Automatic Minimisation of Masking in Multitrack Audio using Subgroups

David Ronan, Zheng Ma, Paul Mc Namara, Hatice Gunes, and Joshua D. Reiss,

Abstract—The iterative process of masking minimisation when mixing multitrack audio is a challenging optimisation problem, in part due to the complexity and non-linearity of auditory perception. In this article, we first propose a multitrack masking metric inspired by the MPEG psychoacoustic model. We investigate different audio processing techniques to manipulate the frequency and dynamic characteristics of the signal in order to reduce masking based on the proposed metric. We also investigate whether or not automatically mixing using subgrouping is beneficial or not to perceived quality and clarity of a mix. Evaluation results suggest that our proposed masking metric when utilised in an automatic mixing framework reduces inter-channel auditory masking as well as improves the perceived quality and perceived clarity of a mix. Furthermore, our results suggest that using subgrouping in an automatic mixing framework can also improve the perceived quality and perceived clarity of a mix.

Index Terms—Auditory Masking; Multitrack Mixing; MPEG; Equalization; Dynamic Range Processing; Subgrouping; Numerical Optimisation; Perceived Emotion

---

1 INTRODUCTION

MASKING is a perceptual property of the human auditory system that occurs whenever the presence of a strong audio signal makes the temporal or spectral neighbourhood of weaker audio signals imperceptible [1], [2]. Frequency masking may occur when two or more stimuli are simultaneously presented to the auditory system. The relative shapes of the masker's and maskee's magnitude spectra determine to what extent the presence of certain spectral energy will mask the presence of other spectral energy.

Temporal masking is the characteristic of the auditory system where sounds are hidden due to a masking signal occurring before (pre-masking) or after (post-masking) a masked signal. The effectiveness of temporal masking attenuates exponentially from the onset and offset of the masker [3].

A simplified explanation of masking phenomena is when a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane. An excitation pattern is a neural representation of the pattern of resonance on the basilar membrane, caused by a given sound [4]. The area around the characteristic frequency (referred to as the frequency bandwidth of the "overlapping bandpass filter" created by the cochlea) of the masker's signal location effectively blocks the detection of weaker signals [3]. Examples of frequency and temporal masking are shown in Figure 1 and Figure 2 respectively.

Mixing is a process in which multitrack material whether recorded, sampled or synthesised is balanced,
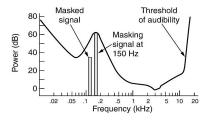

Fig. 1. Frequency masking example of a 150 Hz tone signal masking an adjacent frequency tone by increasing the threshold of audibility around 150 Hz.
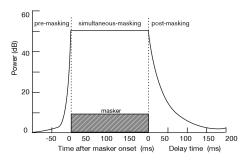

Fig. 2. Schematic drawing to illustrate and characterise the regions within which pre-masking, simultaneous masking and post masking occur. Note that post-masking uses a different time origin than pre-masking and simultaneous masking.[3]

treated and combined into an output format that is most commonly two channel stereo [5].

In the process of mixing, sound sources inevitably mask one another, which reduces the ability to fully hear and distinguish each sound source. Partial masking occurs whenever the audibility of a sound is degraded due to the presence of other content, but the sound may still be perceived.

- D. Ronan is with the Centre for Intelligent Sensing, Queen Mary University of London, UK.
  E-mail: d.m.ronan@qmul.ac.uk
- H. Gunes is with the Computer Laboratory, University of Cambridge, UK.
  E-mail: hatice.gunes@cl.cam.ac.uk
- J.D. Reiss is with the Centre for Digital Music, Queen Mary University of London, UK.
  E-mail: joshua.reiss@qmul.ac.uk

It is often partial masking that occurs within a mix. The mix can sound poorly produced or underwhelming, and have a lack of clarity as a result [6].

Masking reduction in a mix involves a trial and error adjustment of the relative levels, spatial positioning, frequency and dynamic characteristics of each of the individual audio tracks. In practice, the masking reduction process embodies an iterative search process similar to that of numerical optimisation theory [7], [8]. Masking reduction therefore can be thought of as an optimisation problem, which provides some insight to the methodology of automatic mixing in order to reduce masking. Given a certain set of controls for a multitrack, the final mix output can be thought of as the optimal solution to a system of equations that describe the masking relationship between the audio tracks in a multitrack recording.

Frequency processing, dynamics processing and subgrouping are the three main aspects of our masking minimisation investigation. Equalisation can effectively reduce masking by manipulating the spectral contour of different instruments so that there is less frequency domain interference between each audio track. Dynamic range processing is a nonlinear audio effect that can alter the dynamic contour of a signal in order to reduce masking over time. The classic operations of dynamics processing and equalisation control are two separate domains of an audio signal. The combined use of both filtering and dynamics processing implies a larger control space, and can reduce masking much more precisely and effectively in both frequency and time aspects than using either processor alone [5], [9]. Subgrouping allows us to localise the application of the frequency and dynamics processing to specific instrument types that would typically share similar timbre, dynamic range and spectral content.

The two principle aspects of automating a masking reduction process are the creation of a model of masking in multitrack audio that correlates well with human perception, and the development of audio techniques and algorithms to reduce masking without causing unpleasant audio artefacts.

In this article we present a novel intelligent mixing system which uses a psychoacoustic model, numerical optimisation technique and the use of subgroups. Based on this, we propose a novel masking metric for use with multitrack audio. Selected control parameters of equalisation and dynamic range compression effects are then optimised iteratively using the Particle Swarm algorithm [10], toward a desired mix described by the masking metric. We test the hypothesis of whether or not using subgroups is beneficial or not to automatic mixing systems. We also test if subgrouping can have an impact on the perceived emotion in a recording. A formal subjective evaluation in the form of a listening experiment was conducted to assess the system performance against mixes produced by humans.

The structure of this paper is summarised as follows. In Section 2 we discuss the background of masking metrics, subgrouping and measuring emotional response to music. Section 3 describes the methodology of how we formed an automatic multitrack masking minimisation system and how we conducted the subsequent listening test. In section 4 performance evaluations are presented and finally in sec-tion 5 we discuss the most interesting aspects of the research and outline future directions.

2 BACKGROUND

Perceptual models capable of predicting masking behaviour have received much attention over the years, particularly in fields such as audio coding [11]–[15], where the masked threshold of a signal is approximated to inform a bit-allocation algorithm. [16] proposes a method for adjusting the masking threshold in audio coding to make the decoded signal robust to quantisation noise unmasking. Masking models are also often used in image and audio watermarking [17], [18]. Similar models are used in distortion measurement [19] and sound quality assessment [20]–[22], where nonlinear time-domain filter banks are used to allow for excitation pattern calculation whilst maintaining good temporal resolution. Another simple masking model is used in [23] to remove perceptually irrelevant time-frequency components. More advanced signal processing masking models that lie closer to physiology include a single-band model that accounts for a number of frequency and temporal masking experiments [24]. A 'modulation filter bank' was subsequently added to analyse the temporal envelope at the output of a gammatone filter whose output is half-rectified and low pass filtered at 1kHz, simulating the frequency to place transform across the basilar membrane, and receptor potentials of the inner hair cells [25]. Building upon the proposed modulation filter bank, a masking model called the Computational Auditory Signal-Processing and Perception (CASP) model was presented that accounts for various aspects of masking and modulation detection [26].

However, all mentioned models only output masked threshold as a measurement of masking, and only considered the situation when a signal (usually a test-tone signal) was fully masked. [27] explored partial loudness of mobile telephone ring tones in a variety of everyday background sounds e.g. traffic, based on the psychoacoustic loudness models proposed in [28], [29]. By comparing the excitation patterns (computed based on [28], [29]) between maskee and masker, [30] introduced a quantitative measure of masking in multitrack recording. Similarly, a Masked-to-Unmasked Ratio which related the original loudness of an instrument to its loudness in the mix was proposed in [31].

Previous attempts to perform masking reduction in audio mixing include [32]–[35]. [32] aimed to achieve equal average perceptual loudness on all frequencies amongst all multi-track channels, based on the assumption that the individual tracks and overall mix should have equal loudness across frequency bands. However, this assumption may not be valid, and their approach does not directly address spectral masking. [33] designed a simplified measure of masking based on best practices in sound engineering and introduced an automatic multitrack equalisation system. However the simple masking measure in [33] might not correlate well with the perception of human hearing, as is evident in the evaluation. [34] applied a partial loudness model and [27] adjusts the levels of tracks within a multitrack in order to counteract masking. Similar techniques were investigated through an optimisation framework in [35]. However both [34] and [35] only performed basic level

adjustment to tackle masking, which may have additional detrimental effects on the relative balance of sources in the mix [9].

2.1 Masking Metrics

There are a number of different multitrack masking metrics available that can be combined to perform a cross-analysis on multitracks. We can quantify the amount of masking by investigating the interaction between the excitation patterns of a maskee and a masker, where the maskee is an individual track and the masker is the combination of all the other tracks in a multitrack. This is done utilising the cross-adaptive architecture proposed in [36], [37]. All the masking metrics we discuss make use of this cross adaptive architecture. However, the first two masking metrics we will discuss are based on the perceptual loudness work of Moore [38], [39] and the final masking metric we discuss is based on spectral magnitude.

The procedure to derive loudness and partial loudness of each track in a multitrack is summarised as follows [34]. A multitrack consists of N sources that have been pre-recorded onto N tracks. Track n therefore contains the audio signal from source n, given by s_n. The transformation of s_n through the outer and middle ear to the inner ear (cochlea) is simulated by a fixed linear filter. A multi-resolution Short Time Fourier Transform (STFT), comprising 6 parallel FFTs, performs the spectral analysis of the input signal. Each spectral frame is filtered by a bank of level-dependent roex filters whose centre frequencies range from 50Hz to 15kHz. Such spectral filtering represents the displacement distribution and tuning characteristics across the human basilar membrane.
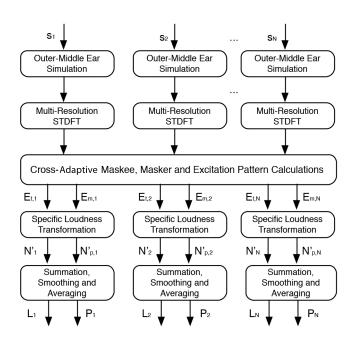

Fig. 3. Flowchart of multitrack loudness model for N input signals.

The excitation pattern E is calculated as the output of the auditory filters as a function of the centre frequency spaced at 0.25 ERB intervals. Equivalent rectangular bandwidth (ERB) gives a measure of auditory filter width. The mapping between frequency, f (Hz), and ERB (Hz) is shown in Equation 1.

$$ERB = 24.7(0.0437f + 1) \tag{1}$$

To account for masking, two excitation patterns, the target track (maskee) $E_{t,n}$ and the masker $E_{m,n}$, with respect to s_n are calculated as described in [28], [29]. The masker here is the supplementary sum of the accompanying tracks related to the target track, as given by [31]

$$s'(n) = \sum_{i=1,i\neq 1}^{N} s_i \tag{2}$$

For a sound heard in isolation, the intensity represented in the excitation pattern is converted into specific loudness N'_n, which represents the loudness at the output of each auditory filter. In a partial masking scenario with concurrent masker $E_{m,n}$, partial specific loudness $N'_{p,n}$ is calculated. The detailed mathematical transformations to obtain specific and partial specific loudness can be found in [28].

The summation of N'_n, and $N'_{p,n}$ across the whole ERB scale produces the total unmasked and masked instantaneous loudness. All instantaneous loudness frames are smoothed to reflect the time-response of the auditory system, as described in [29], and then averaged into scalar perceptual loudness measures, loudness L_n and partial loudness P_n. This is illustrated in Figure 3

Adapting the method of Vega et al [30], the masking measurement M_n can be defined as the masker-to-signal ratio (MSR) based on an excitation pattern integrated across ERB scale and time. This is given by

$$M(n) = \text{MSR}(n) = 10 \log_{10} \frac{\sum_{\text{ERB}} E_{m,n}}{\sum_{\text{ERB}} E_{t,n}} \tag{3}$$

Wichern et al. [40] used a model based on loudness loss to measure masking,

$$L_{loss} = L_{phon} - PL_{phon} \tag{4}$$

where L_{phon} is the loudness of the maskee in isolation and PL_{phon} is the partial loudness of the maskee when masked by the rest of the mix. The loudness unit here is phon as opposed to sones, which was used in Moore's original loudness model we discussed initially. The authors subsequently use a gating procedure to only measure masking when an instrument is actively playing.

In the work by Sina et al. [33], the authors do not use an auditory model to measure masking. They based their measurement on spectral magnitude. Where the amount of masking that track A (masker) at frequency f and time t causes on track B (maskee) at the same frequency and time is given by

$$M_{A,B}(f,t) = \begin{cases} X_A(f,t)X_B(f,t) & \text{if} \\ & R_B(f,t) \leq R_T < R_A(f,t) \\ 0 & \text{else} \end{cases} \tag{5}$$

where $X_N(f,t)$ and $R_N(f,t)$ are respectively the magnitude in decibels and the rank of frequency f, at time t for track N. R_T is the maximum rank for a frequency region to be considered essential.

2.2 Subgrouping

At the early stages of the mixing and editing process of a multitrack mix, the mix engineer will typically group instrument tracks into subgroups [5]. An example of this would be grouping guitar tracks with other guitar tracks or vocal tracks with other vocal tracks. Subgrouping can speed up the mix workflow by allowing the mix engineer to manipulate a number of tracks at once, for instance by changing the level of all drums with one fader movement, instead of changing the level of each drum track individually [5]. Note that this can also be achieved by a Voltage Controlled Amplifier (VCA) group - a concept similar to a subgroup where a specified set of faders are moved in unison by one 'master fader', without first summing each of these channels into one bus. However, subgrouping also allows for processing that cannot be achieved by manipulation of individual tracks. When nonlinear processing such as dynamic range compression or equalisation is applied to a subgroup, the processor will affect the sum of the sources differently than when it would be applied to every track individually. An example of typical subgrouping setup can be seen in Figure 4.
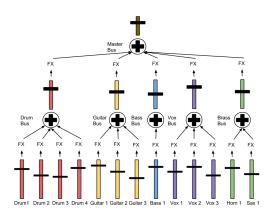

Fig. 4. Typical subgrouping setup.

Very little is known about how mix engineers choose to apply audio processing techniques to a mix, but there have been few studies looking at this problem [41], [42]. Subgrouping was touched on briefly in [41] when the authors tested the assumption *"Gentle bus/mix compression helps blend things better"* and found this to be true, but it did not give much insight into how subgrouping is generally used. In [43], the authors explored the potential of a hierarchical approach to multitrack mixing using instrument class as a guide to processing techniques. However, providing a deeper understanding of subgrouping was not the aim of the paper. Subgrouping was also used in [44], but similarly to [43] this was only applied to drums and no other instrument types were explored.

Although subgrouping is not well documented, it is used extensively in all areas of audio engineering and production. We have in previous work investigated how subgrouping should be implemented when mixing audio [45], [46]. We have utilised these recommendations during the course of this study.

2.3 Measuring Emotional Responses to Music

There are a number of different methods for measuring emotional responses to music. Self-report is one of three methods often used when measuring emotional responses to music, the other two being physiological measurements and facial expression analysis. Perhaps the most common self-report method is to ask listeners to rate the extent to which they perceive or feel a particular emotion, such as happiness. Techniques to assess affect are using a Likert scale or choosing a visual representation of the emotion they are feeling. An example visual representation is the Self-Assessment Manikin [47] where the user is asked to rate the scales of arousal, valence and dominance based on an illustrative picture.

Another method is to present listeners with a list of possible emotions and ask them to indicate which one (or ones) they hear. Examples of this are the Differential Emotion Scale and the Positive and Negative Affect Schedule (PANAS). In PANAS, participants are requested to rate 60 words that characterize their emotion or feeling. The Differential Emotion Scale contains 30 words, 3 for each of the 10 emotions. These would be examples of the categorical approach mentioned previously [48], [49].

A third approach is to require participants to rate pieces on a number of dimensions. These are often arousal and valence, but can include a third dimension such as power, tension or dominance [50], [51].

The methods presented above constitute different types of self-report, which may lead to concerns about the validity of results due to response bias. Fortunately, people tend to be attuned to how they are feeling (i.e., to the subjective component of their emotional responses) [52]. Furthermore, Gabrielsson came to the conclusion that self-reports are "the best and most natural method to study emotional responses to music" after conducting a review of empirical studies of emotion perception [53]. However, one caveat with retrospective self-report is 'duration neglect' [54], where the listener may forget the momentary point of intensity of the emotion attempted to be measured.

We have chosen to use self-report as the measure of perceived emotion (Arousal-Valence-Tension) in our experiment due to it being the most reliable measure according to Gabrielsson [53].

3 METHODOLOGY

3.1 Research Questions and Hypotheses

The main hypothesis we aim to test is *can our proposed automatic mixing system be used to reduce the amount of auditory masking that occurs in a multitrack mix and subsequently improve its perceived quality*. We also tested two further hypotheses, *can using subgroups when generating an automatic mix improve the perceived quality and clarity of a mix* and *can the use of subgroups in an automatic mixing system have an impact on the perceived emotions of the listener over automatic mixes that do not use subgroups*. These hypotheses were evaluated through examination of the objective performance and subjective listening tests.

3.2 Automatic Mixing System

There were two types of automatic mixes generated for this experiment, one which made use of subgrouping and one which did not. The mix process is illustrated in Figure 5.
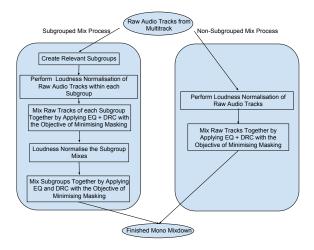

Fig. 5. Automatic mixing process.

3.3 Audio Processing and Control Parameters

3.3.1 Subgrouping

In the multitrack of each song we used for the experiment, we created subgroups based on typically grouped instrumentation such as vocals, drums and guitars etc. This is similar to the approach used in [55]. This allowed us to use the optimisation mixing technique presented here to create a number of sub-mixes and then create a final mix by mixing each of the submixes together. This essentially gave us a multi-layer optimisation framework. When subgrouping was not used in an automatic mix, the optimisation mixing technique was applied to all the audio tracks at once.

3.3.2 Loudness Normalisation

Before we applied the optimisation mixing technique we employed loudness normalisation on each audio track in each multitrack. We performed loudness normalisation on all of the audio tracks using the ITU-R BS. 1770-2 specification [56]. Each audio track was loudness normalised to -24 LUFS except in the case of a lead vocal, where it was loudness normalised to -18 LUFS. We made the lead vocal louder than everything else as it is usually the most important audio track within a mix [57]. Once a subgroup had been mixed, it was also loudness normalised to -24 LUFS except in the case of vocal subgroups, which would be set to -18 LUFS.

3.3.3 Equalisation

We designed a six-band equaliser to be applied in the optimisation process. Six different cascaded second-order IIR filters were designed to cover the typical frequency range used when mixing. The filter specification is shown in Table 1

The gains of the six-band equaliser filter for each track are selected as the control parameters to be obtained

TABLE 1
Six band equaliser filter design specifications

Band No.	Centre Frequency (Hz)	Q-Factor
1	75	1
2	100	0.6
3	250	0.3
4	750	0.3
5	2500	0.2
6	7500	1

through the optimisation procedure. The control parameters in the equalisation cases are given by

$$\boldsymbol{x} = [\boldsymbol{g}_1 \quad \boldsymbol{g}_2 \quad \cdots \quad \boldsymbol{g}_n], \tag{6}$$

in which for each $\boldsymbol{g}_i$ (vector-valued)

$$\boldsymbol{g}_i = [g_{1i} \quad g_{2i} \quad \cdots \quad g_{6i}], \tag{7}$$

contains the six gain controls for each track.

3.3.4 Dynamic Range Compression

The digital compressor model employed in our approach was a feed-forward compressor with smoothed branching peak detector [58]. A typical set of parameters of a dynamic range compressor includes the Threshold, Ratio, Attack and Release Times, and Make-up gain. In the case of adjusting the dynamic of the signal to reduce masking through optimisation, the values of threshold (T), ratio (R), attack (a) and release (r) are control parameters to be optimised. Since dynamics are our main focus here rather than the level, the make-up gain of each track is set to compensate the loudness differences (measured by EBU loudness standard [56]) before and after dynamic processing. The make-up gain for each track is given by

$$g_{\Delta i} = L_{EBUi} - L'_{EBUi}, \tag{8}$$

where L_{EBUi} and L'_{EBUi} represent the measured loudness before and after the dynamic range compression respectively. The control parameters in the dynamic case are given by

$$\boldsymbol{x} = [\boldsymbol{d}_1 \quad \boldsymbol{d}_2 \quad \cdots \quad \boldsymbol{d}_n] \tag{9}$$

Similarly, every $\boldsymbol{d}_i$ is constituted of four standard DRC control parameters denoted as, threshold (T_i), ratio (R_i) attack (a_i), release (r_i).

$$\boldsymbol{d}_i = [T_i \quad R_i \quad a_i \quad r_i] \tag{10}$$

3.3.5 Control Parameters

The notation of the final control parameters to be optimised in the multitrack masking minimisation process is given by

$$\boldsymbol{x} = [\boldsymbol{c}_1 \quad \boldsymbol{c}_2 \quad \cdots \quad \boldsymbol{c}_n], \tag{11}$$

In this case, for each $\boldsymbol{c}_i$

$$c_i = \begin{pmatrix} g_{1,i} & \cdots & g_{6,i} & T_i & R_i & a_i & r_i \end{pmatrix} \tag{12}$$

3.4 Masking Metric

3.4.1 MPEG Psychoacoustic Model

Audio coding or audio compression algorithms compress the audio data in large part by removing the acoustically irrelevant parts of the audio signal. The MPEG psychoacoustic model [59] plays a central role in the compression algorithm. This model produces a time-adaptive spectral pattern that emulates the sensitivity of the human sound perception system. The model analyses the signal, and computes the masking thresholds as a function of frequency [12], [59], [60]. The block diagram in Figure 6 illustrates the simplified stages involved in the psychoacoustic model.
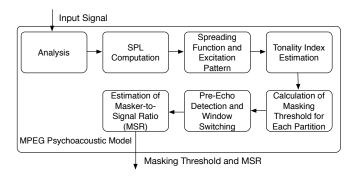

Fig. 6. Flowchart of the MPEG psychoacoustic model [59].

The procedure to derive masking thresholds is summarised as follows. The complex spectrum of the input signal is calculated using a standard forward FFT. A tonality index as a function of frequency is calculated based on the local peaks of the audio power spectrum. This index gives a measure of whether a component is more tone-like or noise-like. This index is then interpolated between pure tone-masking-noise and noise-masking-tone values. The tonality index is based on a measure of predictability, where tonal components are more predictable and thus will have higher tonality indices [61].

A strong signal component reduces the audibility of weaker components in the same critical band and also the neighbouring bands. The psychoacoustic model emulates this by applying a spreading function to spread the energy of a critical band across other bands. The total masking energy of the audio frame is derived from the convolution of the spreading function with each of the maskers. The spreading function, s_f (measured in dB) used in this model is given by

$$s_f(i,j) = \begin{cases} 0 & B(z) \leq 0 \\ x^{\frac{x+B(d_z)}{10}} & \text{else} \end{cases} \quad (13)$$

where the calculation of $B(d_z)$ can be found in [14]. d_z is the bark distance between maskee and masker. Conversion between bark scale and frequency Hz can be approximated by

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan((f/7500)^2). \quad (14)$$

The spreading function is then convolved with the partitioned, renormalised energy to derive the excitation

pattern in threshold partitions. The masking threshold is determined by providing an offset to the excitation pattern, where the value of the offset strongly depends on the nature of the masker. The tonality indices evaluated for each partition are used to determine the offset of the renormalised convolved signal energy [59], which converts it into the global masking level. The values for the offset are interpolated based on the tonality index of a noise masker to a frequency-dependent value defined in the standard for a tonal masker. The interpolated offset is compared with a frequency dependent minimum value, *minval*, defined in the MPEG-1 standard and the larger value is used as the signal to noise ratio. In the standard, Noise Masking Tone is set to 6 dB and Tone Masking Noise to 29 dB for all partitions. The offset is obtained by weighting the maskers with the estimated tonality index. The partitioned threshold derived for the current frame is compared with that of the two previous frames and the threshold in quiet. The maximum of three values is chosen to be the actual threshold.

The energy in each scale-factor band, $E_{sf}(sb)$ and the threshold in each scale-factor band, $T(sb)$ are calculated as described in [14], in a similar way. Thus the final masker-to-signal ratio (MSR) in each scale-factor band is defined as

$$\text{MSR}(sb) = 10\log_{10}(\frac{T(sb)}{E_{sf}(sb)}) \quad (15)$$

3.4.2 Cross-adaptive MPEG Masking Metric

We adapt the masking threshold algorithm from MPEG audio coding into a multitrack masking metric based on a cross-adaptive architecture [36], [37]. The flowchart of the system is illustrated in Figure 7.
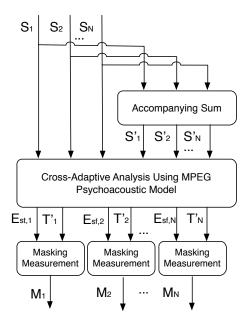

Fig. 7. System flowchart of proposed cross-adaptive multitrack masking model.

To account for the masking that is imposed on an arbitrary track by the other accompanying tracks rather than by itself, we replace $T(sb)$ with $T'(sb)$, which is the masking

threshold of track n caused by the sum of its accompanying tracks. Let H denote all the mathematical transformations of the MPEG psychoacoustic model to derive the masking threshold. We thus can compute $T'(sb)$ as

$$T'_n(sb) = H\Big(\sum_{i=1,i\neq n}^{N} s_i\Big) \quad (16)$$

$E_{sf,n}(sb)$ denotes the energy at each scale-factor band of track n. We assume masking occurs at any scale-factor band where $T'_n(sb) > E(sb)$. The masker to signal ratio in multitrack content becomes

$$\text{MSR}_n(sb) = 10\log_{10}\frac{T'_{sb}}{E_{sf,n}(sb)} \quad (17)$$

We then can define a cross-adaptive multitrack masking, M_n as

$$M_n = \sum_{sb \subset E_{sf,n} < T'_n} \frac{\text{MSR}_n(sb)}{T_{max}} \quad (18)$$

where T_{max} is the predefined maximum amount of masking distance between $T(sb)$ and $E_{sf}(sb)$ for each scale-factor band, which is set to 20 dB.

3.5 Numerical Optimisation Algorithm

The multitrack masking minimisation process is treated as an optimisation problem concerned with minimising a vector-valued objective function described by the masking metric. It systematically varies the input variables, which are the control parameters of the audio effect to be applied, and computes the value of the function until the error of the objective function is within a tolerance value (0.05), reaches the maximum number of iterations or the masking metric is reduced to zero.

3.5.1 Function Bounds

The minimum and maximum values we used for the 6-band equaliser and the dynamic range compressors were set based on audio engineering literature and having consulted a professional practitioner in the audio engineering field [5], [57], [62], [63]. These are detailed in Table 2.

TABLE 2
The minimum and maximum values used for the different types of audio processing used during the optimisation procedure.

Audio Process	Min Value	Max Value
Instrument EQ Gain Bands 1- 6	-6 db	+ 6 db
Subgroup EQ Gain Bands 1- 6	-3 db	+ 3 db
Instrument DRC Ratio	1	6
Subgroup DRC Ratio	1	6
Instrument DRC Threshold	-30 db	0 db
Subgroup DRC Threshold	-30 db	0 db
Instrument DRC Attack	0.005 secs	0.25 secs
Subgroup DRC Attack	0.005 secs	0.25 secs
Instrument DRC Release	0.005 secs	3 secs
Subgroup DRC Release	0.005 secs	3 secs

We used smaller minimum and maximum equalisation gains when we were mixing the subgroups together, since the majority of the inter-channel auditory masking would have been removed when mixing the individual instrument tracks.

3.5.2 Objective Function

A numerical optimisation approach was used in order to derive an optimal set of inputs which would result in a balanced mix. Before defining the objective functions a number of parameters are defined which were used with the optimisation algorithm.

Let A denote the total number of tracks in the multitrack and K denote the total number of the control parameters. The masking metrics are given by $M_i(\boldsymbol{x})$, for $i = 1,\ldots,n$. These describe the amount of masking in each track as a function of the control parameters $\boldsymbol{x}$. Note that $\boldsymbol{x}$ represents the whole set of the control parameters for all tracks. The values of $\boldsymbol{x}$ tend to have multitrack influences, due to the complexity and nonlinearity of the perception of masking. Changes in the control parameter for one track not only affect the masking of that particular track itself but also masking of all other tracks.

The total amount of masking, $M_T(\boldsymbol{x})$, can be expressed as the sum of squares of $M_i(\boldsymbol{x})$, for $i = 1,\ldots,n$,

$$M_T(\boldsymbol{x}) = \sum_{i=1}^{A} M_i^2(\boldsymbol{x}) \quad (19)$$

It is desired to minimise the sum of the masking across tracks and so (19) can be used as the first part of the objective function.

The second objective is that the masking is balanced, i.e., there is not a significant difference between masking levels. Here a maximum masking difference based objective is formed as follows:

$$M_d(\boldsymbol{x}) = \max(\| M_i(\boldsymbol{x}) - M_j(\boldsymbol{x}) \|),$$
$$\text{for } i = 1,\ldots,n, j = 1,\ldots,n, i \neq j \quad (20)$$

This allows this second part of the objective to be used within a min-max framework, similar to that used in [64].

Combining the two objective functions, the following optimisation problem is solved to give $\boldsymbol{x}$:

$$\boldsymbol{x} = \min_{\boldsymbol{x}} M_T(\boldsymbol{x}) + M_d(\boldsymbol{x}) \quad (21)$$

The optimisation problem is a nonlinear, non-convex formulation, and the only information available to the optimisation routine were returns of the function values. Thus a Particle Swarm Optimisation (PSO) approach was used to guide the optimisation routine about the solution space.

3.6 Experiment Setup

3.6.1 Participants

Twenty four participants, all of good hearing, were recruited. 20 were male, 4 were female and their ages ranged from 23 to 52 ($\mu = 30.09, \sigma^2 = 6.2$). All participants had some degree of critical listening skills, i.e, the participant knew what critical listening involved and had been trained to do so previously or had worked in a studio.

3.6.2 Stimuli

There were five songs used in the experiment, where there were five different 30 sec. mono mixes of each song. Two of the mixes were automatically generated using our proposed mix algorithm, where one mix used subgroups and the other did not. There was one mix that was just a straight sum of all

the raw audio tracks. Finally, there were two human mixes, where we selected the low quality mix and high quality mix of each song as determined from a previous experiment. The human mixes were created using standard audio processing tools available in Pro Tools, where we were able to get each mix without the added reverb [42]. The mixes were created with intention of producing the best possible mix. The songs were sourced from the Open Multitrack Testbed [65]. We loudness normalised all of the mixes using the ITU-R BS. 1770-2 specification [56] to avoid bias towards mixes which were louder than others. The song name, genre, number of tracks, number of subgroups and how many of each instrument type there were is shown in Table 3

3.6.3 Pre-Experiment Questionnaire

We provided a pre-experiment questionnaire. The pre-experiment questionnaire asked simple questions related to age, hearing, musical experience, music production experience, music genre preference and each participant's confidence in their critical listening skills. There was also a question with respect to how tired they were when they started the study. If any participant indicated that they were very tired, we asked them to attempt the experiment at a later time once they were rested.

3.6.4 Tasks

We explained to each participant how the experiment would proceed. They were also supervised during the experiment in the event a participant was unsure about anything.

There were two experiment types, where half the participants did experiment type 1 (E1) and the other half did experiment type 2 (E2). Each experiment type had two parts, where the second part was common to both. In E1 (i), we required the participants to rate each of the five mixes of each song they listened to in terms of their preference. In E2 (i), we required the participants to rate each of the five mixes of each song they listened to in terms of how well they could distinguish each of the sources present in the mix (Mix Clarity). In E1 (ii) and E2 (ii) each participant had to listen and compare the automatically generated mixes. They then had to each rate mix for their perceived emotion of each mix along three scales. The scales were Arousal, Valence and Tension (A-V-T). All the songs and mixes used in the experiment were presented in random in order.

After all mixes were rated, participants were asked to provide some feedback on how the experiment was conducted and what their impressions were of the mixes they heard.

3.6.5 Setup and User Interface

The experiment either took place in a dedicated listening room at the university or at an external music studio environment. Each participant was sat at a studio desk in front of the laptop used for the experiment. The audio was heard over either a pair of PMC AML2 loudspeakers or Sennheiser HD-25 headphones, where the participant could adjust the volume of the audio to a comfortable level.

Mix preference and self-report scores were recorded into a bespoke software program developed for this experiment. The software was designed to allow the experiment to run without the need for assistance, and the graphical user interface was designed to be as aesthetically neutral as possible, so as not to have any effect on the results.

4 RESULTS

In this section we present the results related to the optimisation procedure used to generate the automatic mixes. Furthermore, we present the results of the subjective evaluation of the automatic mixes, where the mixes were rated for preference, clarity and the participant's perceived emotion. We have placed all the mixed and unmixed audio used in this experiment in an online repository at https://goo.gl/U2F3ed.

4.1 Results of Optimised Automatic Mixing

In Figure 8 we present the results of the optimisation process used to mix "In the Meantime", for mixing each of the different subgroups, mixing the subgroups and mixing all the tracks together as one. The x-axis on the graph indicates how many iterations of the optimisation process occurred before a solution was found was found. The y-axis indicates masking was present. The results for the other four songs analysed follow a similar trend.
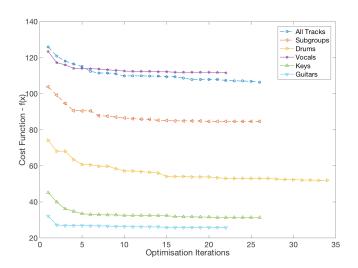

Fig. 8. Cost function value ($f(x)$) for "In The Meantime" plotted against the number of optimisation function iterations.

When the vocal tracks (Vocals) were being mixed, the amount of inter-channel masking that occurred was similar to that of all the tracks being mixed (All Tracks), but took less time to find an optimal solution. This suggests that a lot of the inter-channel masking occurred among the vocalists.

As expected, subgroups with less tracks generally took less iterations to converge. Drums were the instrument type which took the most iterations to converge, with the exception of "Lead Me". This is only partly explained by the number of sources in the drums subgroup, since it often took more iterations than when mixing all raw tracks.

We summarise these results in Figure 4. In this table we present how many iterations were required to mix each type of each song, the change in masking that occurred and the average amount of masking that remained. The numbers in parentheses are the number of tracks used to do the average

TABLE 3
The audio tracks names, genre types, total number of tracks mixed, number of subgroups mixed and the total number of individual instrument tracks mixed.

Track Name	Genre	No. Tracks	No. Subgroups	No. Drums	No. Vox	No. Bass	No. Keys	No. Guitars
In the Meantime	Funk	24	5	10	6	1	4	2
Lead Me	Pop-Rock	19	5	9	2	1	2	5
Not Alone	Funk	24	5	8	9	1	4	2
Red to Blue	Pop-Rock	14	4	9	1	1	0	3
Under a Covered Sky	Pop-Rock	25	5	9	5	1	2	8

calculation. It is clear that applying subgroups to generate stems rather than raw tracks both results in less iterations and a greater overall reduction in masking.

TABLE 4
Number of optimisation iterations required, the change in masking M, and the average masking M where the number of tracks mixed is in brackets.

	No. Iter	ΔM	μM
In the Meantime - All Tracks	26	19.6	4.43 (24)
In the Meantime - Subgroups	25	19.28	16.92 (5)
Lead Me - All Tracks	31	35.3	6.37 (19)
Lead Me - Subgroups	25	16.98	18.66 (5)
Not Alone - All Tracks	26	27.1	6.81 (24)
Not Alone - Subgroups	24	19	20.56 (5)
Red to Blue - All Tracks	37	39.6	7.7 (14)
Red to Blue - Subgroups	24	17.6	26.13 (4)
Under a Covered Sky - All Tracks	51	45.4	25 (4.82)
Under A Covered Sky - Subgroups	25	18.57	19.85 (5)

4.2 Subjective Evaluation Results

4.2.1 Mix Preference

We asked half of the participants to rate each mix based on their preference (E1). The results are illustrated in Figure 9.

In Figure 9 we see the results for each of the five songs used in the experiment, where they are organised by mix type. The figure shows the mean values across all participants, where the red boxes are the 95% confidence intervals and the thin vertical lines represent 1 standard deviation. The songs are ordered for each mix type as follows: "In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".

The mean scores for the summed mixes hover around 0.2, and were never greater than any of the corresponding automatic mixes. However, we see overlapping confidence intervals for all the summed mixes and the automatic mixes without subgroups. Furthermore, there is also some slight overlap with the automatic mixes that use subgroups, but it is not prevalent.

When we compare the two automatic mix types for each song, we see that the automatic mixes that used subgroups were preferred more on average than the automatic mixes that did not use subgroups. This supports our main hypothesis about subgroups improving the perceived mix quality of an automatic mix. However, we see overlapping confidence intervals for "In the Meantime", "Not Alone" and "Under a Covered Sky".

On comparing the automatic mixes to the human mixes, we see the human mixes outperforming the automatic mixes

in nearly all cases except for "Lead Me". In the case of "Lead Me", the automatic mix with subgrouping scores 0.6 on average, while the human low quality mix scores 0.27. There are also overlapping confidence intervals between "Lead Me" for mix types Automatic Mix - S and Human Mix - HQ, "Not Alone" for mix types Automatic Mix - S and Human Mix - LQ and "Under a Covered Sky" for mix types Automatic Mix - S and Human Mix - HQ.

In Figure 10 we see the results for each of the individual mixes, but where we have taken mean across all the different songs. The red boxes are the 95% confidence intervals and the thin vertical lines represent 1 standard deviation. We see there is a trend in increasing means going from Summed mix all the way to Human Mix - HQ. It is apparent that the automatic mixes have performed better than the summed mixes, which supports our main hypothesis, however there is very slight confident interval overlap between Summed Mixes and Automatic Mix - NS. In support of our second hypothesis we can clearly see that there is a preference for the mixes that use subgroups. However, we do not see any confidence interval overlap with either of the human mix types.

4.2.2 Mix Clarity

We also asked the other half of all the participants to rate the mixes in terms of perceived clarity (E2). The results are illustrated in Figure 11.

In Figure 11 we see the results for each of the five songs used in the experiment, where they are organised by mix type. The results are illustrated similarly to Figure 9.

As in Figure 9, the mean scores for the summed mixes are never greater than any of the corresponding automatic mixes. This indicates that the automatic mixes were perceived to have greater clarity on average than the summed mixes. However, we do see overlapping confidence intervals for all the summed mixes and the automatic mixes without subgroups. Furthermore, this also occurred for the songs "In the Meantime" and "Red to Blue" when we compared Summed mix to Automatic Mix - S.

When we compare the two automatic mix types for each song, we see that the automatic mixes that used subgroups had a better clarity rating on average than the automatic mixes that did not use subgroups in only three of the five songs. We also see overlapping confidence intervals for four of the five songs.

On comparing the automatic mixes to the human mixes, we see the human mixes outperforming the automatic mixes in nearly all cases except for "Lead Me". In the case of "Lead Me", the automatic mix with subgrouping scores 0.58 on average, while the low quality mix scores 0.4. There are also
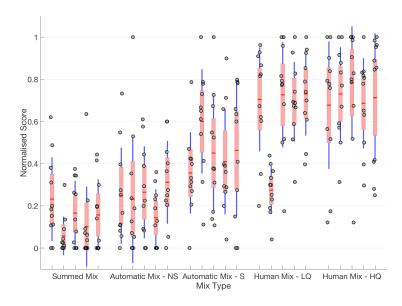
Fig. 9. Results for mix preference based on mix type for each of the individual songs (E1). The songs are ordered for each mix type as follows: "In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".
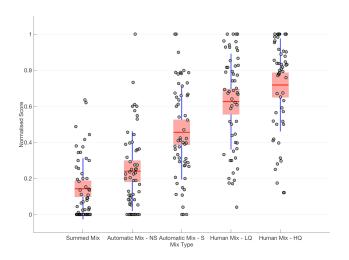

Fig. 10. Results for mix preference based on mix type for all songs (E1).

overlapping confidence intervals between "Lead Me" for mix types Automatic Mix - NS and Human Mix - LQ, "Lead Me" for mix types Automatic Mix - S and Human Mix - HQ and "Under a Covered Sky" for mix types Automatic Mix - S and Human Mix - HQ.

Again we see in Figure 12 there is a trend in increasing means going from Summed mix all the way to Human Mix - HQ. It is apparent that the automatic mixes have performed better than the summed mixes in terms of clarity, which supports our main hypothesis that we are reducing auditory masking. And in support of our second hypothesis, there is a preference in terms of clarity for the mixes that use subgroups.

4.2.3 Perceived Emotion

We asked each of the participants to listen to all the the automatic mixes with subgroups and without subgroups side by side. This was so that they could indicate if they could perceive an emotional difference between each of the two mixes along the three affect dimensions: arousal, valence and dominance. We used the results to test the hypothesis that using subgroups can have an emotional impact on the perceived emotions of the listener. We found our hypothesis to be true in only 1 out of 15 cases (5 songs measured along 3 affect dimensions). The one significant result we found is illustrated in Figure 13.

4.3 Summary

Table 4 and Figure 8 objectively show that our proposed intelligent mixing system is able to reduce the amount of inter-channel auditory masking that occurs by changing the parameters of the equaliser and dynamic range compressor on each audio track. In all mixing cases it was able to reduce the amount of inter-channel masking after a few iterations of the optimisation procedure. Table 4 shows that the reduction in masking was significantly less in four out of the five songs when mixing Subgroups versus All Tracks. This suggests a lot of the masking had been reduced when mixing the subgroups, where the instrumentation would have been similar.

In Figure 14 we present the mean score for each mix type for each of the participating groups, where group 1 evaluated each mix for preference and group 2 evaluated the mixes for clarity. We see that the automatic mixes were preferred more on average than the summed mixes, which agrees with our main hypothesis. However, the automatic mixes never outperformed the human mixes. We also see that the automatic mixes that used subgroups were preferred more on average than the automatic mixes that did not use subgroups. This supports our second hypothesis. However, there were three cases of overlapping confidence intervals. Figure 14 does not show any evidence our second hypothesis is true.

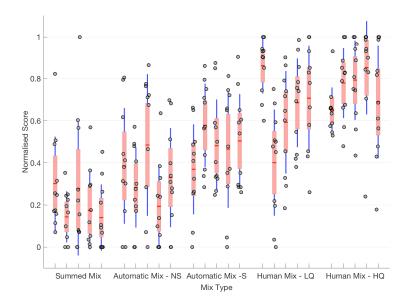When we examine the results for Group 2, which are denoted by the light coloured bars in Figure 14, we see

Fig. 11. Results for mix clarity based on mix type for each of the individual songs (E2). The songs going from left to right for each mix type are "In the Meantime", "Lead Me", "Not Alone", "Red to Blue" and "Under a Covered Sky".
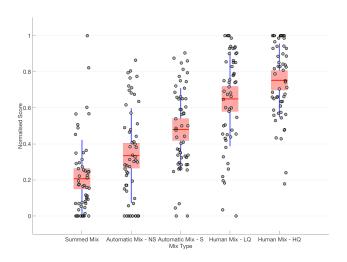

Fig. 12. Results for mix clarity based on mix type for all songs (E2).

Fig. 13. Box plot of perceived arousal for "Not Alone".

that the automatic mixes were preferred more on average than the summed mixes for clarity, which agrees with our main hypothesis. The results do not show any evidence our proposed de-masking method provides any more clarity to a mix than a human can on average. However, one automatic mix with subgroups performed better than human mix. Also, there were overlapping confidence for two automatic mixes and two human mixes with respect to clarity. We see that the automatic mixes that used subgroups had better perceived clarity on average than the automatic mixes that did not use subgroups. This supports our second hypothesis. However, when we examined the clarity results for the individual songs this only occurred for three songs and there were overlapping confidence intervals for four songs.

The results for the mix clarity group are higher on average than the mix preference group. This might suggest that the technique presented here might be better just as a de-masking technique than an overall mixing technique or

just that people are more likely to give higher marks for the word "Clarity" than for the word "Preference".

We were only able to show there was a significant difference in perceived emotions for 1 out of the 15 cases tested. This suggests out third hypothesis cannot be accepted to be true.

5 CONCLUSION

This paper described the automation of loudness normalisation, equalisation and dynamic range compression in order to improve the overall quality of a mix by reducing the inter-channel auditory masking. We adapted and extended the masking threshold algorithm of the MPEG psychoacoustic model in order to measure inter-channel auditory masking. Ultimately, we proposed an intelligent system for masking minimisation using a numerical optimisation technique. We tested the hypothesis that our proposed intelligent system can be used to generate an automatic mix with reduced
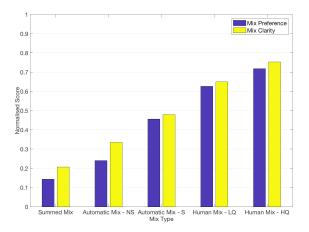
Fig. 14. Mean scores of each mix type for each group, where the blue bars represent mix preference and the yellow bar represents mix clarity

auditory masking and improved perceived quality. This paper also tested the hypothesis that using subgroups when generating an automatic mix can improve the perceived mix quality and clarity of a mix. We further tested to see if using subgrouping or not affects the perceived emotion in an automatic mix. We evaluated all our hypotheses through a subjective listening test.

We were able to show objectively and subjectively that the novel intelligent mixing system we proposed reduced the amount of inter-channel auditory masking that occurred in each of the mixes and it improved the perceived quality. However, the results did not match the results of the human mixes in most cases.

Furthermore, the results of the subjective listening test implied that subgrouping improves the perceived quality and perceived clarity in an automatic mix over automatic mixes that do not use subgroups. However, the results suggested that using subgroups had very little effect if any on the perceived emotion in any of the mixes. It was only shown to be true in 1 out of the 15 cases.

6 FUTURE WORK

It is clear that our proposed intelligent mixing system has scope for improvement. One way in which this could be improved is if the equalisation and dynamic range compression settings changed on a frame by frame based on the inter-channel auditory masking metric. Currently the equalisation and dynamic range settings are static for the entire track. One of our more experienced participants in the subjective listening test mentioned that they could hear this.

We also believe the optimisation procedure could be improved by having a larger optimality tolerance, where once this tolerance has been reached another nonlinear solver begins, using the PSO results as initial conditions. If we examine Figure 8 we see that many of the optimisation procedures find a satisfactory solution in less than ten iterations.

We would also like to see this intelligent system used in combination with panning. We would have liked to have implemented panning, but we believe this would

have removed the majority of the masking present in the mix and would have made it difficult to demonstrate the effectiveness of the inter-channel auditory masking metric.

The process of applying the correct gain, equalisation and dynamic range settings in a multitrack is a challenging and time consuming task. We believe the framework we proposed here could be useful in developing systems for beginner and amateur music producers where it could be an assistive tool, giving initial settings for compressors and EQs on all tracks, that are then refined by the mix engineer.

REFERENCES

[1] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[2] A. J. Oxenham and B. C. Moore, "Modeling the additivity of nonsimultaneous masking," *Hearing research*, vol. 80, no. 1, pp. 105–118, 1994.

[3] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*, vol. 22. Springer Science & Business Media, 2013.

[4] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[5] R. Izhaki, *Mixing audio: concepts, practices and tools*. Taylor & Francis, 2013.

[6] Z. Ma, J. D. Reiss, and D. A. Black, "Partial loudness in multitrack mixing," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Audio Engineering Society, 2014.

[7] J. E. Dennis Jr and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[8] P. E. Gill and W. Murray, *Numerical methods for constrained optimization*. Academic Pr, 1974.

[9] P. D. L. G. Pestana, *Automatic mixing systems using adaptive digital audio effects*. PhD thesis, Universidade Católica Portuguesa, 2013.

[10] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of machine learning*, pp. 760–766, Springer, 2011.

[11] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[12] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.

[13] A. Gersho, "Advances in speech and audio compression," *Proceedings of the IEEE*, vol. 82, no. 6, pp. 900–918, 1994.

[14] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "Iso/iec mpeg-2 advanced audio coding," *Journal of the Audio engineering society*, vol. 45, no. 10, pp. 789–814, 1997.

[15] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[16] M. M. Goodwin, A. J. Hipple, and B. Link, "Predicting and preventing unmasking incurred in coded audio post-processing," *IEEE transactions on speech and audio processing*, vol. 13, no. 1, pp. 32–41, 2005.

[17] A. Robert and J. Picard, "On the use of masking models for image and audio watermarking," *IEEE transactions on multimedia*, vol. 7, no. 4, pp. 727–739, 2005.

[18] C. Maha, E. Maher, and B. A. Chokri, "A blind audio watermarking scheme based on neural network and psychoacoustic model with error correcting code in wavelet domain," in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, pp. 1138–1143, IEEE, 2008.

[19] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 310–319, 2007.

[20] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, vol. 10, pp. 608–611, IEEE, 1985.

[21] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[22] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio qualitytechnology and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1890–1901, 2006.

[23] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 34–49, 2010.

[24] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effectivesignal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[25] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[26] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 422–438, 2008.

[27] B. R. Glasberg and B. C. Moore, "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 906–918, 2005.

[28] B. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[29] B. R. Glasberg and B. C. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[30] S. Vega and J. Janer, "Quantifying masking in multi-track recordings," in *Proceedings of SMC Conference*, 2010.

[31] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, "Describing the transparency of mixdowns: The masked-to-unmasked-ratio," in *Audio Engineering Society Convention 130*, Audio Engineering Society, 2011.

[32] E. Perez-Gonzalez and J. Reiss, "Automatic equalization of multi-channel audio using cross-adaptive methods," in *Audio Engineering Society Convention 127*, Audio Engineering Society, 2009.

[33] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalization based on masking reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, pp. 312–323, 2015.

[34] D. Ward, J. D. Reiss, and C. Athwal, "Multitrack mixing using a model of loudness and partial loudness," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[35] M. Terrell, A. Simpson, and M. Sandler, "The mathematics of mixing," *Journal of the audio engineering society*, vol. 62, no. 1/2, pp. 4–13, 2014.

[36] U. Zölzer, *DAFX: digital audio effects*. John Wiley & Sons, 2011.

[37] J. D. Reiss, "Intelligent systems for mixing multichannel audio," in *17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, IEEE, 2011.

[38] B. C. Moore, "Masking in the human auditory system," in *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, Audio Engineering Society, 1996.

[39] B. Moore, "An introduction to the psychology of hearing, academic," *San Diego*, 2003.

[40] G. Wichern, H. Robertson, and A. Wishnick, "Quantitative analysis of masking in multitrack mixes using loudness loss," in *Audio Engineering Society Convention 141*, Audio Engineering Society, 2016.

[41] P. Pestana and J. Reiss, "Intelligent audio production strategies informed by best practices," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Audio Engineering Society, 2014.

[42] B. De Man, B. Leonard, R. King, and J. D. Reiss, "An analysis and evaluation of audio features for multitrack music mixtures," in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, October 2014.

[43] J. J. Scott and Y. E. Kim, "Instrument identification informed multi-track mixing.," in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pp. 305–310, 2013.

[44] B. De Man and J. D. Reiss, "A knowledge-engineered autonomous mixing system," in *Audio Engineering Society Convention 135*, Audio Engineering Society, 2013.

[45] D. Ronan, B. De Man, H. Gunes, and J. D. Reiss, "The impact of subgrouping practices on the perception of multitrack mixes," in *Audio Engineering Society Convention 139*, Audio Engineering Society, 2015.

[46] D. Ronan, H. Gunes, and J. D. Reiss, "Analysis of the subgrouping practices of professional mix engineers," in *Audio Engineering Society Convention 142*, Audio Engineering Society, 2017.

[47] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[48] C. E. Izard, "Basic emotions, natural kinds, emotion schemas, and a new paradigm," *Perspectives on psychological science*, vol. 2, no. 3, pp. 260–280, 2007.

[49] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[50] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller, "Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music.," *Emotion*, vol. 7, no. 4, p. 774, 2007.

[51] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models.," in *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pp. 621–626, 2009.

[52] P. G. Hunter and E. G. Schellenberg, "Music and emotion," in *Music perception*, pp. 129–164, Springer, 2010.

[53] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?," *Musicae Scientiae*, vol. 5, no. 1 suppl, pp. 123–147, 2002.

[54] E. Schubert, "Continuous self-report methods," *Handbook of music and emotion: Theory, research, applications*, vol. 2, pp. 223–253, 2010.

[55] D. Ronan, D. Moffat, H. Gunes, and J. D. Reiss, "Automatic sub-grouping of multitrack audio," in *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[56] R. ITU-R, "Itu-r bs. 1770-2, algorithms to measure audio programme loudness and true-peak audio level," *International Telecommunications Union, Geneva*, 2011.

[57] A. U. Case, *Mix smart*. Focal Press, 2011.

[58] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 399–408, 2012.

[59] K. Brandenburg and G. Stoll, "Iso/mpeg-1 audio: A generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[60] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 2524–2527, IEEE, 1988.

[61] D. Pan, "A tutorial on mpeg/audio compression," *IEEE multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[62] B. Owsinski, *The mixing engineer's handbook*. Nelson Education, 2013.

[63] Z. Ma, B. De Man, P. D. Pestana, D. A. A. Black, and J. D. Reiss, "Intelligent multitrack dynamic range compression," *Journal of the Audio Engineering Society*, 2015.

[64] P. McNamara and S. McLoone, "Hierarchical demand response for peak minimization using dantzig–wolfe decomposition," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2807–2815, 2015.

[65] B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, "The Open Multitrack Testbed," in *137th Convention of the Audio Engineering Society*, October 2014.