2005-01-0567

Masking Effects of Audio Systems with Non-flat Frequency Response

Richard S. Stroud
Stroud Audio Inc.

ABSTRACT

An automotive audio system typically demonstrates an irregular frequency response. For broad-spectrum signals like rim shots or cymbals, simultaneous masking caused by response peaks masks weaker portions of these signals. This could result in a perceived loss of energy and liveliness. This paper discusses this hypothesis, and experiments focused on relating frequency response irregularities to the perception of listening system liveliness.

INTRODUCTION

Audiophiles tell us that the life-like presentation of a fine listening system can be easily heard. Transient sounds that are subdued on lesser systems seem to be sharper and more real, more lifelike.

Audiophiles also claim that this kind of quality cannot be measured. That seems unlikely, given our very high quality microphones and sophisticated analyzers. But that elusive quality of accurate response to recorded transient sounds just doesn't seem to show up in our measurements.

Or perhaps it does. In fact, we may already be measuring this "liveliness" property when we perform a frequency response measurement. We are perhaps smoothing away the information, or may not be correctly interpreting our data.

Proposed is an hypothesis that states: irregularities in a non-flat listening system's frequency response can create conditions for the psychoacoustic masking of a significant amount of sound information when presented with broad-spectrum sounds. This masking reduces the perceived energy and liveliness of those sounds.

PROPERTY OF "LIVELINESS QUALITY"

The property being described has been called "clarity", "nonmasked clarity", etc. These words are inadequate, and the word "clarity" can be also be used to describe non-linear distortion. Henceforth in this paper, this property will be known as "LQ", short for "Liveliness Quality".

Before this is explained in more detail, let me provide some background. I have heard LQ on high-end audiophile loudspeakers and on $12 headphones. I've heard LQ on non-flat systems that shouldn't have sounded that good. And I've heard LQ in the men's room, standing near one corner, at Kokomo's Olive Garden.

I have also heard LQ destroyed by installing a very good speaker in an environment with very bad nearby reflecting surfaces.

Many, including myself, have chased intermodulation distortion as the cause of all loss of LQ but this has not been totally satisfying. I now believe that what we hear as system "clarity" relates to amplitude nonlinearity, and LQ to simultaneous masking.

Intermodulation distortion produced from amplitude nonlinearity is certainly a clarity killer. It is most easily heard on massed instruments, and to my ears creates a noise floor and/or "grunge" under complex sound passages. Two and three-signal intermodulation tests, Mark Ziemba's POSDAM, and Delphi's Complex-Wave Intermodulation Distortion test, among others, measure this type of distortion.

> Dr. Earl Geddes[1] and Dr Lydia Lee are testing a new distortion metric that may correlate much better with perception than traditional IM measurement. See: http://www.gedlee.com/distortion_perception.htm

But intermodulation distortion doesn't work as a reason why broad-spectrum signals such as rim shots and stick taps don't sound real; don't have LQ. Distortion products would be buried under the undistorted broad-spectrum signal, and thus be inaudible.

SIMULTANEOUS MASKING AND LQ

We need to look away from distortion, I think at the phenomena of simultaneous masking. Restating the paper's hypothesis with more detail; when a broad-spectrum signal is present, sound energy in frequency response peaks creates psychoacoustic masking of energy in nearby response valleys. Especially valleys that are higher in frequency than those peaks.

The overall perception of transient sound energy should seem lower on systems with ragged response than on systems with smoother response. At the very least, perceived frequency response of impulsive or other broad-spectral sounds could be significantly altered. This alteration could in fact be much greater than would be inferred from looking at a frequency response plot. The consequence is that listening systems that produce this simultaneous masking may seem less lifelike reproducing broad-spectral sounds.

A "masker-based" analysis of frequency response should be able to determine if, and how much, sound energy is being masked. Note that non-flat response characteristics would not necessarily cause this masking. Measurement of frequency response can be combined mathematically with appropriate psychoacoustic masking curves. This will determine frequency response ranges that are masked by peaks on broad-spectrum signals, and potentially give us a way to measure LQ.

Masking curve information is commonly used in the design of perceptual coders. There is not just one curve. Zwicker's[2] and Moore's[3] maskers don't quite align at lower frequencies, individuals vary, and masking curves change with sound level. For the example I will shortly discuss, I will assume a masking curve having a downward masking slope of 100 dB / octave and upward slope of 45 dB / Octave (noise maskers tend be even flatter than this, but we'll start here). On the response chart, I also rounded the top to create a masking "hairpin".

All masking curves indicate a stronger masking of frequencies higher than the masker. I suspect this is a reason that some listeners feel the need to boost treble, thus reducing the effect of this higher frequency masking. Also, it's been observed that a system with very good LQ will tolerate a considerable amount of treble cut and not lose its lifelike sound.

A MASKING EXAMPLE

Figure 1 shows an undesirable, but somewhat typical automotive frequency response plot. One might wish to boost treble and cut bass, but typical user equalization could not address the many peaks and valleys in response. Response peaks are filled with energy on broad-spectrum signals and are thus capable of masking sounds in valleys near these peaks. This response is measured with a moving microphone and using a 1024 bin FFT.

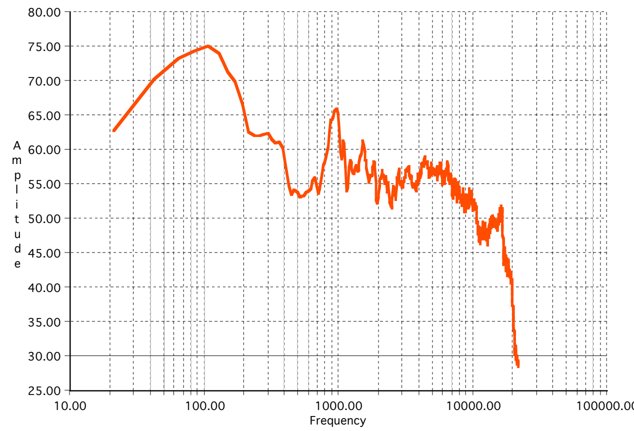

Figure 1. Frequency Response of 2003 Pontiac Vibe

Now let's examine the measured response curve again for broad-spectrum signal masking. For each of the midrange and treble frequency peaks, a candidate masking curve (those "hairpins") has been added. The parts of the response curve that are on or below the masking curves represent spectral regions that cannot be heard when broad-spectrum sounds are present (Fig. 2).

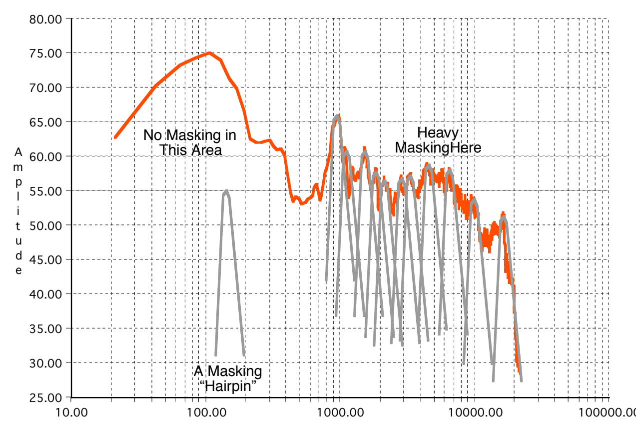

Figure 2. Response with Masking "Hairpins".

Looking at Figure 2, the reader can see that in the perception domain, the frequency response would drop to zero in masked regions, thus potentially impacting broad-spectrum fidelity rather dramatically. In essence, the response is likely much different in the perception domain than in the measurement domain. An instrument that produces discrete harmonics would certainly change timbre with such a response, but broad-spectrum instruments could seem attenuated, lack LQ.

> One would also expect that instruments other than the "impulsive masker" would be temporarily inaudible while a peak excitation exists. This may not be a perceptible effect.

The masking produced is likely worse than that shown in Figure 2. These illustrations presume a spectrally flat stimulus. Real world, broad-spectrum signals have energy that tends to diminish with frequency. This makes upward masking worse than it would appear from the illustration in Figure 2. And since masking curve slopes decrease as sound levels increase, energy peaks created by impulsive sounds would result in even stronger masking, especially upward masking.

ARE HOME SYSTEMS BETTER?

Perhaps the reader is thinking that a fine home system would never feature a ragged response like that of the above example. You should then look at the home speaker's response at the listening position, with all the effects of floor, walls, etc.

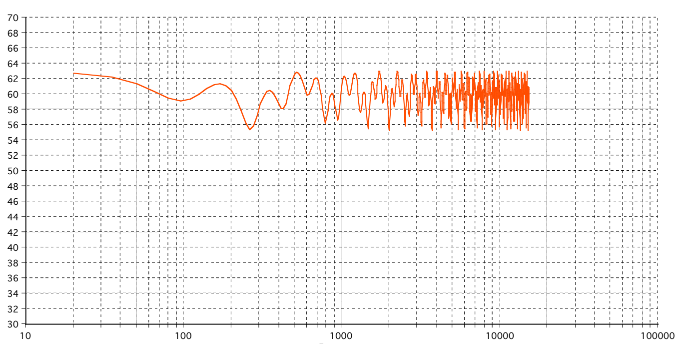

Figure 3. Effects of Reflection from Nearby Surfaces.

Figure 3 is the result of a simulation of a home speaker mounted 1 M above a hardwood floor and 2 M from a drywall covered wall (no ceiling bounce included). The listener is 3M from the speaker. A speaker's actual frequency capabilities and power response are not considered and it is assumed perfectly flat for the illustration below. The floor and wall are assumed to have a reflections coefficient of 0.3 and 0.6 respectively (no frequency-selective absorption effects are calculated by this simple program).

Now imagine your favorite speaker's response, with cabinet diffraction and time alignment issues, convolved with the curve above. The automobile may not seem quite that bad after all.

MORE THOUGHTS AND OBSERVATIONS ON LQ

Note that time-related interference effects become more dense on a logarithmic plot, while the masking curves appear relatively constant on this scaling. This shows why simultaneous masking is likely a mid-to-high frequency problem, mostly in the 1 kHz to 10 kHz region.

Based on past observation, I suspect that reflections arriving within the first 11 mS should be included in any LQ analysis of response (fusion of impulsive sounds occurs with intervals under 11 mS).

Systems with a second set of speakers delayed more than about 18 mS produce to my ears a strong sensation of LQ on impulsive signals, This LQ improvement is likely because peaks from one system fill in valleys from the other, and the fill is beyond fusion range for impulsive sounds. Unfortunately, delays of this magnitude also produce unacceptable artifacts on most music.

Large-panel home speakers like electrostatics generally provide a higher ratio of direct to reflected sound to the listener (with proper rear-wall consideration), and I have almost always found these speakers to have more lifelike impulsive signals and high levels of LQ.

Moving microphones, commonly used in automotive measurement, may "smooth away" LQ characteristics. A simple experiment with headphones can determine if sounds masked by one ear and supplied to the other ear can still produce the perception of high LQ. If not, moving microphone measurement may not be suitable for LQ assessment.

Additional literature research and experimentation will be necessary to determine what bandwidths and amplitude differentials of masker peaks to masked valleys will produce the masking effect. Excess smoothing of frequency response data, common for marketing presentation, should almost certainly be avoided for this evaluation.

A single tone can provide a good deal of masking; I suspect even an energized narrow response peak can do likewise.

> If one is perhaps thinking that a very narrow peak would not contain enough energy to do masking, realize that the potentially masked region is filled with even less energy per Hz. Broad-spectrum maskers behave differently than tones and their noise-masking curves are very aggressive.

EXPERIMENT ON MASKING HYPOTHESIS AND ORTHOGONALITY

The goals of the first LQ experiment are to 1) determine if the irregular frequency response produced by comb filtering produces an audible effect on a pure impulsive sound, 2) determine if the addition of said filtering materially changes the perception of the frequency response of non-impulsive music, and 3) see if this irregular response reduces the audibility of impulsive sounds, using music, pink, or red noise as maskers.

In parts 1 and 2 of the experiment, the listener is asked to comment on differences heard. For the music listening element, the listener is asked if there is a need for tone adjustment between the two music samples, combed and non-combed, to make them more alike.

Basically, part 3 of the experiment uses a pure impulse as a probe signal. Music and noise are used to establish a masking threshold for the listener. The experiment's methodology is to adjust the level of the impulse to determine if the impulse is less audible with a combed frequency response than one with a flatter response.

The experiment was setup as a randomized, single blind test with minimal communication between the test administrator and the subject. Also, the administrator was seated out of the subject's view.

For this experiment, Ableton Live®, a software-based "sequencing instrument", is employed (Fig. 4). This flexible software records test music and test signal "clips", and allows various gain and track selection setups to be saved as "Scenes" that can quickly be switched among the conditions to be evaluated.

Figure 4: Ableton Live® Screen Shot: LQ Experiment One

Music that is relatively free from impulsive or other broad-spectrum sounds is recorded into an 18 second clip and this clip is placed on tracks where needed. The first track is low-passed and the other two are high-passed, with a crossover around 2.5 kHz. One of the two high-passed tracks is comb filtered using a time delay of 2 mS and a 50-50 "dry-wet" balance. For testing, the LF and one of the HP tracks are combined. The HP-LP setup is used, as combing of the HF range more closely resembles that of real-world characteristics.

A pure narrow impulse stream is recorded on a second 18 second clip and this clip is put on two additional tracks. The narrow impulse produces a brief, flat spectrum for each "tic". One of the two tracks is comb filtered as in the previous paragraph.

Since the experiment is about impulse audibility, the tone is bandpassed in the treble range (Fig. 5). This is required because masking produced by the comb used would not occur at lower frequencies, and audibility of the pulse probe signal would not change nearly as much.

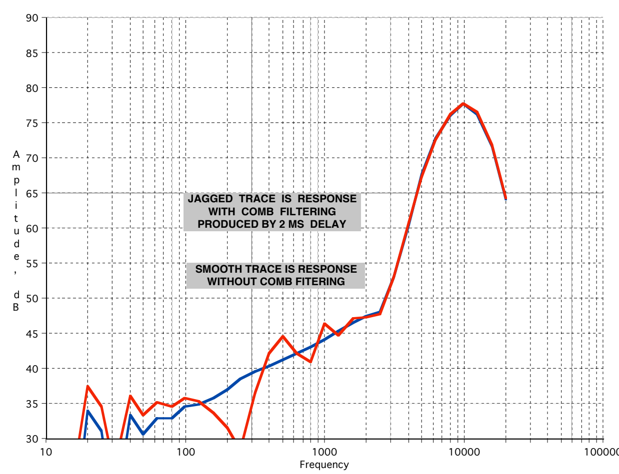

Figure 5: Bandpass Characteristic of Impulse Probe Signal

Note also that there are two curves in Figure 5. These represent the third-octave (Real Time Analyzer) spectrum and show that for higher frequencies, there are no broad spectral differences caused by the comb. This relates to orthogonality comments made later in this paper.

> A 3dB gain shift was needed to match the curves, this due to the implementation of Ableton's "Simple Delay" plug-in set at 50%.

Although the focus of this paper is automotive sound, a reference listening room and high-grade speakers were used for this experiment. This was to optimize system parameter control, and to provide the subjects with at least one high LQ test element.

The evaluators included four highly trained listeners, two "audiophiles" and two with perhaps lesser qualification. Playing the recorded material through the listening room's speakers, trained listeners performed an evaluation of the non-combed and comb filtered sounds of the music tracks. Again, these parts 1 and 2 of the experiment were to determine that there was, or was not, a major frequency response effect caused by processing the music with a dense comb.

The listeners did not perceive a significant frequency response impact from combing the music clip, but did mention a loss of "air" and "phasiness". The guitarist thought it sounded like a flanging effect. There was little indication that treble boost would be needed to match the two signals. The effect is subtle, and I suspect differences heard relate to broad-spectral breath sounds in the music.

Conversely, combing was perceived to have a major impact on the impulsive test signal. References to "timbre shift", "duller" and "higher pitch" (for non-combed signal) was common.

The effect of combing on impulsive signals is not a subtle effect, and can be heard by casual observation. I would describe it as the difference between a "snap" that sounds like a spark and something that sounds like the "recording of a spark", but that is not all that much quieter.

Next came part 3 of the experiment. Three different maskers are provided, and the non-combed and combed impulse track adjusted to a "just audible" level, and the level settings are noted.

Red noise, pink noise and music masker tests all showed only about 1 dB of difference in audibility. While this was generally in favor of the non-combed signal the small difference is a disappointing result.

I had thought that the acoustic combing observed in the listening room could be affecting the results, and changes were made, including only using one speaker, and installing a "floor bounce" shield. This significantly reduced combing in the reference system, but did not improve the masker test data.

I have later repeated the experiment on myself, using very good headphones and flat-panel speakers, still getting about a 1 dB result.

Perhaps I should not be surprised, as the levels of the two signals were matched in level (by third-octave measurement), but I still think masking would make the perception domain's result greater than a 1 dB difference.

Summarizing the results of this first experiment:

- Response combing produces little effect on non-impulsive music.
- Combing produces significant perception differences on impulsive test signal.
- Masking tests intended to show improved audibility of non-combed signal show very little in the way of encouraging results.

While the first two points above lend support for the hypothesis, the third suggests the need for additional thinking and testing.

EXPERIMENT ON EVALUATING AND RANK ORDERING SPEAKERS FOR LQ

A second experiment could be run to determine if speakers can be subjectively evaluated and rank ordered by listeners and that same rank order be determined by response-masker analysis.

A panel of listeners would sort speakers listening to an impulsive test signal similar to that used in the first experiment.

Second, the speaker's measured response in the listening position would be analyzed with "hairpin" masker algorithms (see figure 2) to determine a second rank ordering.

The rank orderings are then compared. If correlation is strong enough, a third experiment will be performed. The same speakers rank ordered by the tests above are again rank ordered on well-recorded music by a panel of trained listeners. These listeners will have no knowledge of the first rankings, and will be listening to music in the same position as used in the first part of the experiment. The rank orders are again compared.

If the rank orderings of all three experimental sections correlate well, the hypothesis will gather strength as a basis for a potential new LQ test.

The probe signal test and "hairpin" masking analysis can be modified if necessary to better correlate with expert listening data. But any masking test modifications should conform to a psychoacoustically valid rationale or they should not be done. In any case, the test would need to be validated a number of times with many home and automotive audio systems before it could become an accepted standard.

This second experiment is a longer-term task and will hopefully be completed by late 2005.

A word about smoothing: it is hard to imagine relevant masking occurring within a 1/12 octave band. Peaks and valleys contained within this small bandwidth should perhaps be eliminated from LQ measurement by data smoothing. This remains to be evaluated.

CLOSING THOUGHTS

I would expect most garden-variety speakers to have lower LQ scores and many audiophile speakers to do better. But I think my relatively inexpensive, great sounding Mission Acoustics speakers would score well too. Likewise, I would expect automotive systems, because of their reflective environment, to generally have lower LQ scores than home listening systems. That said, I have heard several automotive systems with very good LQ.

Can equalization add LQ to a lesser system? This is likely, especially if the equalizer has sufficient adjustability. And since maskers have a stronger effect on higher frequencies, the EQ of a dull system could almost surely improve LQ scores. Future experimentation will test whether parametric equalization can make a bad speaker better and a good speaker worse, both in LQ measurement and in listening evaluation.

Distributed-Mode loudspeakers, with their somewhat different response curves, could score poorly in a non-reflective environment. Or possibly not. Their spectral ripple density is very high and the response valleys are so narrow they may not affect the LQ perception. Also, the spatial averaging between the two ears may ameliorate this issue (would learn with the moving microphone headphone test already mentioned). DML's may also do better in a reflective environment like an automobile, where the differently colored early reflections can fill in response valleys.

What about sounds other than rim shots, sticks and castanets? The initial sounds of many instruments that start suddenly, like a piano, are broader in spectrum than their sustaining sounds. Reproducing systems with good LQ scores may impart a sense that those instruments are more real.

CONCLUSION

It is hoped that proposed measurement and analysis techniques shown in this paper could encourage better improved automotive and home system design. These better systems could provide significant broad-spectrum sound improvement. Cymbals might sound more lifelike. The salty sense of ocean surf would excite us. Sibilants and plosives would sound more balanced with harmonic instruments.

And singers with breathy voices, like Nora Jones, would come a bit closer. That's a good thing.

ACKNOWLEDGMENTS

I would like to acknowledge David Clark for his listener training and Mark Ziemba for sharing my belief that "clarity" (which I now call LQ), and a measurement for it, exist.

REFERENCES

1. Geddes E., Lee L. (2004). "Perception". Retrieved November 30, 2004, from the World Wide Web: http://www.gedlee.com/distortion_perception.htm

2. Zwicker E., Zwicker T., "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System", Journal of Audio Engineering Society, vol. 39, No. 3, March 1991

3. Moore B., "Masking in the Human Auditory System", in: N. Gilchrist, C. Grewin, eds., Collected Papers on Digital Audio Bit-Rate Reduction, Audio Engineering Society, 1996

CONTACT

Please feel free to contact the author:
Email: rsstroud@surbest.net
Mail: Richard S. Stroud
Stroud Audio Inc.
4707 Wexmoor Dr
Kokomo IN 46902-9596
Phone: 1-765-963-3030