Audio Engineering Society

Convention Paper

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Cross-Adaptive Polarity Switching Strategies for Optimization of Audio Mixes

Pedro Duarte Pestana[1,3], Joshua D. Reiss[2], and Álvaro Barbosa[1]

[1] *CITAR - Catholic University of Oporto, Rua Diogo Botelho, 1327, 4169-005 Oporto.*

[2] *C4DM - Queen Mary University of London, Mile End Road, London E1 4NS.*

[3] *also member of CEAUL - Universidade de Lisboa, and ILID, Universidade Lusíada de Lisboa*

Correspondence should be addressed to Pedro Duarte Pestana (ppestana@porto.ucp.pt)

ABSTRACT

Crest factor is an often overlooked part of audio production, yet it acts as an important limit to overall loudness. We propose a technique to optimize relative polarities in order to yield the lowest possible peak value. We suggest this is a way of addressing loudness maximization that is more sonically transparent than peak limiting or compression. We also explore additional uses that polarity analysis may have in the context of mixing audio. Results show this is a fairly effective strategy, with average crest factor reductions of 3 dB, resulting in equivalent values for loudness enhancement. While still not comparable to the amount of reduction peak limiters are typically used for, the approach is seen as more transparent via subjective evaluation, through a multi-stimulus test.

1. BACKGROUND CONTEXT

In this work we present a method to achieve loudness enhancement of an audio mixture by using relative polarities in a multi-track context in order to reduce crest factor[1]. To our knowledge, this is a novel perspective, but there are several approaches that address polarity in a multi-track context (but not for loudness optimization) and many works that focus on crest factor minimization, normally to enhance loudness (but do so based on mixed signals, rather than multi-track content). In this section we will briefly review these past approaches.

One should note that the use of polarity during the mix-

[1] It is widely accepted that loudness perception correlates with a signal's root mean square (RMS) value, while peak value has little bearing. This means that a signal with a lower peak-to-RMS level (usually termed crest factor) will be louder than a counterpart with a larger value, if peak values are made equal.

ing phase of an audio production is usually reserved for the corrective purpose of avoiding cancellation between audio tracks that contain strongly correlated signals. The engineer will judge relative polarity on a pair of tracks that represent different spatial instances of the same sound source, and choose the one that will result in the stronger sum. In this corrective perspective, two closely related issues are those of time-alignement and bleed-removal. In these cases it is relative phase and not polarity that is the issue, and instead of an incorrect setting resulting in cancellation, it will result in comb-filtering, a detrimental effect to the sonic quality of a mix.

A number of works have proposed corrective solutions for cancellation and comb-filtering in a multi-track context. [1, 2] focused on both problems, and [3, 4] in time-alignment alone. These works, much like the present article, are rooted in recent interest on intelligent production systems as an aiding device for audio-related tasks such as mixing [5, 6]. However, instead of working from knowledge engineering in order to automatize a process that is typical of manual mixing, we propose to explore an established technique and repurpose it[2].

In this sense, we turn to strategies for loudness enhancement through crest factor reduction. In the realm of single signals (as opposed to multi-track signals), dynamic range compression (DRC) [7] and peak limiting [8] are the traditional approaches used in order to achieve this goal. There is a long standing discussion on the adverse effects of the excessive use of compression and especially peak limiting in the audio industry [9]. In [10] the authors find evidence that this is indeed harmful for the listening experience in a series of subjective tests, and this confirmation provides additional reasons for finding less aggressive strategies for crest factor reduction.

Lynch [11] looks at the problem with an alternative motivation in mind, namely its application in speech in the context of AM and shortwave broadcast, where signal-to-noise ratio is severely compromised. The approach therein is the use of quadratic phase dispersion, but there is no subjective evaluation of audio quality transparency, which will necessarily suffer slightly from the method described. [12] pursues the same goal, driven by a different goal, also grounded on signal-to-noise problems; that of transfer function measurement. They propose the use of nonlinear Chebyshev approximation strategies.

Parker and Välimäki [13] suggest the use of golden ratio all pass filters, and report average reductions in peak amplitude of 2.5 dB for a function that is akin to the one proposed herein. Their suggested approach will cause some degree of transient smearing and the authors offer no subjective evaluation to understand the system's aural transparency[3]. As the proposed strategy needs optimization for short-length signals, the authors suggest segmentation into manageable units, something that is relevant for the real-time solutions proposed in the current article.

Tsilfidis et al. [15] suggest a technique they call Hierarchical Perceptual Mixing, where the perceptually irrelevant spectral components of a mix are discarded prior to summation according to a masking model. The authors present data that shows that this strategy also serves as an approach to lower the crest factor without the artifacts of a squashed dynamic range, even if the result is a consequence of the undertaken approach and not the other way around.

The interest of polarity in mixing is then four-fold:

1. Recent analyses of the loudness wars problem in music mastering [16, 8] brought to the research community's attention the general interest in loudness maximization strategies, which, as stated, have so far been the domain of peak limiting and compression in professional production contexts. Even though there is a strong bias against maximization practices in audio production, supported by recent initiatives in loudness metering [17], these opinions are based on the adverse effects of peak limiting [18], and not on the fact that loudness is, by itself, a problem. In this work we suggest a different, more transparent approach to achieve loudness maximization through the use of polarity.

2. Traditional mixing is a complex task, and engineers will, during the summing process, often hit a clipping point. At this juncture the gain structure may make it difficult to equally lower all signals, and having a way in which to optimize polarity relationships without altering overall balance can prevent the digital overflows. In [19], it is proposed that a

[2] One of the main advantages of automatic systems in musical production is their ability to extend the possibilities of manual operation to a point that would not be feasible for a sound engineer, out of either technical or temporal constraints, and that is what is presented herein.

[3] The ability to perceive slight amount of phase distortion is still a topic that elicits passionate responses - see [14] for an early view - much similar to the problem of absolute polarity we discuss below.

polarity switch on a single track can prevent clipping in this context, a suggestion that sparked our interest in this pursuit.

3. Some tracks within an audio mixture exhibit a polarity problem as a result of strong cross-correlation. When summed, they will cancel out or strongly thin the low-end. This scenario typically arises from acoustical polarity inversions that result from practices such as placing microphones on opposite sides of an acoustic membrane. A system such as the one we propose must take this into consideration and not arbitrarily flip any track in a mixture.

4. Some tracks within an audio mixture exhibit unwanted phase interactions as a result of two microphones picking up copies of the same acoustic source at different points in time.

With all this in mind, we will presently propose a system that automatically tests a polarity-reversing schema that evaluates all possible polarity permutations to achieve optimal sum peak reduction in a mixture, leading to a reduced crest factor and extended headroom, to allow for enhanced loudness. In addition, in order to build a functioning system, we will explore how polarity can harness different goals in the context of automatic mixing.

In the next section we shall introduce concepts that are relevant to this study and justify proposed approaches, and the motivation behind them. In section 3 we will examine the useful case of automatically checking for cancellations between tracks as a result of polarity inversions, and in section 4 extend the idea to bleed detection. Section 5 introduces the core proposal of this text, namely the optimization of track polarities in a multitrack context, and section 6 proposes an extension of the idea to real-time systems. Section 7 provides some objective and subjective evaluation of results, and finally, in section 8 we summarize and propose directions for further research.

2. POLARITY, PHASE, PEAK VALUES AND CREST FACTOR

When several signals in a multi-track mixture are added, the overall peak of the resulting sum is a function of the interaction between all the separate local amplitude values on each of the tracks, and the signal change brought along by a polarity flip on any of the individual tracks will affect the overall result. In a song with M uncorrelated tracks, there will be 2^{M-1} valid polarity permutations[4], each of which will likely result in a maximum peak at a different location in time, and with a different value. If all tracks are uncorrelated, one should expect the resulting RMS to be fairly robust to these inversions. The maximum peak value, however is more prone to serendipity. We performed a quick analysis in order to understand the sort of variations which would result, and found strong indications that the method herein would be interesting, leading to peak reductions of up to around 6 dB, for short segments.

Crest factor is defined as the ratio between the maximum absolute peak of an audio signal and its RMS value. It is usually presented in dB. For sample number n, let $y(n) = x_1(n) + x_2(n) + ... + x_m(n)$, be the audio sum of m individual signals that constitute a mix, and $\Upsilon = \{|y(0)|, |y(1)|, ..., |y(N-1)|\}$, with N the total number of samples. The crest factor is given by:

$$cf(y) = 20 \times \log_{10}\left(\frac{\max(\Upsilon)}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} y^2(n)}}\right) \quad (1)$$

A peak reduction without change in the overall RMS value will provide additional headroom in a mix, and thus perceptual loudness can be increased as a result of a lowered crest factor. Note that concepts are intertwined and while it is easier and cheaper to calculate peak value and increase headroom by lowering it [10], perception of loudness correlates with RMS, and thus only when reducing the crest factor, can we in fact guarantee loudness increase. We have, however, observed that throughout our processing, RMS is kept marginally unchanged, and in this context there is a direct relationship between peak reduction, crest factor minimization and loudness enhancement.

From this point on, it should be noted that the following discussions ignore the idea of *absolute polarity*, otherwise none of the concepts herein could be implemented. Following [20] and [14], [21] runs a test that concludes that the majority of people consider that acoustic sources that are reproduced with a polarity consistent with the

[4] Notice the exclusion of all permutations that would result in all-opposite polarities, and thus a sum that is equal but inverted, necessarily yielding the same crest factor.

original source sound better than those that are not. This means that an acoustic compression (a positive pressure wave) should be represented as a positive voltage, a positive number in the digital media and again a positive pressure wave through the loudspeakers. Suggested practices for the standardization of such a measure are outlined in [22], though the recommendation is targeted towards manufacturers and not practitioners, for the sake of consistency. This is still a controversial topic - see [23, 24, 25].

Absolute polarity is not among the strongest concerns of many top manufacturers [26] and engineers [9], with several sources considering it a myth. For instance, [18] states that *"it is debatable whether the human hearing mechanism can detect absolute polarity. If both speakers are moving inward when they should be moving outwards, can you hear the difference? Many listeners claim to be sensitive to absolute polarity reversals, but scientists have shown that this may only be due to a nonlinearity in the loudspeaker driver or magnet structure"*. There are solid physiological reasons for absolute phase to be a concern in asymmetric waveforms, as neuron spikes seem to respond to a half-rectified version of a signal. But when dealing with contemporary practices in audio production, one almost inevitably opens Pandora's box when defining best practices. Johnsen [21] states that some of the misunderstanding may come from *"some confusion. Recorded polarity comes two ways, but only one is correct for a given system. Much confusion results from freely mixing them. Daily exposed to mixed and incoherent signals, even otherwise careful listeners become accustomed to ignoring polarity"*. For the purpose of the current exploration, we shall accept that ignoring absolute polarity is standard practice in mixing.

A much clearer question regards the *relative* polarity relationships of sources that were captured on the same acoustic environment, especially those that represent sources with acoustic inversions (such as the top and bottom of a membrane or plate). Here, individually recorded tracks must maintain a locked polarity, lest they are cancelled out in the final mix. This situation also introduces the problem of phase interference, which results in potentially severe comb-filtering, when a signal from the same acoustic source is picked up in different microphones at different times [4]. These shall be explored over the next two sections.

3. POLARITY CORRECTION TO PREVENT CANCELLATION

For the 120 multi-track files that were examined in the course of this work[5], we explored the concept of polarity-linked tracks, which are pairs of tracks where in the polarity space $\{[1,1],[1,-1]\}$ there is an option that yields strong cancellation.

We have first explored a blind approach to this problem, as it is simpler than detecting potential sources of conflict and devising more intricate strategies for phase-reversal detection, as done in [1]. It leans on two notions:

1. A sum of two uncorrelated tracks should give rise to an increase in SPL of 3 dB (corresponding to a $\sqrt{2}$-fold gain in amplitude) [27], and this is independent of relative polarities.
2. When a polarity reversal problem happens, it happens for the whole duration of a pair of tracks, and it can be observed at any non-silent point along those tracks.

This would mean choosing the first significant period of time where two arbitrary tracks x_f and x_g from within the track pool had substantial signal level, that is, finding time ν such that:

$$\sqrt{\frac{\sum_{n=\nu}^{\nu+N'} {x_f}^2(n)}{N'}} > k \wedge \sqrt{\frac{\sum_{n=\nu}^{\nu+N'} {x_g}^2(n)}{N'}} > k, \qquad (2)$$

where we use the RMS values to ascertain whether we are above level k for length N' on both tracks. We have found that a sensible choice for N' would be such that it would lead to at least $400\ ms$ and k could fall around 0.05 for a typical track count. This would fulfill point 2 above.

Several approaches will test whether polarity reversal will cause cancelation, following point 1. The simplest one is to check whether the RMS value of the sum is too different from the RMS value of the difference, as

[5] It should be noted that a tight methodology for dealing with multi-track content is hard to implement, as access to multi-track files is still a difficult task. Our files cannot be considered to be either randomly selected, or chosen through a thought-out criteria, and thus the nature of this work remains exploratory.

this will be a sign of a problem. Let a and b be arbitrary track numbers with $a \neq b$, let the sum value be $s(n) = x_a(n) + x_b(n)$, and the difference value be $d(n) = x_a(n) - x_b(n)$. One now wants to check whether:

$$20\log_{10}\sqrt{\frac{\sum_{n=\nu}^{\nu+N'} s^2(n)}{N'}} - 20\log_{10}\sqrt{\frac{\sum_{n=\nu}^{\nu+N'} d^2(n)}{N'}} < 1 \quad (3)$$

and a negative occurrence would give us a flag that would henceforth link a and b's polarity, with a polarity vector that would reflect the relationship (sum or difference) yielding the larger RMS value.

4. POLARITY FOR BLEED DETECTION

Rather than blindly using the simple rule above, one can explore a learning approach such as training a naive Bayes classifier [28] to find class separations. While doing this we extended the class categorization so that it would include hybrid cases where there is no linked-polarities, but there is some degree of correlation, as in cases of bleed.

For the 120 multi-track songs we had available, we had manual tags for all combinations that had been recorded simultaneously in the same acoustic environment in the original track sheets (that is, that would likely present bleed). Let us call the set of excluded track-pairs (those that would not have a cause for bleed) set 'd' of uncorrelated pairs. There were 34534 cases in this set. We then listened through all the correlated cases and tagged those that we considered to be prone to a polarity link (set 'a' with 151 occurrences[6]) and those that were part of a left-right pair (set 'b' with 63 occurrences)[7]. The remaining tracks are those that will exhibit bleed but not a polarity issue, and we shall call them set 'c' with 2026 occurrences.

We then looked at two different features for all pairs, their squared correlation, and the squared difference between sum RMS and difference RMS, as they are both candidates to express the same concept of inter-dependence between tracks. Figure 1 depicts how the two features interact for our dataset, showing clear clusters for the different tags. Note that all instances that were tagged as uncorrelated are squashed together very close to the origin, as expected.

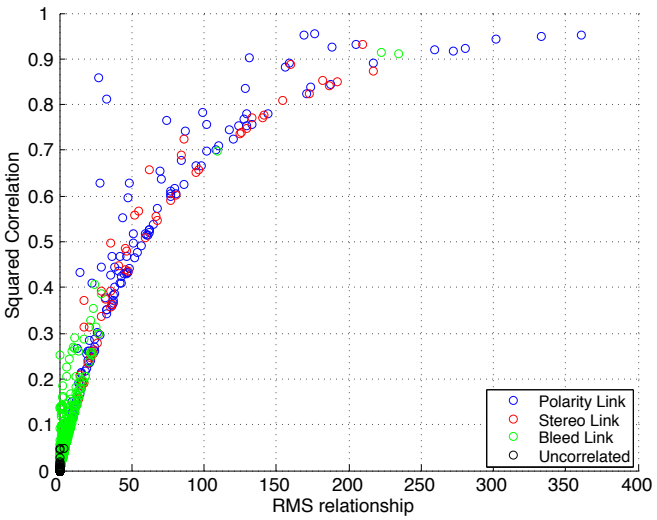

Fig. 1: Suggested crossfade for minimization of digital polarity switching artifacts.

We trained a naive Bayes classifier on the dataset (even though the features are clearly not independent, there are indications that this classifier is still useful [29]). Below is the confusion matrix ordered 'a'->'b'->'c'->'d':

$$\begin{bmatrix} 124 & 2 & 25 & 0 \\ 52 & 9 & 2 & 0 \\ 19 & 0 & 2004 & 3 \\ 0 & 0 & 0 & 34507 \end{bmatrix} \quad (4)$$

The classifier is very successful in predicting uncorrelated tracks, and also seems to separate between bleed issues and polarity issues, if we consider stereo-linkage to be a polarity issue. Precision (p) is a measure of exactness of a classification scheme, relating relevant to irrelevant items (it is the ratio of true positives to the sum of all positives), and recall (r) a measure of completeness (the ratio of true positives to the sum of true positives and false negatives). For our set we see that $p = \{0.72, 0.24, 0.99, 1.0\}$ and $r = \{0.82, 0.14, 0.99, 1.0\}$,

[6] About a fourth of these corresponded to microphones in opposite sides of a membrane, another fourth to tracks that served as comparisons for two microphones, and one-half to microphone-DI Box combinations in guitars and bass guitars.

[7] This is a small set as most of the stereo files in our set were not saved as multi-mono. We decided to tag them separately as their polarity relationship will depend on the stereophonic technique used, and there is room for much more investigation there

so other than very poor recall for left-right links, the approach looks interesting. If we concatenate the polarity cases and the stereo-linkage ones, we arrive at $p = \{0.88, 0.99, 0.99\}$ and $r = \{0.89, 0.99, 1.0\}$. An overall accuracy measure of the f-score, given by:

$$f = \frac{2 \times p \times r}{p + r}, \tag{5}$$

would yield values of $f = \{0.88, 0.99, 1.0\}$ for this three-class situation with sets 'a+b', 'c' and 'd' respectively, which is very promising indeed. This analysis serves only as a pointer to a possible application in polarity and bleed detection, and can, in the future, be studied much more thoroughly.

5. MULTI-TRACK POLARITY OPTIMIZATION

A static optimization of all polarities is fairly easy to achieve, even if it results in a very computationally-intensive process. For the purpose of this section, *optimization* means lowering the crest factor (or lowering the maximum peak value in order to gain headroom, without a change in RMS), as this allows an increase in loudness that is proportional to its decrease. As mentioned, crest factor depends on the ratio of peak-to-RMS, and the latter is rather (but not totally) insensitive to phase or polarity, whereas the former is extremely dependent on it. In a very simple three track case, if we take:

$$\begin{bmatrix} y_1(n) & y_2(n) & y_3(n) & y_4(n) \end{bmatrix} = \begin{bmatrix} x_1(n) & x_2(n) & x_3(n) \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}, \tag{6}$$

finding the best possible solution is just a question of finding the iteration that minimizes the crest factor:

$$k : \frac{\max|y_k|}{RMS_{y_k(n)}} = \min\left(\frac{\max|y_1|}{RMS_{y_1(n)}}, \frac{\max|y_2|}{RMS_{y_2(n)}}, \frac{\max|y_3|}{RMS_{y_3(n)}}, \frac{\max|y_4|}{RMS_{y_4(n)}}\right). \tag{7}$$

Extension to M tracks requires the creation of a permutation matrix built such that:

$$\begin{aligned} &p(m,k) = (-1)^{\left\lfloor \frac{k-1}{2^{(m-1)}} \right\rfloor}, \\ &1 \leq m \leq M, \\ &1 \leq k \leq 2^{M-1}, \end{aligned} \tag{8}$$

where k is the permutation number, m the track number and $\lfloor\ \rfloor$ stands for the floor function, or the integer less than or equal to the evaluated expression. The choice of the correct k is done through the minimization of the peak-to-RMS relationship as:

$$\min_k \left(20\log_{10} \left(\frac{\max_n \left| \sum_{m=1}^{M} p(m,k)\, x_m(n) \right|}{\sqrt{\frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=1}^{M} p(m,k)\, x^2{}_m(n)}} \right) \right). \tag{9}$$

6. REAL-TIME POLARITY CORRECTION AND OPTIMIZATION

For real-time applications we do not have the luxury of knowing how different starting polarity relationships will influence crest factor, but we can look ahead and flip polarities at every point where we are close to clipping, which is very useful as an alternative to revising all gain structure.

Let $p(m,k)$ be the polarity permutation k for track m defined by equation (8), which assumes value 1 for no flip and -1 for flip. Let us also consider that column $k = 1$, that is $p(m,1)$ is the row of no polarity change (a vector of all ones). Given arbitrary coefficients $g_m(n)$ which are the time-varying gains applied by the user to track m, the audio mixture can be written with a polarity flipper in mind as:

$$\begin{aligned} y(m,k) = g_1(n)\, p_1(k)\, x_1(n) + \\ + g_2(n)\, p_2(k)\, x_2(n) + ... + \\ + g_m(n)\, p_m(k)\, x_m(n) \end{aligned} \tag{10}$$

It can be seen that for $k = 1$ all polarity-flipping functions p can be left out and the sum simplifies to a mixture with time-varying gains. We now want to look ahead for potentially clipped values, such that:

$$|y(n + \eta, k)| > 1, \tag{11}$$

where look-ahead lag η is a value reached through heuristics and should be set so that it is over 100 ms according to our data. Let us now imagine that clipping happens at discrete time $n = a$ and we want to consider

a window of size w around the time of the peak. Whenever the condition in equation (11) arises the optimal approach is to evaluate which permutation minimizes:

$$|\max\left(y\left(n,k\right)\right)|_{a-w/2<n<a+w/2-1}. \qquad (12)$$

While optimal, it can be seen that in all likelihood this will lead to a substantial number of simultaneous polarity changes. The sub-optimal but feasible solution is thus to check which single-track polarity flip will yield the best result. That means building a simpler polarity flipping function:

$$p\left(k,l\right) = \begin{cases} -1, & k = l \\ 1, & k \neq l \end{cases}, 1 \leq k,l < M, \qquad (13)$$

which is just a matrix of all-ones with a diagonal of minus ones, so that only one track is flipped for each permutation. If we now consider the sum in equation (10) to incorporate this polarity-flipping function instead of the original, exhaustive one, and again try to minimize the peak within the interval given by equation (12), we arrive at a decision of which track to flip[8].

All this is necessary as it is a complicated task to alter the polarity of one single track in real-time without the addition of sonic artifacts, let alone a greater number of instances. The problem is now how to flip a single instance inaudibly and although we defer a more elegant solution to future explorations, we outline below the heuristic result of our investigations into the best process to achieve this. It can easily be seen (or heard) that the discontinuity brought about by a real-time polarity flip causes an audible glitch[9].

[8] Note that a further simplification can be made: the track that has an amplitude value closer to half the amplitude of the sum of all tracks is the most likely candidate for the polarity flip.

[9] It is curious to notice that though this is a trite point, and more than solved in the analogue domain as the mechanical switching introduces a natural crossfade, it is not solved in the Digital Audio Station world. As an example, we reverse engineered the polarity switch in two different plug-ins by a renowned DAW-manufacturing brand. The polarity switch on the built-in time adjustment tool implements a one sample switch between normal and reversed, while the polarity switch in the built-in EQs will perform a crossfade where the fade out is 256-sample linear before the crossfade point and the fade in 1024-sample inverse exponential. Evidently the latter introduces a much more acceptable artifact than the former.

We have identified three methods to reduce the artifacts associated with changing the polarity half-way through a track[10]:

1. One may perform a cross-fade between a normal-polarity and an inverted-polarity copy of the same signal. We use a fade-in starting approximately $80\ ms$ before peak point and lasting close to $92\ ms$. The ramp-in time follows an exponential onset of type $(q/Q)^4$, where q is the sample number after fade-in start and Q the total number of samples of the fade-in. The ramp-out time is linear starting approximately $11\ ms$ prior to peak point, and lasting $22\ ms$. This is illustrated in figure 2.

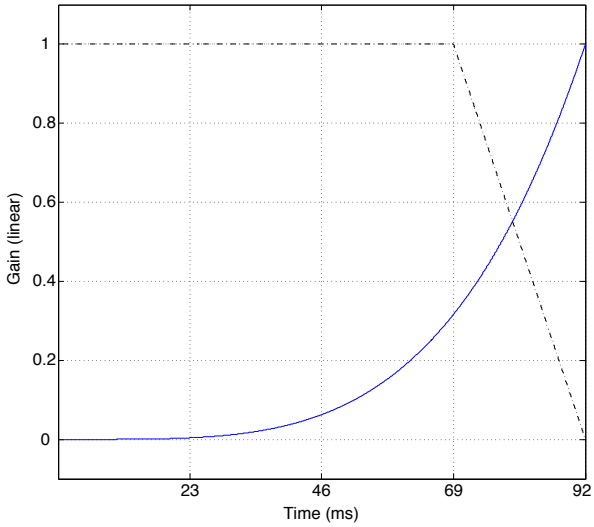

Fig. 2: Suggested crossfade for minimization of digital polarity switching artifacts.

2. One may crossfade into a parallel version that is phase shifted by a constant of $2 \times \pi/3$ over the whole frequency range, which is the lag value for which the sum of two signals has an amplitude equal to any of the signals alone, and given by:

$$y(n) = \mathcal{F}^{-1}\left\{X(k) \times e^{j2\pi/3}\right\}, \qquad (14)$$

with $\mathcal{F}^{-1}$ the Inverse Fourier Transform, and $X(k)$ the complex valued spectral response. Or, in other words, a complex valued all-pass filter of the form:

[10] A zero-crossing flip would be a first choice, but unless other conditions are met, it still causes a glitch.

$$H(z) = e^{j \cdot 2 \cdot \frac{\pi}{3}} \quad (15)$$

This approach has the advantage of allowing for smoother crossfades, but two disadvantages: it is harder to calculate, and the sound quality of the parallel copy is compromised in comparison to the simple version. It also requires a new evaluation stage as we are no longer minimizing the peak after a polarity inversion, but after a constant phase-shift.

3. By introducing a function that looks at the threshold of masking (see [9] for details), it is possible to optimize the polarity flip point by hiding it where it can be masked by the other elements of the audio mixture. This also brings about new problems, namely we have found it quite likely that during a fast read-ahead time there is no point where the intended signal is masked (particularly because a masking function operates on windows that are typically much longer than what would be desirable for real-time). A possible alternative that we have explored is to choose the track to flip as a function of it being masked, but this yields results that are sub-optimal.

The success of these approaches is highly dependent on the number of digital overflows in a given situation and the ability to solve them with a polarity flip.

7. EVALUATION

The success of the proposed approach depends on two factors: whether high crest factor reductions can be achieved, and whether the polarity permutations are sonically of perceptual similar quality. Figure 3 shows the range of crest factors obtainable on a subset of 26 songs, when twenty second long clips are used, and all polarity-related tracks are pre-linked, as described in Section 3. Notice that even in the cases where the range is small, it is still over the 1 dB just noticeable difference.

In total, we used 120 multi-track files for analysis. An equal-loudness mix of all tracks would yield on average a crest factor of roughly 21 dB. The optimal polarity-relationship would on average result on a crest factor of 18 dB. The most extreme case we found yielded a reduction in excess of 6 dB. This situation improves for shorter segments as can be seen in Figure 4.

Because the real-time approach is currently a compromise solution, however, the fact that it runs on shorter segments does not make it more effective, and the average crest factor after optimization is roughly 19.4 dB. The number of tracks also has a bearing on the situation, with fewer tracks marginally allowing for better results, as seen in Figure 5.

For the assertion of sonic quality, we have performed a multi-stimulus subjective evaluation (similar to the pop-

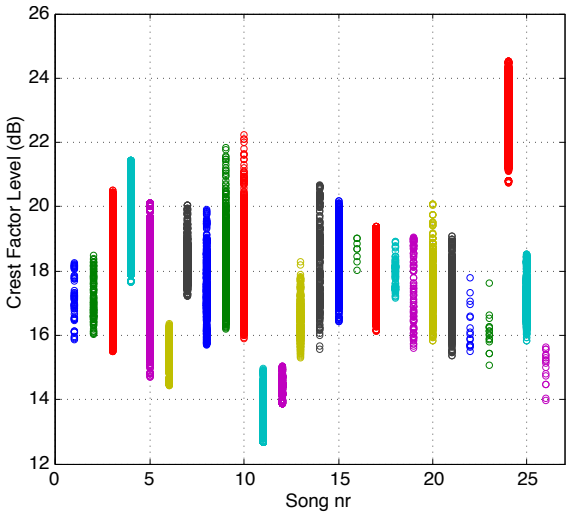

Fig. 3: Crest factor range for a dataset of 26 songs mixes with no compression applied individually on any track. Dot density depends on track count and thus number of permutations.

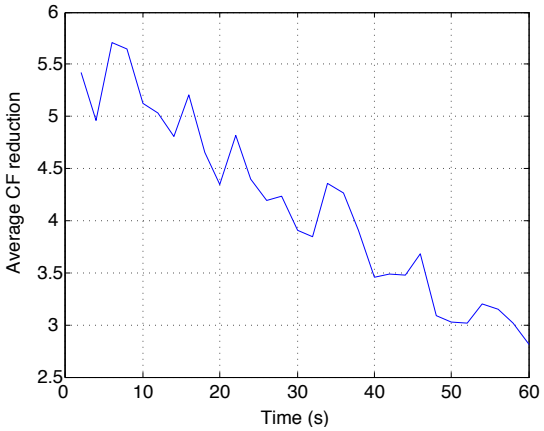

Fig. 4: Average crest factor reduction for randomly selected time segments within different songs in a 120-song pool.

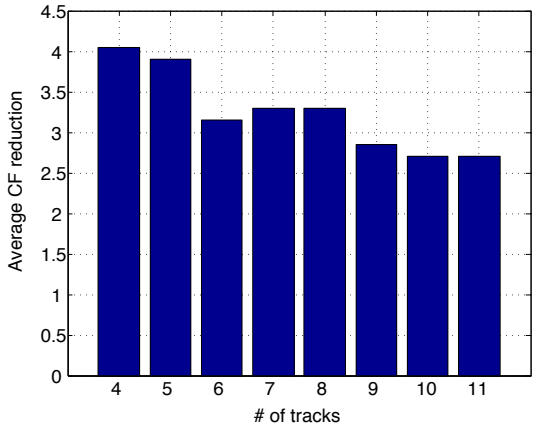

Fig. 5: Average crest factor reduction as a function of number of tracks, for randomly selected 20-second segments.

ular MUSHRA tests [30] in terms of interface, but without reference or anchor) to ascertain differences in polarity preference on the four songs that yielded largest original-to-optimized crest factor differences. Five conditions were tested:

- Condition α: simple balance of unprocessed tracks, done by a professional sound engineer. Relative track loudness balance would be kept as best as possible during the next conditions, lest it became a confounding factor.
- Condition β: crest factor reduction by choosing the best permutation described above. Peak level matching with condition α, which results in dramatically increased loudness (by about $4-6\,dB$, as seen above). This condition was inserted with the expectation that it would be rated highest, as test subjects are known to prefer louder material [18, 31].
- Condition γ: similar to condition β but with overall loudness that matches that of condition α. This tests for the hypothesis of absolute polarity being perceptually relevant in a complex mix.
- Conditions δ and ϵ are situations where peak limiting was used in order to achieve crest factor reductions similar to β in the case of δ and double that of β in the case of ϵ. These are loudness matched to α.

Twenty four subjects experienced in audio engineering collaborated. Tests were performed through headphones with a controlled signal path, at an $83\,dB$ listening level measured with a dummy head. Subjects listened to four songs with the same condition set and rated for perceived audio quality. Overall results are presented in Figure 6, where the complete equivalence of mean and confidence intervals between α and γ indicates absolute lack of preference in terms of polarity relationships. As expected, enhancing loudness is clearly preferred. While this does not prove that arbitrary polarity flipping is indistinguishable from normal polarity in a multi-track context, it does show that in terms of subjective user preference there is no clear favorite, and any concern is outweighed by the large benefit of gaining headroom in order to achieve a loudness boost. The peak limiting cases are judged to be subtly worse (and worsening with increase in peak limiting amount), even if there seems to be less agreement, as confidence intervals are much higher[11].

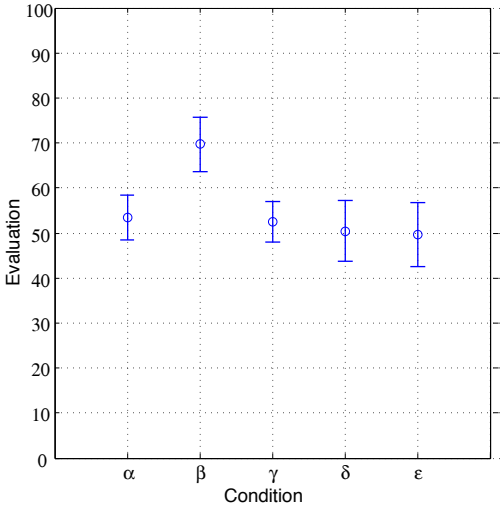

Fig. 6: Mean and confidence intervals for the evaluation of each condition, considering inter-song differences to be irrelevant.

[11] We analyzed this spreading of data by clustering user ratings and found that the larger variance can be explained by a polarization of preference into two clusters, whose main factor is subject age. It appears that younger subjects are more prone to classifying peak limiting as being of better quality, whereas older subject take exception to the sonic results of this practice. This can be a sign of adaptation to a characteristic that has become a standard over the last couple of decades.

8. CONCLUSIONS AND FURTHER WORK

It has been seen that arbitrary polarity reversals can lower the crest factor level by more than $6\ dB$ in offline implementations and $3\ dB$ in real-time, with modest track counts. At a point in time where production techniques routinely demand crest factor reductions of up to $12\ dB$ or even more in extreme cases, it is interesting to know that half of that value can be achieved with '*transparent*' techniques (causing less change in perceptual sound quality), allowing for a reduction in compression and peak limiting amount. It would be interesting to investigate the perceptual effects of compression after this phase of crest-factor optimization, to see whether a two-fold approach will reap benefits.

All our tests were performed with monophonic material, and an extension to stereophonically-recorded tracks is useful in the future. As far as the static version goes, the concept is extremely straightforward and the only sophistication to pursue is a computational optimization for the long calculation process[12]. For real-time systems, the integration with a proper loudness balancing mechanism is the logical next step, and optimization of flip points with more sophisticated measures of masking can help with artifact reduction. Automatically segmenting long signals into shorter streams can also be an advantageous strategy to explore.

The learning approach in Section 5 is clearly still only proof-of-concept. The naive Bayes classifier was used for its simplicity, and there might be more accurate algorithms for the task. Feature selection was also not fully considered and the features used were too dependent on one-another. The problem of class imbalance was also not addressed, so there are many paths to explore along a direction that was left as a pointer here, as it stands as a peripheral side-effect of our main concept.

There is still much to be done on the analysis side, as it was not explained why the performance of this strategy works better on some songs than others, or how phase-shifting the signal to flip to in real time affects performance and quality. The non-random selection of multitracks can be a cause for bias in results and interpretation, and we would hope to be able to extend our analysis with access to larger (and especially more diverse and publicly-available) collections. Finally, it would be interesting to ascertain whether in this context the strategy is truly transparent through a large scale ABX-type test.

[12] Though the proposal herein could be very useful in a D.A.W. environment, if performed as a background, offline task that the user could switch on or off.

9. REFERENCES

[1] E. P. Gonzalez and J. D. Reiss, "Determination and Correction of Individual Channel Time Offsets for Signals Involved in an Audio Mixture," in *Proceedings of the 125th AES Convention*. San Francisco: Audio Engineering Society, 2008.

[2] N. Jillings, A. Clifford, and J. D. Reiss, "Performance optimization of GCC-PHAT for delay and polarity correction under real world conditions," in *Proceedings of the 134th AES Convention*, 2013.

[3] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY: IEEE, 2009.

[4] A. Clifford and J. D. Reiss, "Reducing Comb Filtering On Different Musical Instruments Using Time Delay Estimation," *Journal on the Art of Record Production*, vol. 5, 2011.

[5] J. D. Reiss, "Intelligent Systems for Mixing Multichannel Audio," in *17th Intl Conf on Digital Signal Processing (DSP)*, 2011.

[6] P. D. Pestana and J. D. Reiss, "Intelligent Audio Production Strategies Informed by Best Practices," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Jan. 2014.

[7] B. A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *Audio and Electroacoustics, IEEE Transactions on*, vol. 17, no. 1, pp. 22–32, 1969.

[8] E. Vickers, "The Loudness War: Background, Speculation and Recommendations," in *Proceedings of the 129th AES Convention*. San Francisco: Audio Engineering Society, 2010.

[9] P. D. Pestana, "Automatic Mixing Systems Using Adaptive Digital Audio Effects," PhD, Universidade Católica Portuguesa, 2013.

[10] M. Wendl and H. Lee, "The Effect of Dynamic Range Compression on Loudness and Quality Perception in Relation to Crest Factor," in *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[11] J. T. Lynch, "Reduction of peak/rms ratio of speech by amplitude compression and quadratic phase dispersion," *Journal of the Audio Engineering Society*, vol. 36, no. 3, pp. 147–152, 1988.

[12] P. Guillaume, J. Schoukens, R. Pintelon, and I. Kollar, "Crest-factor minimization using nonlinear Chebyshev approximation methods," *Instrumentation and Measurement, IEEE Transactions on*, vol. 40, no. 6, pp. 982–989, 1991.

[13] J. Parker and V. Valimaki, "Linear dynamic range reduction of musical audio using an allpass filter chain," *Signal Processing Letters, IEEE*, vol. 20, no. 7, pp. 669–672, 2013.

[14] D. Stodolsky, "The Standardization of Monaural Phase," *IEEE Transactions on Audio and Electroacoustics*, vol. 18, no. 3, pp. 288–299, 1970.

[15] A. Tsilfidis, C. Papadakos, and J. Mourjopoulos, "Hierarchical Perceptual Mixing," in *Proceedings of the 126th AES Convention*. New York: Audio Engineering Society, 2009.

[16] AES, "Loudness Trumps Everything," *Journal of the Audio Engineering Society*, vol. 54, no. 5, pp. 421–423, 2006.

[17] EBU, "Tech Doc 3343 "Practical Guidelines for Production and Implementation in Accordance with EBU R 128"," European Broadcast Union, Geneva, Tech. Rep., 2011.

[18] B. Katz, *Mastering Audio — the Art and the Science*. Oxford: Focal Press, 2007.

[19] M. P. Stavrou, *Mixing with your Mind*, 1st ed. Mosman, Australia: Flux Research, 2003.

[20] W. A. Rosenblith and M. R. Rosenzweig, "Electrical Responses to Acoustic Clicks: Influence of Electrode Location in Cats," *Journal of the Acoustical Society of America*, vol. 23, no. 5, pp. 583–588, 1951.

[21] C. Johnsen, "Proofs of an Absolute Polarity," in *Proceedings of the 91st AES Convention*. New York: Audio Engineering Society, 1991.

[22] AES, "AES26-2001 (r2006): Recommended Practice for Professional Audio — Conservation of the Polarity of Audio Signals," Audio Engineering Society, New York, Tech. Rep., 2006.

[23] J. Vanderkooy and S. P. Lipshitz, "Polarity and Phase Standards for Analog Tape Recorders," in *Audio Engineering Society Convention 69*, 1981.

[24] R. A. Greiner and D. E. Melton, "Observations on the audibility of acoustic polarity," *Journal of the Audio Engineering Society*, vol. 42, no. 4, pp. 245–253, 1994.

[25] S. Sakaguchi, T. Arai, and Y. Murahara, "The effect of polarity inversion of speech on human perception and data hiding as an application," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2000, pp. II917—-II920.

[26] D. Self, *Small Signal Audio Design*. Oxford: Focal Press, 2010.

[27] D. M. Howard and J. Angus, *Acoustics and Psychoacoustics*, ser. Focal Press. Oxford: Elsevier Science & Technology, 2009.

[28] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[29] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[30] ITU, "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunications Union, Geneva, Tech. Rep., 2003.

[31] S. Bech and N. Zacharov, *Perceptual Audio Evaluation — Theory, Method and Application*. Chichester: John Wiley & Sons, 2006.