Echo test

Report written by Marcel van de Gevel, experiment done by Marcel van de Gevel and Ken Newton, July 2021

1. Introduction

According to Lagadec and Stockham [1], linear-phase filters having equally spaced ripples in their passbands have a pre- and a post-echo. The FIR filters typically used in digital audio are linear phase and have ripples in their passbands that are approximately evenly spaced. Unlike the ultrasonic pre- and post-ringing, these passband-ripple-related pre- and post-echoes affect the part of the audio band that is traditionally regarded as audible for humans. The present experiment was meant to check if they can be big enough to indeed be audible.

2. Experimental set-up

Eight 44.1 kHz sample rate, 16 bit stereo recordings were processed using the expression evaluator of the GoldWave digital audio editor. Four of these tracks were really used in the experiment. The music was selected by Ken Newton, except for a harpsichord track selected by Marcel van de Gevel. Three versions were made of each:

A. One that was only dithered (2 LSB peak-peak triangular PDF) and rounded to 16 bits

B. One that got a pre- and post-echo similar in level to that of the SAA7220, got dithered and rounded to 16 bits

C. One that got a pre- and post-echo equivalent to approximately 0.2 dB peak-peak ripple, got dithered and rounded to 16 bits

Besides, whenever the peak level of the original signal was greater than -3.01 dBFS, the signal was first multiplied by 0.7071 to prevent intersample overshoot issues.

According to [2], the SAA7220 has a peak-to-peak passband ripple of about 0.02 dB to 0.03 dB and according to [3], its length is 120 taps at 176.4 kHz sample rate. The echoes usually occur close to the ends of the filter, so at ±60 taps from the centre at 176.4 kHz, equivalent to ±15 taps at 44.1 kHz. The expression used for the SAA7220-like pre- and post-echo was:

864.21542E-6*(wave(n)+wave(n+30))+wave(n+15)

giving a peak-peak ripple of 0.030025949 dB.

For the 0.2 dB peak-peak ripple (actually 0.2011647148 dB) filter:

5.78972713E-3*(wave(n)+wave(n+88))+wave(n+44)

This corresponds to a filter of almost 2 ms long, which is a bit on the long side for an interpolation filter and should reduce forward and backward masking compared to a shorter filter.

Dithering was done with the expression

wave(n)+rand(1/y)+rand(1/y)-1/y with y = 32768

The order in which these versions were created was randomized. They got _1, _2 and _3 added to their file names, indicating which was generated first, second and third (which didn't say anything about which was which). Only Marcel van de Gevel knew which was which, and he made sure not to communicate this to anyone else until the experiment was finished. As he and the participants were never in the same room together and also didn't hold any video or audio conferences, unintended nonverbal communication (as in cold reading) wasn't possible.

The files were put on a file sharing service. Anyone from the diyaudio.com community could download and listen to them via their own equipment. The participants were asked to rate which version they liked best, which medium and which worst, and to send the results to Marcel by personal message.

The hypotheses to be tested and the way to calculate the probabilities of excess were agreed upon between Marcel van de Gevel and Ken Newton before the experiment started, although there was some miscommunication about hypothesis 0.

2.1. Methodological weaknesses

Undefined length

The end of the experiment, either in time or in number of received reports, was not defined in advance. This is a weakness, because it would have allowed the experimenters to stop the experiment at a moment when the total results look nice. However, even though it was not in advance, Ken Newton chose a length of six reports while knowing nothing about the results that had come in so far.

Possibilities for participants to manipulate the outcome
As the number of participants wasn't very large, participants could deliberately give wrong or random answers to prevent the results of the whole group from becoming significant.

Due to the difference in group delay (compared to the file without echo, +340.1 µs for the SAA7220-like echo and +997.7 µs for the larger echo), it might have been possible to distinguish between the audio files by measuring the delay from the start of the file to the music using an audio editor, rather than by listening. In retrospect, it would have been better to use wave(n+44) as the centre sample in all three cases and to cut off the first and the last millisecond of each track.

As there is no logical reason why anyone would want to manipulate the outcome, we don't regard these issues as a problem.

Miscommunication about hypothesis 0

As already noted, there was some miscommunication about hypothesis 0. Ken Newton was under the impression that this was the main hypothesis from the very start, but Marcel van de Gevel had failed to put it on the list of hypotheses to be tested and only realized his mistake after the first two reports came in.

The problem with adding hypotheses after data become available is that people often see improbable-seeming patterns in random data, as there are many random patterns that would seem improbable. They can then come up with hypotheses that fit the patterns that they've seen. The methodologically proper solution is to obtain new data and apply the hypotheses to the new data.

Uncontrolled listening environment

As the participants listened at their own home to their own equipment, the experimenters had no control over the listening environment. Then again, an advantage was that the participants did not have to get used to the equipment and room acoustics.

Small sample size

Due to the small number of participants, the test was not very sensitive.

Majority listened via oversampling DACs that probably had pre- and post-echoes of their own

See table 1.

3. Raw results

In the "Correct answer" row of table 1, A means no echo, B means SAA7220-like echo and C means 0.2 dB peak-peak echo. In all other rows, A means most preferred, B means next preferred, and C means least preferred.

	Domenico Scarlatti, K189	John Williams, Raiders of the lost ark	Keith Jarret, Paris-London testament	Camille Saint-Saëns, Danse macabre (Alexander Gibson, Witches' brew)	OS or NOS DAC used?
Correct answer	1C 2B 3A	1A 2C 3B	1C 2A 3B	1B 2C 3A	
Participant #1	1B 2A 3C	1C 2A 3B	1C 2B 3A	1B 2A 3C	OS
Participant #2	1C 2B 3A	1A 2C 3B	1C 2A 3B	1C 2A 3B	OS
Participant #3	1A 2C 3B	1B 2C 3A	1B 2A 3C	1C 2A 3B	OS
Participant #4	1C 2A 3B	1C 2B 3A	1A 2C 3B	1A 2B 3C	OS
Participant #5	1B 2C 3A	1C 2B 3A	1B 2C 3A	1A 2C 3B	Quasi-NOS, sigma-delta with zero order hold interpolation
Participant #6	1B 2C 3A	1C 2A 3B	1C 2A 3B	1C 2B 3A	OS

Table 1: Raw results of the pre- and post-echo listening test

Two participants complained about the quality of the Saint-Saëns recording, saying the differences were much more difficult to hear on that recording than on the others. One of them added that it was best audible on the Keith Jarret and Domenico Scarlatti tracks. In fact this participant had all results correct, except those for Saint-Saëns. Another participant regarded the John Williams track as overprocessed, and only had a clear preference in the cases of the Keith Jarret and Domenico Scarlatti tracks.

4. Hypotheses and statistical analyses

       0. To be tested for each participant, for the whole group and for each piece of music: people can hear the difference between the large echo, small echo and no echo at all and rank them in order of increasing preference.

The probability of guessing large echo, small echo and no echo correctly is 1/6. Binomial probability calculations will tell what the probability is of by chance having at least as many cases where all three are correct as in the test result.

1. To be tested for the whole group: people can hear the difference between the case with the largest echoes and the rest, but do not necessarily know what is what

If this is true, then the case with the largest ripple will either get preference A or preference C, not preference B. When you can't hear any difference, the probability of giving the case with the largest ripple either preference A or preference C is 2/3. Binomial probability calculations will tell what the probability is of by chance having at least as many cases where the largest ripple is given preference A or C as in the test result.

There is no point in testing this per participant, as the result would still not be significant at the 5 % level if all four answers would match the hypothesis (as $(2/3)^4 > 0.05$). Something similar applies to testing per piece of music.

2. To be tested for each participant, for the whole group and for each piece of music: people can hear the difference between the case with the largest echoes and the rest and rate the largest echo as the worst sound

In this case you would only count cases where the largest ripple is rated worst as correct replies. The probability of this happening by chance is 1/3.

3. To be tested for each participant, for the whole group and for each piece of music: people can hear the difference between the case with the small (SAA7220-like) echoes and the others

If this is true, then the case with the small echoes will get preference B. The probability of giving it preference B by chance is 1/3.

	Hypothesis 0	Hypothesis 1	Hypothesis 2	Hypothesis 3
Participant #1	100 %		80.246914 %	40.740741 %
Participant #2	1.62037 %		11.111111 %	11.111111 %
Participant #3	100 %		80.246914 %	100 %
Participant #4	100 %		80.246914 %	80.246914 %
Participant #5	100 %		80.246914 %	100 %
Participant #6	51.774691 %		80.246914 %	40.740741 %
Whole group	58.448667 %	74.616512 %	57.619535 %	57.619535 %
Whole group, only Scarlatti	66.510202 %		64.883402 %	91.22085 %
Whole group, only Williams	66.510202 %		64.883402 %	31.961591 %
Whole group, only Jarret	26.322445 %		31.961591 %	31.961591 %
Whole group, only Saint-Saëns	100 %		91.22085 %	91.22085 %

Table 2: p values for the different hypotheses, that is, probability of getting an equal or better result by pure chance when you don't hear any difference. Results that are significant at the 5 % level in green.

5. Conclusions

The test has produced one significant result, namely the result of participant #2 for hypothesis 0. Either participant #2 was the only one who could hear a clear difference or he made very lucky guesses. It's interesting that he complained about the quality of the one recording that he got wrong. On the other hand, the more hypotheses are tested, the more likely it gets that at least one of them produces a significant result due to pure chance; if the tests were independent, which they are not, one out of twenty would on average produce a significant result due to chance.

Most participants used an OS rather than a NOS DAC for the test, even those who prefer NOS over OS DACs. The reasons varied from the NOS DAC being broken down to being stuck abroad with only an OS DAC. The pre- and post-echoes of the interpolation filters of the OS DACs might have made the difference harder to hear. In fact the same holds for the decimation filters and sample rate converters, if any, used during recording.

All in all, the experiment is inconclusive. Most participants don't seem to hear a clear difference, if any at all, but it would definitely be interesting to do a follow-up experiment with participant #2.

6. References

[1] R. Lagadec and T. G. Stockham, "Dispersive models for A-to-D and D-to-A conversion systems", Audio Engineering Society preprint 2097, presented at the 75th convention, March 1984

[2] See abraxalito's comment in post #999 in the "What do you think makes NOS sound different?"

thread on diyaudio.com, https://www.diyaudio.com/forums/digital-line-level/371931-makes-nos-sound-100.html#post6722475

[3] Philips, SAA7220 datasheet *SAA7220 Digital filter for compact disc digital audio system*, September 1985, page 7, can be found on http://www.datasheetcatalog.com/datasheets_pdf/S/A/A/7/SAA7220.shtml