What is the distribution of these excitation levels across the filterbank? Fig. 7(d) shows this excitation spectrum as a function of channel frequency when one pure tone is applied at a time (independent plots for 7 kHz and 14 kHz are shown on the same graph). The 14 kHz input frequency makes negligible contributions to the channels in the 7 kHz vicinity. The tail of the 7 kHz input does extend into the 14 kHz territory, but falls 22 dB below the direct excitation from a 14 kHz input of equal magnitude. Thus for the 7 kHz tail to not drown out the 14 kHz's own excitation, the latter needs to have a power level that is no more than 22 dB below the fundamental. The implied degree of nonlinearity, $b^2 > 10^{-22/10} = 0.006$, seems plausible.

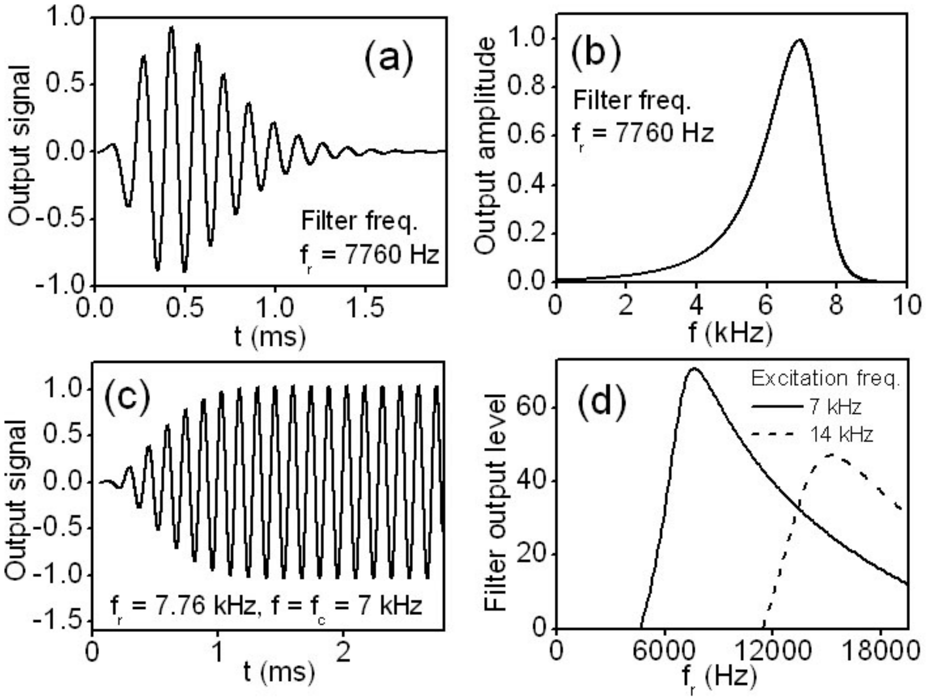

FIG. 7: Modeling the cochlear basilar membrane motion as a bank of gammachirp filters. (a) Gammachirp impulse response of a single filter channel with f_r=7760 Hz. (b) Amplitude spectrum of this filter channel as a function of stimulus frequency. Note that the center (peak) frequency f_c=7 kHz is lower than f_r=7.76 kHz. (c) Time domain filter response to a semi-infinite (turned on at t = 0) sine wave whose f = f_c. (d) Distribution of relative output levels across the filterbank for single pure tones (two independent curves, for 7 kHz and 14 kHz, are plotted on the same graph); here the abscissa is the frequency of filterbank channel.

Please also look at the work by Zwicker, E. (1981).

Some other more specific questions

1. “Why no high resolution graphs on distortion from the playback chain? Seems like at least one visible IMD product in the graph. RS-1 [pair of headphones] has ...”
>>> All the analysis is carried out on the actual measured acoustic output of the headphone. Therefore how this output relates to the input (i.e., the amount of distortion present) does not matter. As the paper reports, the effect of the low-pass filter on the harmonics is exactly as per its theoretical response – please see the paragraph "The measured attenuations in harmonic levels..." after Table 1.

2. “At one place a 2MS/s was mentioned and at another place 20MS/s. That's surprising and I assume a typo that found its way to print even though peer reviewed?”
>>>This is not a typo. For smaller frequency steps in the measured spectrum, a longer duration scan has to be analyzed. Since the digital scope has a fixed total number of samples (record length) this meant using a lower (2MS/s) sampling frequency to obtain the 20 Hz spectral resolution. On the other hand, for the time-domain graph it was desirable to use the higher sampling rate but with a shorter time window.

3. “How does Dr. Kunchur discount that perception of a sliding loudspeaker isn't a consequence of the ear's sensitivity to loudness? … How was the mechanical device able to slide the loudspeaker without the person hearing the slide.”
>>>The fractional change in intensity of the signal from the moved speaker is negligible (given approximately by the inverse square law): ~[(4.3m + 0.002m)/(4.3m)]^2; this corresponds to (10 x log of that fraction) ~ 0.004 dB change in SPL. Of greater importance is the change in intensity level (loudness) caused by the interference between the two (slightly out of phase) signals from the individual speakers. This change (of 0.2 dB) is also subliminal (four times less than the accepted loudness JND of 0.7 dB). Please see the earlier discussion about JNDs. The speakers are mounted on top of each other separated by smooth lubricated rails so that there is no audible noise during their motion and the operator running the experiment does not talk to the subjects during

the testing.

4. “Is it plausible that transducers with minimal distortion could yield lower levels of ultrasonic audibility...”
>>>No. As per the answer to question no. 1 above, the analysis (changes in harmonic levels and relative phases) is carried out on the actual measured acoustic output signals fed to the listener. The conclusion that we have observed a new reduced upper bound of ~5 microseconds (versus the previous 12.5) is therefore solid.

5. “...novel power spectrum measurement algorithm written in C without the use of an FFT, but does not go into any more detail. What is the nature of this algorithm? What windowing methods are used?”
>>>The power spectrum measurement was not claimed to be novel. It is an exact discrete transform (basically following classic textbook equations). Shown are signal peaks not sums of powers in a band around each frequency component; the frequency steps were made smaller and smaller until the peak heights stopped changing and were sufficiently resolved. Hence the reason for the longer record lengths used (as per answer to question no. 2 above).

6. “My question is a not specifically about the test methodology or results …but rather why the Physics and Astronomy Dept …is apparently doing research in a separate, only peripherally related field, ie, the branch of psychology studying the limitations of human perception…”
>>> I teach a course on Musical Acoustics (syllabus at http://www.physics.sc.edu/kunchur/p155-syllabus.htm), am an audiophile, an electronics designer, and have experienced noise-induced hearing loss myself (mainly from riding sports motorcycles and listening to too much music). These factors give me an additional personal incentive into delving deeply into what goes on with human hearing and sound perception. Beside this research in psychoacoustics and auditory neurophysiology, I also have an active research program in another area, which is investigating high-speed electromagnetic responses of superconductors and nano-materials (please see my web site for details: http://www.physics.sc.edu/kunchur). For the superconductivity research, I routinely have to develop instrumentation for generating and detecting fast (sub-nanosecond) signals.

7. “Question two would be how do *these* results relate to the body of work already extant from those actually working in that sepcific field? I'll admit that I'm not up to the minute in my readings in this area, but the conclusions drawn here seem to be at odds with existing research and thinking…”
>>> Absolute true! My results created a stir in the field (in the words of one of the editors, the results "...draw into question 100 years of research on the temporal resolving capability of the ear..."). It was a very long process to convince the professional community and to satisfy them that all checks and cross checks had made. The instrumentation, test procedures, and general approach were new to the field. Prof. Thomas Rossing (author of ~16 books and a former president of the Acoustical Society of America (ASA) and American Association of Physics Teachers (AAPT)) visited our home once and inquired why the subwoofers (VMPS) were coplanar (time aligned) with the main speakers (ProAc Studio 1). When I told him I could hear a difference if they were moved by half an inch, he thought that that was preposterous; he expected a movement of at least a foot to hear a difference. This was one of the events that made me realize the importance of looking into the possibility of establishing anecdotal HiFi claims as scientific facts in refereed scientific journals. It took 5 years to achieve this. Each experiment had to be carefully thought out and then submitted as a proposal for approval to an Institutional Review Board (IRB), which is also responsible for legally approving the consent forms that must be used. Then optimum equipment, methods, and a multitude of cross checks must be developed (some details are given in the papers). It takes about half a year to conduct each sequence of controlled blind tests; several sequences were conducted. The results, analysis, and conclusions were then carefully considered and discussed with colleagues who are experts in their related inter-disciplinary fields; for this I went in person to various universities and research institutes and met with people in departments of physics, engineering, psychology, neuroscience, music, communications sciences, physiology, and materials science (as noted in the acknowledgements lists at the end of the papers). After that the results and conclusions were presented at conferences of the Acoustical Society of America (ASA), Association of Research in Otolaryngology (ARO), and American Physical Society (APS). Seminars were also given at numerous universities and research/industrial institutions (please see the list on my web site). After each presentation, the audience is free to tear apart the conclusions and ask all possible questions. Eminent people such as presidents of the above

mentioned societies and corporations were present at my presentations and engaged in the discussions. After this oral presentation process, written manuscripts were then submitted to journals where more than a dozen referees and editors were involved in the refereeing process. Now with the three refereed publications in print the psychoacoustics and auditory neurophysiology communities have accepted the results and methods.

8. This is a totally serious question. I would like to know what playback format Dr. Kunchur personally prefers for his own listening? Is he a vinyl guy? A tape guy? A PCM (compact disc) guy? Or SACD perhaps? What does he feel is the most musical delivery medium out there?
>>>My own system is quite modest (picture posted at http://www.physics.sc.edu/kunchur/HiFiPic.jpg). I listen equally to cassettes (Nakamichi CR3 deck), LPs (Thorens turntable), and CDs (Theta DSPre DAC, Monarchy jitter buster, and a G&D Transforms transport). I listen to all kinds of music from zero to infinity.

I hope this answers the bulk of the questions asked on the forums. All the best to everyone out there!

Sincerely,
Milind Kunchur
P.S.: A succinct and highly readable article by George Foster, summarizing the essence of my work and conveying the main conclusions in an easy to understand manner, has appeared in the HiFi Critic magazine (Volume 3, Issue 2, Page 7, April/May/June 2009).

****
Milind N. Kunchur, Ph.D.
Professor of Physics
Department of Physics and Astronomy, University of South Carolina
Columbia, SC 29208. Phone: 803 777 1907 FAX: 803 777 3065
Email: kunchur@sc.edu   Web: http://www.physics.sc.edu/kunchur
****